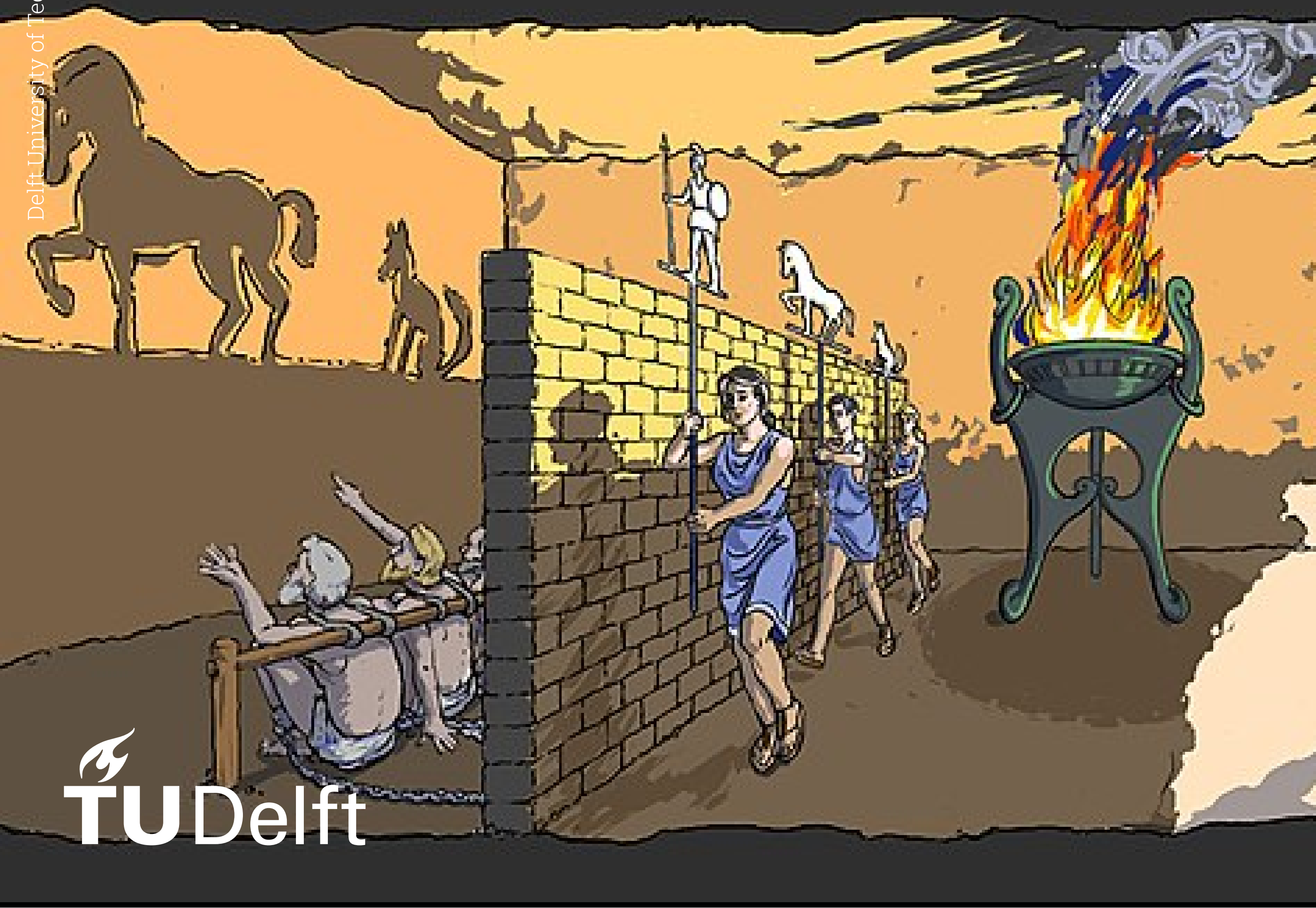
Finding values in green hydrogen using topic modelling

Building a framework for explorative modelling

EPA 2942

Niels de Boer

Delft University of Technology



TU Delft

Finding values in green hydrogen using topic modelling

Building a framework for explorative modelling

Master thesis submitted to Delft University of Technology
in partial fulfilment of the requirements for the degree of
"Master of Science"
in *Engineering policy analysis*

To be defended in public on the 14th of March 2023 at 1100 hours
in the faculty of Technology, Policy and Management, Delft.

By

Niels de Boer

4249496

First supervisor:	Dr.Ir. E.J.L. Chappin, section E and I
Second supervisor:	Prof.Dr.Ir. I.R. van de Poel, section EPT
Advisor:	Ir. T.E. de Wildt, section EPT
Project Duration:	March, 2022 - March, 2023
Faculty:	Faculty of Technology, Policy and Management

Cover:	An Illustration of The Allegory of the Cave, from Plato's "Republic" by Wikimedia Commons user 4edges is licensed under CC BY-SA 4.0.
Style:	TU Delft Report Style, with modifications by Daan Zwaneveld

TU Delft

Preface

Samuel Johnson wrote a series of essays, one of them, called No. 43. The inconveniences of precipitation and confidence. It was published in the Rambler in 1750. Let's look at a quote from this essay.

"There is, indeed, some danger lest he that too scrupulously balances probabilities, and too perspicaciously foresees obstacles, should remain always in a state of inaction, without venturing upon attempts on which he may perhaps spend his labour without advantage."

The version of this quote in modern English is more commonly found today: *"Nothing will ever be accomplished if all objections are considered"*. To some, this is still an exhausting read and for those, there is good news! The spirit of Samuel's message continues to live on in Nike's slogan: "Just do it."

This inspirational quote was love at first sight. To me, it is about dealing with setbacks. You may recognize this when you find yourself putting your own work in doubt and start contemplating. I wrote it down on my whiteboard before starting this thesis in the hope that it would help keep me in check.

The quote is however a bit out of context. It continues with *"But previous despondence is not the fault of those for whom this essay is designed"*. Samuel continues by stating that those who are precipitating (hasting) would never overthink anyways. Samuel compares these dichotomous perspectives to cowardice and arrogance respectively. He admires perseverance instead.

Niels de Boer
Scheveningen, March 2023

Abstract

Increased pace of developments strain the ability of policy makers to be timely and sufficiently informed. While there are already sufficient methods available for gauging what plays a role, topic modelling is a novel method that has the potential to be deployed at high speed with low effort. Values play an important role as it shapes policy making and in turn affects stakeholders. A semi-supervised topic modelling method called correlation explanation (CorEx) was used for this purpose as it allows steering the model to find aforementioned values. Green hydrogen is used as a case study where topic modelling is used to find values in scientific literature on this subject. Green hydrogen is envisioned to play a more important role in the context of decarbonisation. Such a transition has a major influence on society engaging business and households alike. Three different value sets are used reflecting different perspectives. These perspectives focus on corporate values, public values and the context in which hydrogen is discussed respectively. It was found that using topic modelling to identify values is highly constrained by its given inputs and processed outputs however. A methodological framework is therefore proposed. It suggests how topic modelling can be conducted for the purpose of identifying values playing a role in a certain domain. This framework consists of five components which are value definition, corpus selection, language processing, topic modelling and result interpretation. Utilising this framework helps with structuring the topic modelling procedure and identify bottlenecks in result quality.

Summary

Policy makers experience problems being timely and sufficiently informed regarding new technological developments hampering the generation of appropriate policy actions. This problem is partly caused by increased output in scientific literature and acceleration of technological change. Furthermore, values are playing an increasingly important role in policy making. Understanding which values are held by stakeholders and how they change over time enhances cooperation and could lead to more efficient decision-making. A computational method for identifying values in large data sets was brought forward called topic modelling.

A case study was undertaken to apply topic modelling for the purpose of finding values in the domain of green hydrogen. This was done by using a corpus of scientific articles on hydrogen. The goal of this case study was to discover which values play a role in green hydrogen. Various difficulties were encountered in this process and because of that, this research focuses on the how question: How can topic modelling be used to interpret what values are related to hydrogen technologies in scientific articles?

Green hydrogen is envisioned to play a more important role in the context of decarbonisation. Such a transition has a major influence on society engaging business and households alike. Values therefore play an important role in this environment affecting the shape of policy. While there are already sufficient methods available for gauging what plays a role, topic modelling is a novel method that has the potential to be deployed at high speed with low effort.

Various topic modelling tools exist, in this research a specific implementation was chosen. Traditional topic models allow the specification of a specific number of topics after which the model proceeds to classify all documents in these topics. Note that documents can be member of several topics. There is however not a lot of influence the modeller has on the outcomes apart from specifying the number of topics and selecting a good set of documents. The implementation used in this research, correlation explanation or CorEx, has an additional input. This input is called ”anchors” which allows the modeller to give certain words more weight in the model nudging the results in a certain direction. If the right words are chosen it is possible see these anchors back in the results, creating a different result by human input.

In order to complete the goal of the research, various value sets are used as input in the topic model. The corpus and topic modelling method were left unchanged in this process. If the opposite decision was to be legitimised it requires the variation of each value and corpus set on each topic modelling method to see how the results are affected. Since the focus of the research is on on values, it has little added benefit and it is furthermore unrealistic given the time it would take. Going back to the objective, two terms require some explanation before proceeding. These are ”values” and ”importance”. A value is considered to be some subjective judgement on what is important or relevant to a subject in relation to green hydrogen. Constructing a value is complicated as it requires the knowledge on what their relevant elements are which then have to be translated into the model in the form of anchors. Importance refers to placing these judgements in a hierarchy. In practice it meant determining if the value could be found or not. It does not allow the exclusion of topics if they are not found.

Three value sets were used in the case study resulting in three different results. It is not useful to compare the results of different runs that use different values as inputs. The first value set that was used represents values held by corporations or organisations and is called core values. The second value set represents the context in which hydrogen is discussed distinguishing (allowing the identification of green hydrogen documents). Lastly, the third value set, generic values, represents values held generally by the public. The methodology for obtaining results was similar for the first and last set. The second set started without anchors. The model was inclined to discover this context ”naturally”. The words that it provided were fed back into the model to improve the result.

The process of finding if these values are present requires the interpretation of model results. Interpreting the results contains two steps. The first step is interpreting the raw topic modelling data and second is verifying these results by manually reading documents. Interpretation of the raw topic modelling data was done by a manual and automatic method. Criteria have to be established for both

methods. A manual method goes through the data and manually checks if the topic passes certain criteria and if it is worth verifying if the topic is truly present. An automatic method is an algorithm that uses a KPI to do this automatically. Verification corresponds to observation of the documents. Only after the verification steps results are presented. Interpretation of the results happens after presenting these results and is a subjective judgement referring to the ”importance” component of the goal.

Results of the topic model are of course the value sets that were found in the corpus. However the methodology that was used to legitimise this is considered to be more important. An overview of this methodology was made in a framework with distinct components. This framework also highlights the importance aspects that had little to no attention in this research such as corpus selection and language processing of this corpus. Another development is using an automatic method for interpreting the results of the topic model which produces more consistent results and increases the phase at which a topic modelling exercise can be completed. Many opportunities exist for improving topic modelling. These opportunities are related to presentation (visualising), workflow (standardisation, creating a framework) and quality (topic modelling results).

Contents

List of Figures

List of Tables

Nomenclature

Definitions

Term	Definition
Anchor	Anchors are consisting of anchor words which fit in to some overarching idea, set of assumptions or concept.
Anchor strength	Anchor strength is the predefined bias allocated to an anchor word during each learning step in the model. It overrides whatever learning took place in the previous step. Anchor strength allows n-grams to be excluded from the corpus, nudging the model to allocate the n-gram to a specific topic or nudge the model to include the n-gram in any of the latent topics. Anchor strength corresponds to the strength of this nudge and typically is several times stronger than the maximum strength that can be given to ordinary n-grams in the learning process.
Anchor set	An anchor set are a group of anchors that can collectively be described by the same overarching idea, set of assumptions or concept that are used in (one specific) topic modelling exercise(s).
Anchor word	Anchor words are n-grams used as an input in the topic model and part of an anchor part of an anchor that are ideally independent from other anchor's anchor words. Anchor words have a predefined anchor strength.
Corpus	The corpus is the set of documents that is used as input in a topic model.
Document	A document is a collection of n-grams that can contain metadata. A document is in practice synonymous to a scientific article and it can be chosen to include publisher, title or keywords as well. Documents are best understood as the abstracts of scientific articles.
Green hydrogen	Green hydrogen (GH) is hydrogen produced from electrolysis using renewable energy sources only.
Keyword	A keyword is a n-gram allocated to a latent variable with a degree of informativeness. The informativeness is measured in natural unit of information (nat). Direction of informativeness is a distinct property and can be positive or negative meaning that the presence of a word is indicative or counter indicative for a document to be part of a topic respectively.
Latent Topic	A latent topic is the output of the topic model and consists of and only of keywords and member documents. CorEx uses "latent variables" and updates these until the optimisation criteria or maximum iterations (sometimes called layers) has been reached. This last latent variable is equal to the model´s output and are called latent topics.
Mutual information	Is the informativeness of an n-gram over topics that is measured in the unit nat.

Term	Definition
N-gram	A n-gram is a group of tokens of size n that are in the same order as the document they originate from. N-grams are referred to by their number n, such as unigrams, digrams and trigrams. Unigrams are best understood as all the individual words in a document. Digrams as unigrams plus the sets of two words that occur next to each other, etc. for higher n-grams.
Token	A token is a group of symbols in a document that are separated from other each other by a blank space.
Topic	A topic is a category that reflects an abstract concept and an overall idea. A topic can be inferred from a latent topic.
Topic model	A topic model is an algorithm that classifies documents in a corpus into clusters
Value	Schwartz’s definition: ”Values are concepts or believe pertaining to desirable states or behaviours that transcend specific situations, guide the selection or evaluation of behaviour and events and are ordered by relative importance.” (Schwartz & Bilsky, 1990)

Symbols

Symbol	Name	Definition or description
α	N-gram topic distribution matrix	Is the distribution of n-grams over topics, which corresponds to mutual information (of n-grams to topics) parameter, ”mis”. In CorEx α is the ”adjacency matrix” or weight distribution of n-grams over topics.
θ	Document topic distribution matrix	Is the distribution of documents over topics, which corresponds to the $p(y\|x)$ parameter. In CorEx θ is used to indicate four intermediate parameters.
λ	Scalar for word-topic association	λ refers to two things in this research. In CorEx λ refers to the sensitivity in learning word-topic memberships. In the pyLDAvis implementation λ refers to the relevance of word-topic associations.
ϵ	Convergence criterion	Is used in CorEx and can be any real number
X	Corpus	All the documents
x^l or θ	A Document	Is a set of n-grams. The θ symbol is sometimes used instead. This is done in context of $p(y_j\|\theta)$ which refers to the document topic distribution matrix.
x_i or α	An n-gram	See definition of n-gram. N-grams are in the unit interval, meaning that they have a value of or are between zero and one. The α symbol is sometimes used instead. This is done in context of $p(y_j\|\alpha)$ which refers to the n-gram topic distribution matrix.
x	All n-grams in the corpus	$x = \sum_{i=0}^{i} x_i$
x_i^l	All n-grams in a document	Contains all n-grams of the corpus, i, specifying if it is present in a specific document, l
i	Number of n-grams	Is a natural number.
j	Number topics	Is a natural number.

Symbol	Name	Definition or description
l	Number of documents	Is a natural number. There are two ways in which i, j and l are counted in this research which are the human and computer way of counting. When a human counts it starts from 1, while a computer starts from zero. The human way of counting is used when talking about the total number of documents, topics, etc. and the computer way of counting is used when displaying equations or when talking about a specific index or identifier, such as in table 3.4.
y_j	A topic	y is a topic where j specifies the specific number of the topic. The total number of topics, j is a model parameter

1 Introduction

Policy makers experience problems being timely and sufficiently informed regarding new technological developments hampering the generation of appropriate policy actions. One of the causes are emerging transformative technologies. Concepts introduced by these transformative technologies interact with governance increasing the complexity of the interaction (Popper, 2002). Since technologies emerge at an accelerated pace, these problems become more apparent over time. An example of this effect is the digitisation, increasing the availability of information complicating the completeness of information retrieval.

Acceleration of technological change means that it is important to collect all relevant information to that policy in a timely manner. Policy can otherwise not keep up with technological change. While not all technologies are transformative, the rapid successive introduction of technologies, which will happen in case of accelerated technological change, implies that it is affecting the interests of all stakeholders. To understand the impact of policy and technology on stakeholders it is important to understand which values are relevant. This introduces a moral aspect and makes the qualitative identification of values related to policy or technology relevant.

1.1. Relevance of finding values in policy making

1.1.1. Challenges in the identification of values

Identification of values is becoming more complicated due to the increased availability of information. Firstly this is caused by the growth of the yearly output of scientific literature making it harder to create a good representation of all information available. Secondly, the production of literature and even literature reviews are biased through forward and backward citing. Otherwise relevant research is excluded by this bias. Similar patterns can be seen in the information produced by media and polities. More information is yielded due to the growth of the number and size of institutions. Media consumption on digital platforms has increased the availability of information, but the diversity of the provided information is much smaller through filter bubbling creating narrow target audiences.

1.1.2. Values and contentious politics

Values play an increasingly important role in policy making. Growth of polities (in the Netherlands) is associated with more institutions, employees and arenas where decision-making takes place. The traditional view of political process theory sees the government as a unified actor. Contention is from this perspective done by social movements that engage in claim-making targeting the government directly or indirectly. Recent efforts have shown that governmental players may just as well be claim makers for legitimising their policies. In this process called contentious governance governmental players seek to mobilise and demobilise social movements in order to legitimise their desired policies (Verhoeven & Bröer, 2015). Such a system is part of the multi-actor system perspective, which is a situation where no single actor is able to dictate or impose their desired policy without cooperating with other actors. Understanding which values are held by stakeholders and how they change over time enhances cooperation and could lead to more efficient decision-making.

1.1.3. Values

Contentious governance was adduced to explain the context in which values are used in this research. We would like to explain the meaning of values in this research. A popular definition of values is the one given by Schwartz defining ”Values as 1) concepts or beliefs, which 2) pertain to desirable end states or behaviours, that 3) transcend specific situations, 4) guide the selection or evaluation of behaviour and events, and 5) are ordered by relative importance” (Schwartz & Bilsky, 1987, 1990). To reiterate this definition, values are concepts or beliefs that relate to positive states or behaviours. Furthermore, values are ordered in a hierarchy and are persistent throughout different situations. While Schwartz’s definition of values is one of the most well-established definitions, values continue to be poorly understood. Particularly regarding delineating the meaning of value from the meaning of norms and whether values are properties of individuals or groups. Understanding values is best done in relation to norms because values overlap with the meaning of norms but differ in some key aspects (Maltseva, 2018).

Norms guide behaviour and operate on the group level where the enforcement of norms is a self-reinforcing process (Searle, 1995). Entitativity is the acknowledgement of the existence of a group and therefore their norms. Norms are not arbitrary but are negotiated and agreed upon within a certain group. Values legitimise norms and the discourse on these values is called culture which in turn gives rise to these norms. Norms are expressed in terms of ”shoulds” and are correspondingly expressed by the word ”shouldness”. Values are positive and learned through rewarding affective states (emotions), while norms are strongly associated with biological adaption, namely learning through stress response and emotions of disgust and anger. The behavioural outcome of a mismatch between a value (on the individual level) and a norm (in a situation) is the result of the valuation of the learned positive and negative affective states of value and norm respectively (Maltseva, 2018).

Values on a group level can be seen as a reflection of culture, which is the consensus and discourse on values held by individuals; resulting from this discourse and consensus are (group) norms (Maltseva, 2018). Going back to policy analysis and contentious governance where values relate to the following. First are the resulting policies which reflect norms and not values. Second is the process of mobilising and demobilising social movements which is an aspect of culture, the discourse on values. Agents, those who do this process, therefore play a central role in culture. A description of a possible mechanism for how this discourse and mobilisation take place is given. Individuals associate and dissociate with social movements based on their values and existing group identity. These roughly correspond to the values held by the individual and the norms it endorses. This results in a set of values and norms held by the group that legitimises the implementation of policy.

Finally, we give some examples of values and norms in relation to this research. Some words which are used in this research are taken and explained in terms of value and norm. First, let us take the word ”flexibility”. It is a value when it is expressed in the form of ”the ability to be flexible is important to me”. Expressing flexibility in terms of a norm corresponds to, for example, ”a (business, organisation or individual) should be flexible”. This means that the word ”flexibility” can imply both value and norm at the same time. Another, but a more devious example is the word ”industry”. It can be a value in: ”I endorse industry”. Values are associated with positive affective states, thus endorse can be replaced by words such as like, support, pride and joy. A norm relating to ”industry” is less obvious and more cumbersome in reasoning. Such a norm would be expressed in terms of policies encouraging large-scale and effective production processes which are implied to be opposed to decentralised, small scale and artisan production processes.

All in all, a word could relate to both a value and a norm but in this research, the detection of these words is seen as the framing of these words in terms of a value. Recall Schwartz’s definition of value to assign the following properties to these words. First, words exist in a hierarchy with other values (held by some individual or group). Second, these words apply (to that same subject) independent of the context. This research does not attempt to illuminate the hierarchy, but to identify the presence of values.

1.2. Finding values using topic modelling

de Wildt et al. (2018) have brought forward a computational method for identifying values in large data sets. This method is called topic modelling and is used to classify large sets of documents in (latent) topics. This method classifies documents based on their word content. No exact match of words is required, but the overall similarity in the usage of words among groups of documents will favour classification into the same topic. Topic modelling is especially interesting if the topic of interest is dispersed among different scientific fields. In such cases, a topic modelling approach creates bridges between disciplines as it allows the identification of other relevant scientific literature and fields. Moreover, topic modelling could be used for identifying or detecting value change (de Wildt et al., 2022).

How is this identification or detection done by topic modelling? This question refers to the validity of the method. It addresses the process of using the topic model, which includes handling data inputs and result interpretation. This is relevant because of two reasons. The first is for providing a qualitative result because the result is dependent on the inputs used and that includes values. Therefore the second reason is to understand and communicate to who the results (values) cater. The answer to this question is in the form of a framework allowing topic modelling to suffice in value identification. One notable paper was found that exhibits a topic modelling framework, which is for general purpose (Maier et al., 2018). The framework in this paper is for a specific topic modelling technique serving a specific purpose, namely value identification. It is therefore interesting to look at the creation of a topic modelling framework for value identification.

1.2.1. Topic modelling framework for policy makers

Topic modelling helps with identifying values playing a role in policy and technology but doing so effectively remains challenging. The framework addresses several challenges that are encountered in the process of topic modelling. First, why are values not found in the results and what can I do about it? The framework lists the methods for allowing values to be found. This prevents loss of time by attempting ineffective solutions. It also prevents users from having to go through the source code of the algorithm to understand how the result is obtained to solve this problem. Second, for who are the values relevant? Answers to which group of stakeholders do the results relate. Third, how do I interpret values in the latent topics? Which lists several methods for interpreting the results of the topic model, the latent topics, to determine the presence of values. Employing the principles of the framework as a guideline in the process of topic modelling is necessary for creating meaningful, high-quality results in a timely manner.

1.2.2. Purpose of topic modelling for policy makers

The ideal use of topic modelling, for the purpose of finding values, is compared to other methods to achieve the same goal. Let us first consider consulting expert opinions through interviews or questionnaires. Such methods are more expensive than conducting a topic model as it demands the time and money of one or multiple experts. It is furthermore unknown if this expert is aware of the latest developments while a topic model is guaranteed to be aware of these developments. Topic modelling should not be seen as a replacement, but as an complement to expert consultation. A true expert is on the forefront of development in the field and to use an algorithm that reflects on these actions of these persons as authority over these persons is rubbish. On the other hand, the advantage of the topic model is that it is a machine; where its bias and partiality is made explicit by the input values.

Checking the presence of a single value could be done using a search engine, but doing so for larger sets of values is time consuming and the result will be inconsistent in quality. Topic modelling can check the presence of values in a fraction of the time and does not require manual feeding of each value to the search engine. Feeding the model with hundreds of values is unrealistic as this complicates the interpretation of the results. Another issue is maintaining the same level of quality between values. Judging the presence of each value has to be done in the same manner without bias. Doing so manually will result in propagation of bias and inconsistencies in judgement style. This framework suggests the use of a quantitative method that ensures the maintenance of the same judgement criteria (whether a value is present or not) based on informativeness of the documents. Lastly is the aspect of reproducibility, which is maintained by topic modelling as the method and corpus are stored. Storing search queries is not enough to ensure reproducibility because search engines are ever-changing and not transparent in the way they deliver their results.

1.3. Towards a methodology for value identification

This process of value identification through topic modelling is done with the use of a case study on green hydrogen. This case study aims to identify the values present in this corpus. During this process, it was discovered that the identification of values was influenced by many factors outside the topic modelling process. While an overview of values playing a role in this case study was interesting, it was deemed more relevant to create an overview of the factors that influence and limit the identification of these values. Such an overview, call it a framework, does not exist yet for this purpose. Frameworks for topic models do exist, the most notable is Maier et al. (2018). This paper stands out because it distils a standardised process for conducting topic modelling focusing on the entire process that includes handling of inputs and outputs of the topic model. Other topic modelling reviews typically compare topic modelling methods. Here the mathematical details are compared while the overall process is regarded to be out of scope. However, the topic modelling framework by (Maier et al., 2018) holds for a specific topic modelling technique called LDA (latent Dirichlet Allocation) and regards the general-purpose use of this method. The method used in this research is semi-supervised, unlike LDA which is an unsupervised method. In this semi-supervised approach, the values that one wishes to identify are used as additional input. Therefore, the processing of results is comparatively of different nature as they refer back to the inputs used in the model, while in LDA these results are ”standalone”. These key differences legitimise the creation of a framework for conducting and evaluating the validity of identifying values using topic modelling.

What results from the previous deliberations and motivations leads to the **main research question** of this thesis. **”How can topic modelling be used to interpret what values are related to hydrogen technologies in scientific articles?”** The answer to this question is a framework that illustrates a methodology for conducting topic modelling for the purpose of value identification.

This answer is developed through the use of a case study; although this case study was initiated for a different reason. This initial research objective was to identify **”which values play a role in scientific literature on green hydrogen”**. Identifying values proved to be more challenging, explaining the change of this main effort (to the main research question). The reason for this is that under the given conditions (of inputs and the specific topic modelling method), the model does not prefer to create topics based on ”values”. It heavily favours context instead. This aspect is handled by the second research question. Since it takes a lot of effort to identify values heavy constraints are emplaced on the validity of the results. Validity is here related to the trustworthiness and reproducibility of the results. It is therefore more appropriate to improve this validity instead of seeking to answer this research objective. The creation of a framework helps to improve the validity by targeting critical processes. This also helps others understand the process of topic modelling to identify values.

The **first subquestion** relates to the entire topic modelling process: **”How can it be determined which values are the most important or relevant?”**. It includes the processing of inputs and results as these affect the results. The answer to this question is the framework for creating an overview without going into detail in each step. The last subquestion handles the interpretation compartment. The first compartment, defining values, is varied by each section in the results. The remaining three compartments of the framework were left constant in this research. Implications of varying these compartments are discussed instead.

The **second subquestion** seeks to identify the context in which green hydrogen is discussed. **”How is green hydrogen distinguished in the corpus?”** This was done because of the initial difficulties with identifying values. It relates to the initial goal and not to the main research question of this research. When this question is related to the main research question it seeks to answer the question **What sensible topics will be produced by the topic model if no values are used as input?**. Under these conditions, the model will, for this corpus, produce topics containing the context in which hydrogen is discussed. The sensible and coherent topic outputs are corresponding to disciplines and technical terms. This research question is handled in section 3.2.

The **third and last subquestion** relates to the processing of the results of the topic model in order to create a judgement regarding the presence of values in the corpus. **”How can topic modelling results be interpreted?”** This relates to the last step, interpretation, in the framework generated by the first subquestion. The process of result interpretation consists of two steps. First is the processing of the α and then the θ signature of the topic model. These signatures are large matrices where each entry is filled with the probability that a word (α) or document (θ) belongs to a topic. The interpretation of the θ matrix is done manually by reading the documents belonging to a topic. The interpretation of the α matrix can be done manually or automatically.

2 Approach

This chapter explains how the research questions are answered. In the previous chapter, it was explained that there is a main research question and a research objective. The main research question relates to the main effort of this research and the research objective relates to the reason why this was done.

The research objective existed before the main research question and the research setup was made from this objective. During the execution of this setup, it was discovered that providing an answer to this objective is not straightforward. It was at this point that the decision was made that a framework is more useful than answering the objective. The main research question was a consequence of the objective and it, therefore, makes no sense to omit the research objective. Furthermore, the research methodology is attempting to answer the research objective. Following that methodology also happens to answer the main research question. This chapter explains how this research was done.

A short reflection is given on this methodology which should relieve the objections held by the reader that may exist after reading the previous paragraphs. Considering the option to compare different topic modelling exercises to construct the topic modelling framework, similarly to (Maier et al., 2018). This methodology is ideally more suited for creating a framework because it uses more topic modelling exercises. In this research only a few factors were varied and both the corpus and topic model techniques remained constant. While this statement holds in hindsight, it does not consider the constraints in the lack of availability and experience in topic modelling exercises for this specific purpose. The creation of such a framework ”on the run” is therefore acceptable.

2.1. Research setup

This section of the methodology describes the steps that were taken to answer the research questions. The goal of this research is to identify values playing a role in green hydrogen. Two quantitative methods were used for this purpose. These are data analysis and topic modelling. The application of these methods can be seen as a process and a simplified overview of this process is shown in figure 2.1. As seen in the figure, the application of these methods results in an output of values as indicated by these methods. Note that this is not to be confused with quantitative values. Values instead refer to Schwartz’s definition of values mentioned in the introduction.

To find values a case study is employed in this research. This case study applies to the specific set of documents on which quantitative analysis is performed to find values. The case study thus follows the steps of figure 2.1 where the input is left constant. This section expands on the previous figure which results in figure 2.2 and explains each step. The first observation in this figure (figure 2.2 is the presence of two ”rows” or ”flows”. Each row represents one of the quantitative methods. The setup of each quantitative method is now described. This row is thus referred to as the research setup which is part of the case study.

2.1.1. Data analysis research

The first quantitative method indicated by the first row in figure 2.2 is the data analysis. It plays a minor role in this research and is utilised only in the first section of the results. This quantitative method

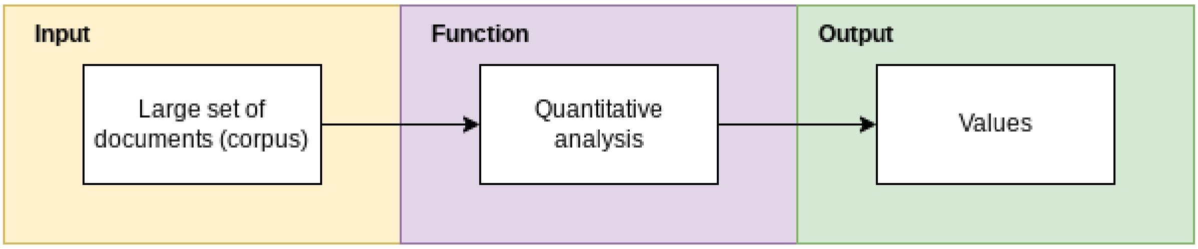


Figure 2.1: Simplified overview of the research methodology. Note that it is in the form of f(x) = y. In this research, the main method topic modelling is used to process a large set of documents to find values. This figure implies several important aspects related to this research. First is that there is information in the documents. Second that this information can be processed by a method that can detect values. Third that these outputs, our values (not to be confused with quantitative values) are a proper reflection of the input.

produces value frequencies. It is performed in section 3.1 and its implications are mentioned in 4.2.3 It answers the research objective by providing an answer to how frequently each value occurs. Value frequencies were found by counting the unique occurrences of each value divided by the total number of documents. A unique occurrence takes on when a value is present in a document. Therefore the first quantitative method is summed up by the following equations:

$$y_v = \frac{\sum u^l}{l} \tag{2.1}$$

$$u_v^l = \begin{Bmatrix} 1 \text{ if } V_v \in x^l \\ 0 \text{ else} \end{Bmatrix} \tag{2.2}$$

Where y_v is the value frequency for a specific value indicated by the index v. V_v corresponds to a specific value in the set of all values V. l corresponds to the index of a document. x^l is a specific document from the set of all documents X.

2.1.2. Topic modelling research

The second quantitative method indicated by the second row in figure 2.2 is topic modelling. Topic modelling is a tool for classifying large sets of documents into a specified number of topics. Each document is specified into a topic with a specific degree of certainty which is called the θ matrix. Several topic modelling techniques exist and in this research one specific method was used with the name Correlation explanation or simply CorEx. A topic model requires the processing of words into numbers. Several different tools exist for this purpose. This process is called word vectorization. The word vectorization ”scheme” refers to the specific tool used for this purpose. Topic modelling is described in more detail in section 2.3. This section provides a list of the inputs used in this model among other things.

In this research, CorEx is used because it allows steering of the model results through ”anchoring”. A topic model ”learns” which words belong to what topic. This is a distinct property from the informativeness that belongs to each word. That informativeness is predefined by the word vectorization preceding the topic modelling. Anchoring binds a group of words together in one topic overriding model learning for that group of words to that topic. The binding strength through anchoring is predefined to be three times stronger than what is possible through regular learning by the model. By anchoring groups of words that correspond to a value, it is possible to identify these values in groups of documents.

Topic modelling and the main research question

Topic modelling is used to find values in a group of documents given a predefined group of values. Answering the research objective is not straightforward because of three considerations. The first consideration is how the lists of values are created that serve as anchors of the topic model. These anchors determine what one is looking for and therefore directly affect the results of the research. Multiple anchor sets are used since it is given that multiple sets of values are relevant to policy decision-making (see introduction). What the anchors are and how the anchors are constructed corresponds to the ”values” in 2.2. These values are represented by each section in the results corresponding to corporate values (called core values), the topics generated by the topic model itself (hydrogen context) and generic values (general representation of society). The identification of hydrogen context is incorporated in the second sub-question. As a whole, this should be seen as the variation of values ultimately resulting in the answering of the research objective and an illustration of the variety in values 4.2. All in all multiple anchor sets are used in the results because 1) anchors are a reflection of values and 2) values represent the interest (of a group) of people. 3) Since it is assumed that policy-making occurs in a multi-actor system, 4) finding values has to occur with some consideration of value plurality.

The second consideration is how one ensures that these lists of values can be found given the corpus and word processing methodology. This consideration is incorporated in the first sub-question. The answer to this consideration is found in the second and third steps of the resulting framework.

What methodology is used for interpreting the results of the topic model back into words is the third and last consideration. Results of the topic model are in the form of large matrices containing the probability that a word (α) and document (θ) is a member of a topic. The third subquestion of this research provides several answers and examples for doing so.

Allegory of the cave

Topic modelling can be explained by using the allegory of the cave. Imagine the topic model to be the prisoner. In this allegory, objects pass along but the prisoner cannot see them. Only the shadows of these objects reflect on the cave wall which the prisoner uses to infer what these objects are. Objects are similar to n-grams (the object resulting from processing a word into a number). The processing of text data into numbers is similar to the shadow of these objects. While the prisoner has a similar frame of reference as the prisoner, a human stands outside the cave and has the ability to understand these objects. This extract is shown on the cover of this report.

2.2. Case study

The aforementioned research setups are applied to a case and this section explains the steps performed to obtain the results. Using figure 2.2 as a guide, research setups corresponding to the rows, are reflected by the purple boxes. The yellow input box, the case study, involves varying inputs corresponding to the boxes in the process. This section explains each step in this figure. Scientific literature on green hydrogen is used as the subject of the case study. Some background information is given on this theme in appendix D.

2.2.1. Text data input (Corpus)

As an input, sets (groups or collections) of documents are used and this collection is called the corpus. This corpus was made by extracting the scientific articles that were found by using the search term ”Hydrogen” on Scopus. Each scientific article has a set of data entries related to it. The most important ones are the title and abstract and these two were the only data entries used in this research. The input contains 63 thousand articles and credit for collecting this database goes to Tristan de Wildt.

2.2.2. Value inputs (Anchors)

Values are used as input in step three. A value (see definitions) can be indicated by a single word (see introduction). This word then implies that this word is desirable behaviour, such as competition, or an end state, such as capability. Each value is accompanied by groups of words that indicate that this value is present. The process of defining corresponds to listing all relevant values and words that indicate each value. The phrasing groups of words means that there is a set of words and each entry can consist of one or of multiple words.

Each entry is called an anchor word. An anchor consists of a group of anchor words and an anchor refers to a value. The set of all the anchors is called the anchor set. In this research, the anchor set represents the values that are relevant to an individual or group. Anchor words are not to be confused with keywords. Anchors relate to input and keywords to output. A keyword is a word or group of words that are informative to a topic. Analyses of the α signal (the machine learned the weight of each term and the informativeness conveyed by each term) show that every term contains some information about every topic, therefore every term is a keyword to every topic. Nevertheless, keywords are useful because they are in a hierarchy.

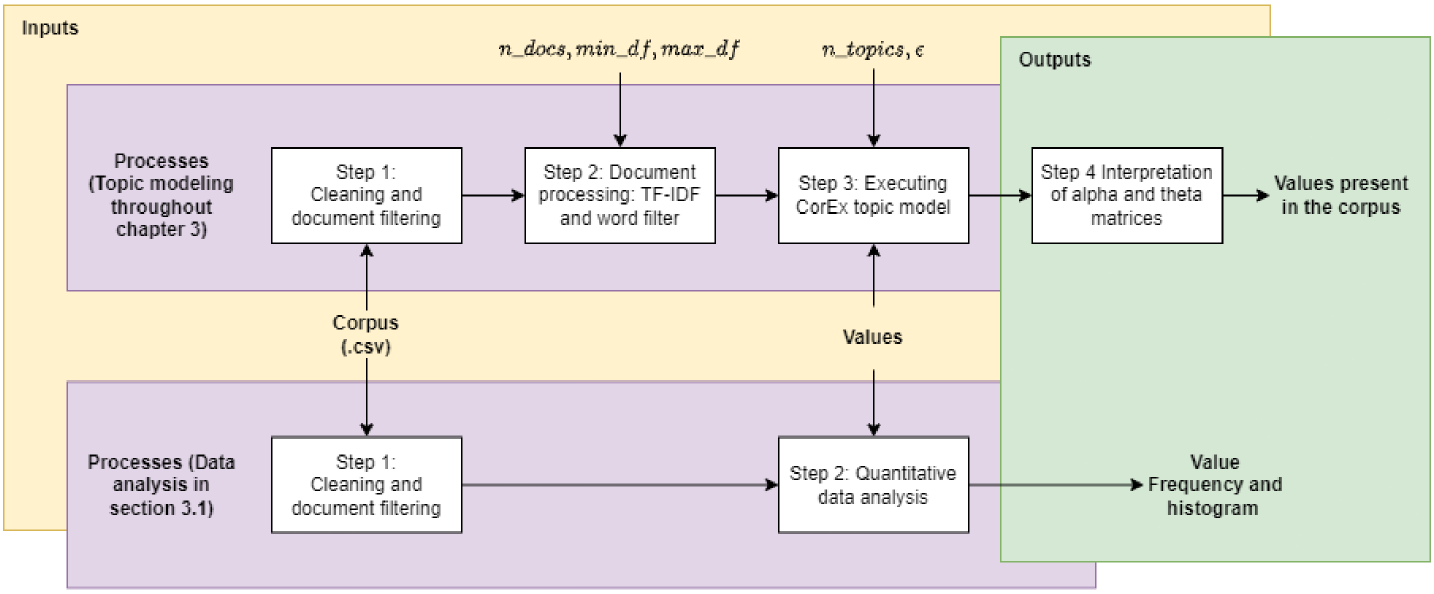

Inputs
n_docs, min_df, max_df
n_topics, ϵ
Outputs
Processes (Topic modeling throughout chapter 3)
Step 1: Cleaning and document filtering
Step 2: Document processing: TF-IDF and word filter
Step 3: Executing CorEx topic model
Step 4 Interpretation of alpha and theta matrices
Values present in the corpus
Corpus (.csv)
Values
Processes (Data analysis in section 3.1)
Step 1: Cleaning and document filtering
Step 2: Quantitative data analysis
Value Frequency and histogram

Figure 2.2: Overview of research methods of data analysis and topic modelling. The first step involves cleaning and filtering the documents. In topic modelling the second step is word filtering proceeded by vectorization of the corpus using TF-IDF (term frequency inverse document frequency) available in python's Scipy package. This vectorized data can be used to construct topics using CorEx.

2.2.3. Step 1: Text data processing

Text data processing is indicated by step one in the methodology figure. Here the abstracts, titles and keywords from these articles were filtered for punctuation, upper-cases, nouns, articles and other non-semantic symbol sets to obtain input words x_i in the topic model. Commonly used words that ”clutter” the input were filtered using a database commonly used for text data mining purposes. This database is the NLTK stopword list which can be imported into python where the text data processing occurred.

2.2.4. Step 2: Word vectorization

While the top row refers to the data analysis (see research setup), the bottom row continues with an intermediary step namely document processing. In this step, four operations are performed over the documents. These processes are indicated in the purple (indicated by TF-IDF) part of figure 2.3. First, a sample is taken from the documents. This sample is of the size n_docs and selected at random.

Words are afterwards transformed into n-grams. An n-gram is a group of tokens of size n that are in the same order as the document they originate from. A token is a group of symbols in a document that are separated from each other by a blank space. N-grams are referred to by their number n, such as unigrams, digrams and trigrams. Unigrams are best understood as all the individual words in a document. Digrams as unigrams plus the sets of two words that occur next to each other, etc. for higher n-grams.

After n-grammisation the resulting n-grams (no longer words) are filtered by a bandpass filter. A bandpass frequency filters out everything that has either a too low or high frequency only leaving the values in the middle. The variable min_df is the lower and max_df is the highest end of the filter. The values provided can be absolute numbers (number of unique occurrences in documents) or frequencies (occurring in a certain percentage of the documents). This filter improves the performance (speed) of the model.

Lastly, the n-grams are vectorized using a (word) vectorization scheme. In this research, the term frequency inverse document frequency (TF-IDF) is used. This algorithm attributes a value between zero and one (the unit interval) to each n-gram. It does so by taking the term frequency (TF) and inverse document frequency (IDF) into account. The more often a term occurs in a document (TF), the higher the number. The more often a word occurs in different documents the lower the value. Therefore terms that are used in specific disciplines (high TF, low document frequency) have a high value. The IDF works as a counterweight. Terms typically occurring in many documents will have a high document frequency and therefore the resulting (TF multiplied by IDF) ”should” reduce the value of this term close to zero.

2.2.5. Step 3: Quantitative data analysis

The quantitative analysis is performed after the text processing and reports the role of values in the corpus. What hides behind these results are assumptions. The most important dependent input variable that reflects these assumptions is the value set in data analysis. In data analysis, a set of values and its associated keywords are in its entirety fed into the analysis.

2.2.6. Step 3: Running topic models

Topic modelling occurs after word vectorisation, unlike the quantitative analysis which was performed after text processing. In topic modelling values are fed to the model as anchors and the results are in the form of large matrices where each entry is filled with a probability. These are called α and θ matrices explained in section 2.4. These matrices require interpretation in order to create a tangible result that can be communicated to others.

The number of topics are after anchors the most important dependent variable as it determines topic resolution (how detailed the topics are). Other input variables are less relevant (in this study) but shortly highlighted. A minimum number of documents are required to generate representative topics. This value depends on the number of topics for a specific corpus. Finding this minimum value of n-docs for which the model yields stable results is therefore complex and out of scope. Finding a stability criterion is not urgent unless topic modelling is performed in an automated environment. Filter bandwidth primarily affects performance and as with the word filter scheme, it doesn't seem to affect the results whatsoever. Lastly is the document filtering scheme which was slightly varied, but not experimented with. In the end, the same corpus is used throughout chapter 3. What can be said about changing the corpus is that it affects both term and document frequencies, making a document filter arguably more important than a word filter.

2.2.7. Step 4: Interpretation of alpha and theta matrices

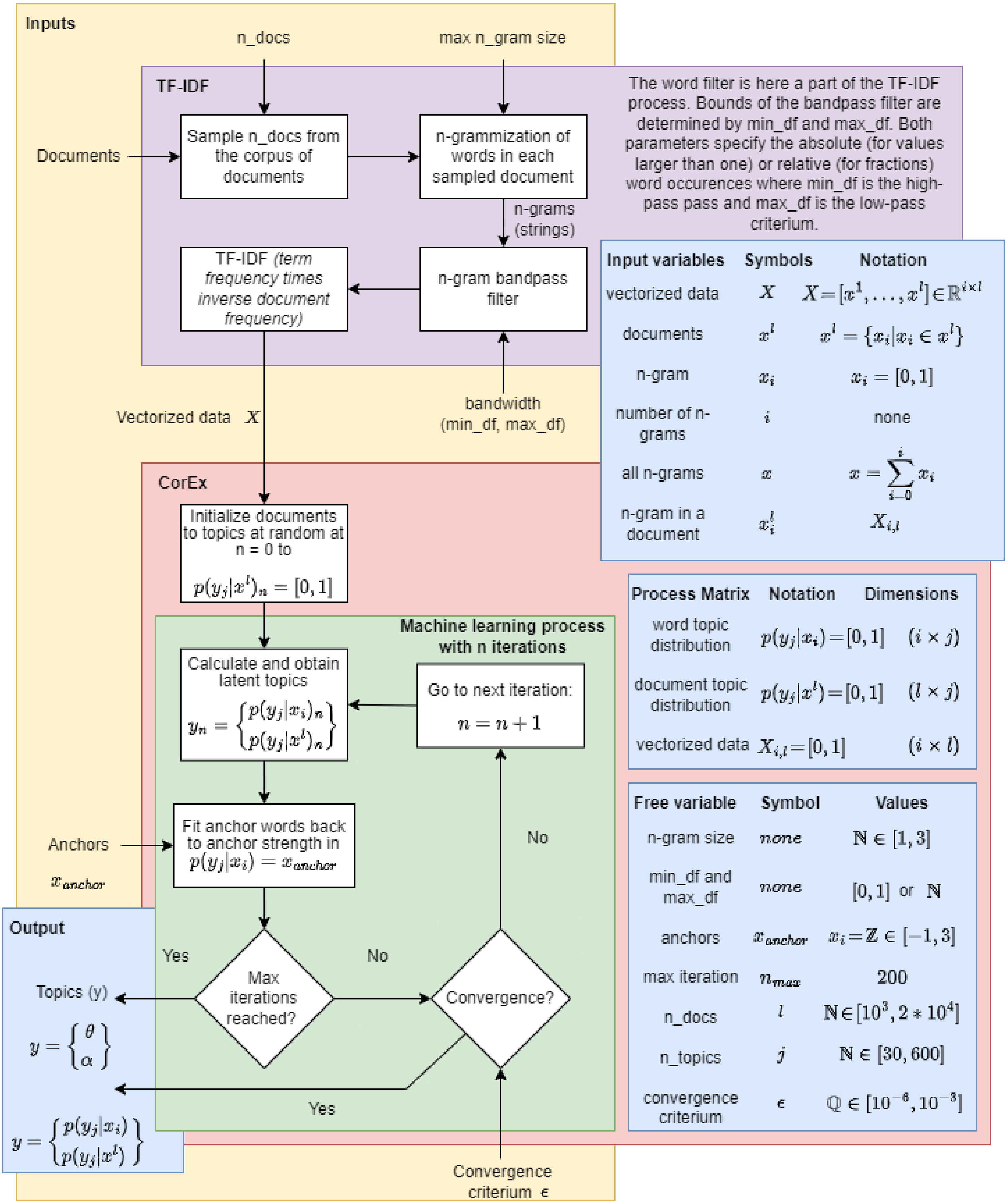
Interpreting the topic model results is required to create a result that can be communicated to others. The first step in this interpretation is analysing the α matrix. This contains the n-gram topic distributions. In this analysis, topics are selected and forwarded to the next ”round” if ”relevant” terms (n-grams) are members of that topic. Anchor sets make up all the anchors, where a single set communicates a single value. A relevant term is part of any of the anchors or relates to or communicates the same idea. This first analysis round can be done manually or automatically. Manual inspection is done by human judgement based on some formalised criteria. Automatic inspection is done by use of an algorithm and selects any topic if the information contribution of any set of relevant terms compared to all other terms to that topic is higher than a predefined criterion. A set of relevant terms correspond to a single set of anchors which in turn corresponds to a single value. A subset of topics is obtained after this first round of interpretation. This is followed by the interpretation of the θ matrix. This contains the documents that are members of each topic. These documents are read manually to determine if and how it relates to the corresponding value.

2.3. Topic modelling using Correlation Explanation (CorEx)

Topic modelling is used for the classification of sets of documents in topics. Unsupervised machine learning models are useful for large data sets with an unknown number of relevant subjects and only allow the specification of the number of topics. There is next to this no other influence on the result apart from selecting and processing the documents used as input. A semi-supervised topic model gives more control over the process by allowing the usage of anchors. Using anchors corresponds to binding a group of words to a specific topic. In this research, the method correlation explanation (CorEx) is used. An implementation of the CorEx model is available in python (Jupyter notebook) (Gallagher et al., 2017). Its specifications are described in section A.1 and a more thorough overview, expanding on figure 2.2 is shown in figure 2.3.

2.3.1. Why is CorEx used

CorEx stands out because it allows the manipulation of keywords namely filtering, anchoring and including keywords. This labelling of keywords determines their output value which is a form of supervision.


Inputs
n_docs
max n_gram size
TF-IDF
Documents
Sample n_docs from the corpus of documents
n-grammization of words in each sampled document
n-grams (strings)
TF-IDF *(term frequency times inverse document frequency)*
n-gram bandpass filter
Vectorized data X
bandwidth (min_df, max_df)
CorEx
Initialize documents to topics at random at n = 0 to $p(y_j|x^l)_n = [0,1]$
Machine learning process with n iterations
Calculate and obtain latent topics $y_n = \begin{Bmatrix} p(y_j|x_i)_n \\ p(y_j|x^l)_n \end{Bmatrix}$
Go to next iteration: $n = n + 1$
Anchors x_{anchor}
Fit anchor words back to anchor strength in $p(y_j|x_i) = x_{anchor}$
No
Output
Yes
No
Topics (y)
$y = \begin{Bmatrix} \theta \\ \alpha \end{Bmatrix}$
Max iterations reached?
Convergence?
Yes
$y = \begin{Bmatrix} p(y_j|x_i) \\ p(y_j|x^l) \end{Bmatrix}$
Convergence criterium ϵ

The word filter is here a part of the TF-IDF process. Bounds of the bandpass filter are determined by min_df and max_df. Both parameters specify the absolute (for values larger than one) or relative (for fractions) word occurences where min_df is the high-pass pass and max_df is the low-pass criterium.

Input variables	Symbols	Notation
vectorized data	X	$X = [x^1, \dots, x^l] \in \mathbb{R}^{i \times l}$
documents	x^l	$x^l = \{x_i \mid x_i \in x^l\}$
n-gram	x_i	$x_i = [0,1]$
number of n-grams	i	none
all n-grams	x	$x = \sum_{i=0}^{i} x_i$
n-gram in a document	x_i^l	$X_{i,l}$

Process Matrix	Notation	Dimensions
word topic distribution	$p(y_j\|x_i) = [0,1]$	$(i \times j)$
document topic distribution	$p(y_j\|x^l) = [0,1]$	$(l \times j)$
vectorized data	$X_{i,l} = [0,1]$	$(i \times l)$

Free variable	Symbol	Values
n-gram size	$none$	$\mathbb{N} \in [1,3]$
min_df and max_df	$none$	$[0,1]$ or $\mathbb{N}$
anchors	x_{anchor}	$x_i = \mathbb{Z} \in [-1,3]$
max iteration	n_{max}	200
n_docs	l	$\mathbb{N} \in [10^3, 2 * 10^4]$
n_topics	j	$\mathbb{N} \in [30,600]$
convergence criterium	ϵ	$\mathbb{Q} \in [10^{-6}, 10^{-3}]$

Figure 2.3: Schematic working of the CorEx topic model performed during each model run. Note that α and θ correspond to words and documents respectively. These symbols should not be confused with the python code where θ corresponds to $P\_y\_given\_x$ and α to the word topic distribution.

Since some and not all inputs are labelled CorEx can be utilised as a semi-supervised topic model. Filtering removes a word from the corpus and anchoring sets the n-gram to a predefined topic. Various dependent variables exist in CorEx for influencing the results. The most important variables in topic modelling (which can be said for other topic modelling algorithms too) are the corpus and number of topics. The corpus consisting of the documents and how they are processed determine what can be found by the model. The number of topics is the number of compartments in which the documents are allowed to be classified. Therefore the number of topics defines the level of detail that is present in the result.

2.3.2. Topic model mechanism

The mechanism of the topic model is illustrated in figure 2.3. A vectorized dataset is used as input for the topic model. It was explained at the beginning of this chapter how this is obtained. During initialisation, documents are attributed to topics at random. The degree they are attributed to a topic follows the unit interval. A document can in principle be a member of all or no topic. After initialisation, the model enters a loop. This loop starts calculating and obtaining the latent topics. At this step, both the word and document membership of each topic is calculated. These are called the α and θ matrices respectively. Since this adjusts the anchors, the next step involves resetting the anchored n-grams back to their anchored value. This iteration is repeated until one of two conditions is reached. Either the maximum number of iterations is reached (by default 200) or the model converges. Convergence occurs when the change in total correlation (some degree of informativeness of the entire system) is smaller than a certain condition. This is influenced by the convergence criterion, ϵ. If one of two conditions is reached the latent topics (α and θ matrices) are reported.

2.4. Important concepts in CorEx

The input for a topic modelling is called the corpus. The corpus corresponds to a set of documents. When a topic model is performed, a topic modelling exercise, not the entire corpus but a sample of the corpus is taken. After finishing the exercise the results of the sample are interpolated over the corpus. A single document, in our case scientific articles pertaining to hydrogen, contains data and meta-data. Data is the abstract of the article, while title, authors, keywords, year, publisher, etc. are meta-data. All sorts of meta-data can be used by the topic model. Data used by the topic model are numbers and not words.

These numbers are obtained by creating n-grams, filtering the n-grams and using a TF-IDF (tfidf) scheme to represent the n-grams as numbers. An n-gram is a set of symbols (typically words) divided by a number of punctuation spaces of size n that occur next to each other in the text. Thus 1-grams, unigrams, are individual words and 2-grams, digrams, are sets of two words. Next, a filter is introduced since too high occurring n-grams are not useful and too low occurring n-grams are clogging model performance. After filtering the n-grams are vectorized by a tfidf scheme, which stands for term frequency times inverse document frequency. For the height of the number that is allocated to the n-gram; term frequency means that it is beneficial to occur more in a single document, but inverse document frequency means that occurring in a multitude of documents is not.

The documents with vectorized n-grams are fed into the CorEx model. The model randomly attributes topic membership to all documents at initialisation. This distribution is called θ or $p(y|x)$, the chance a document belongs to a topic. With this data, four intermediate parameters can be calculated, which are collectively called θ in CorEx. Since it is non-canonical they will be called ”four intermediates” and remember: θ is the word topic distribution. Given the θ distribution and the vectorized data, one can calculate mutual information of n-grams that signal topic membership. The distribution of n-grams among topics is called α. This distribution is combined with total correlation to execute one learning step. In this learning step the model tries to maximise total correlation. In doing so it introduces a weight to each n-gram belonging to each topic. This matrix is called α in CorEx but is not canonical and is therefore called weight, weight matrix or n-gram weights. Weights allow CorEx to be executed in a semi-supervised way by giving custom weights to specific n-grams. After obtaining α latent topics can be calculated by using ”the four intermediates” and α that also yields a θ variable. The initialised θ variable is now updated and the entire process is repeated (called an iteration) until two conditions are met. Either the maximum number of iterations is reached or the change in total correlation is sufficiently small to break the loop. What results are latent topics, which consist of a final α and θ matrix of size

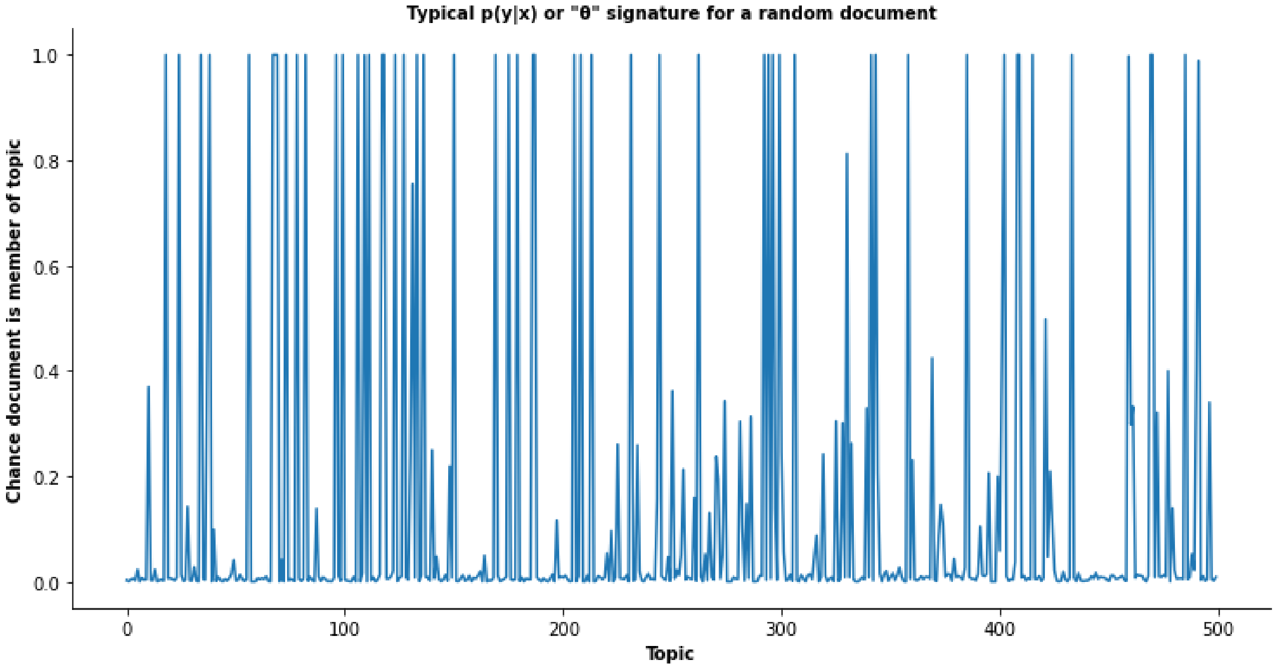


Figure 2.4: Typical ”θ”$p(y \mid x)$ signature showing, for a given document, the chance it is a member of a document. A document was chosen to represent this distribution. Note that θ is a matrix and this distribution exists for every document.

(topics x n-grams) and (documents x topics) respectively.

2.4.1. N-gram topic distribution θ

Figure 2.4 shows the document topic distribution. In this figure, a random document is chosen. For this document, the figure shows the chance that it is a member of the topic indicated on the x-axis. Values of theta range between zero and one and are typically close to zero or close to one. The figure illustrates that a document is typically a member of tens of topics. θ is used in other topic modelling techniques to indicate the distribution of documents (x) among topics (y) or $p(y|x)$. In CorEx this is called $p(y|x)$ and not θ. Within CorEx θ instead refers to four ”marginal” parameters. Every time θ is mentioned it is referred to as $p(y|x)$.

2.4.2. Document topic distribution α

Figure 2.5 shows the final document topic distributions from a topic modelling exercise. The n-gram topic distribution signature is divided into a mutual information and weight signature. Weight and mutual information are closely related to each other but differ in a few important regards. Mutual information represents the information of an n-gram to a topic and is based on the tfidf adjusted by the document topic memberships learned so far throughout the model. Mutual information can be seen as the data that indicates topic membership. Weight represents the ”learning” or ”optimising” of the model which seeks to maximise total correlation. Mutual information is the primary source of information for this learning. The weight of anchored n-grams is fixed in every iteration. Weight is introduced by human (through anchor) and machine (learning) to change the fit of documents to latent topics.

In both weight and mutual information many zero values are observed, while weight takes on values of either zero or one, mutual information is more gently distributed. To elaborate on this first weight and then mutual information is described. Looking at the figure shows that the weight of n-grams typically takes on three values. Normal values are fitted between 0 and 1 and tend to approach these values. We assume that it is either zero or one. Anchored n-grams form the exception and take on the ”anchor strength” of three. The majority of n-grams take on a value close to zero. Out of all the n-grams (close to fifty thousand), the topic with the highest number of n-grams (weight is 1 or higher) is 716 and the lowest is 2.

Unlike alpha, mutual information is more gently distributed, hence many n-grams can alter the re-

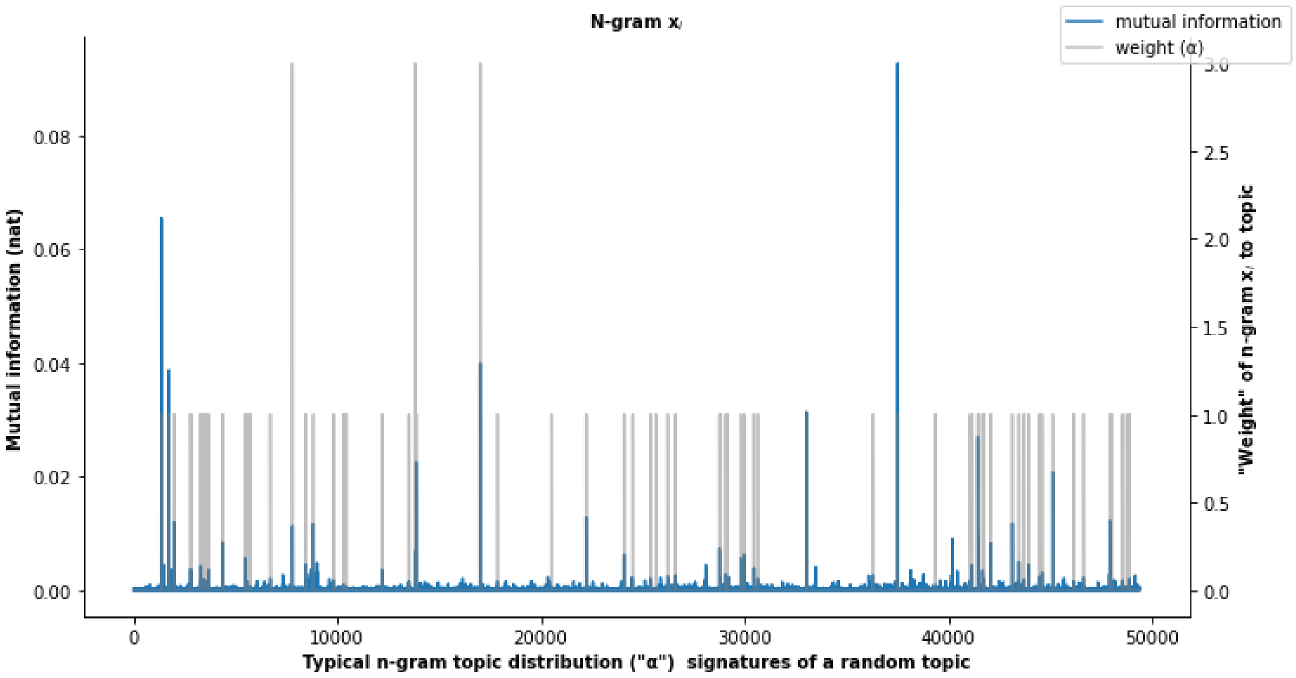


Figure 2.5: Typical α signature showing both the weight and the mutual information signatures of a random topic. Nat is the unit of information corresponding to a chance of $1/e$. Note the anchor strength of three giving n-grams a weight higher than one. A random topic was chosen to represent this distribution. Note that α is a matrix and this distribution exists for every topic.

sults (latent topics). The distribution of mutual information is nevertheless still dauntingly harsh. Around fifty percent of all the n-grams take on a value of zero (typically between 25-95%) depending on the topic. Many values are close to zero, but as the figure indicates there is still a significant portion of n-grams that have a reasonable amount of informativeness. Let's take 0.01 nat of informativeness as a decent amount of information (this corresponds to $1/e$% of the topic occurring when the n-grams are observed). Typically dozens of n-grams have more than 0.01 nat informativeness. Every time this value is scaled down by a factor of ten the number of n-grams is increased by a factor of ten too. It should be observed that some topics consistently have a low number of n-grams that persist during this operation. In these cases, we speak of zero, several or tens of n-grams.

These observations indicate three things. Firstly, some topics have a negligible amount of mutual informativeness. This could indicate that topics are fully disentangled meaning that adding more topics does not increase total correlation. Furthermore, all information is already contained in other topics explaining the low levels of mutual informativeness of these topics. Secondly, the higher number of zeros in alpha relative to mutual information indicates that other distributions could be possible, but doesn't happen however because of the total correlation maximisation criterion. This pattern in weight reflects the ability of the model to use all the data available to it. This optimisation criterion could furthermore be responsible for the harsh distribution of weight, while it could also be caused by the corpus. Lastly, weight is clearly a product of interpretation, while mutual information is closer to the original data (tfidf) and the informativeness of the n-gram topic distribution in the most recent iteration.

Alpha is used in canonical topic modelling to indicate the n-gram topic distribution. Both variables say something about this distribution. Weight is specific for CorEx however, and values higher than 1 for anchored n-grams are generally not accepted (because it implies a 300% chance). The mutual information parameter conveys the informativeness of n-grams across topics in a way that can be used for other purposes. Therefore alpha, which indicates weight in CorEx does not correspond to the n-gram topic distribution, but mutual information does.

2.4.3. Total correlation

Maximisation of total correlation (TC) is CorEx's optimisation objective. TC is a complex calculation using the tfidf and θ as inputs. TC is updated each iteration until the optimisation criteria are reached. Figure 2.6 shows the typical oscillatory behaviour of TC. This oscillation is apparently created when

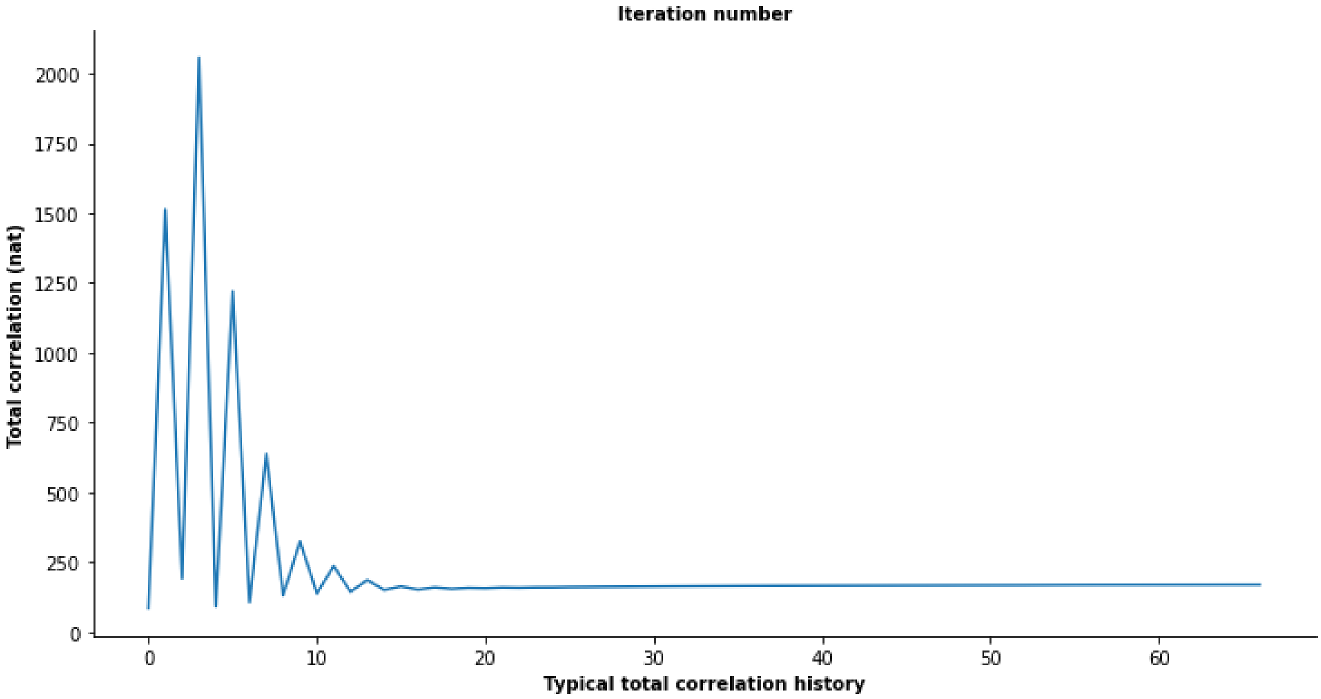


Figure 2.6: Typical development of the total correlation over given the number of iterations. Total correlation is reported in nat.

TC is updated each iteration. Oscillation continues indefinitely and is a consequence of θ because tfidf remains static throughout the run. The optimisation criterion is reached if a maximum number of runs is reached or the difference between runs is small enough. This latter calculation is problematic due to its design and perpetual oscillation. First, it takes the last five runs and the five runs before that. The difference between the mean of both sets should be smaller than a constant, called ϵ. The problem is that an odd number, five, was chosen. Because of this it effectively calculates the oscillation height at the end of the run divided by five. Consequently, the loop will never break if epsilon is small enough.

3 Results

Topic models were used for finding values in green hydrogen. Latent topics, the results of topic models, can show these values. Anchors and model parameters are varied to influence these results. The central theme in the result section is the selection of anchors. Most attention will be paid to anchors because they are the most important variable influencing the model's results. The way in which anchors are constructed seeps through the model into the results. Anchors follow a rationale or a set of assumptions. A description of this ensures that anchors fit in some artificial boundaries. An anchor set is a group of anchors that can collectively be described by the same overarching idea, set of assumptions or concept that are used in (one specific) topic modelling exercise(s). Three topic modelling exercises are performed each with a distinct way of dictating how these anchors are constructed.

More specifically, assumptions of the anchor set determine which anchor words are included in each anchor. Anchor words are n-grams and ideally independent from other anchors' anchor words. Anchors are comprised of a set of anchor words and fit in some overarching idea, set of assumptions or concept. In practice, anchors reflect ”values” (documents in topic x describe value x) or ”context” (in which context in hydrogen is described). The construction of the anchor words depends on some set of assumptions. Each section in this chapter reflects these assumptions leading to different results. In the first section anchors are created using a large dataset, ”core values”. In the next section, the anchors were selected based on fitting existing topics and their informativeness and lastly, a dataset of anchors common in topic models is used to construct anchors.

3.1. Core values in literature

A dataset called core values was used to identify the values playing a role in the hydrogen corpus. First, data analysis was performed on the corpus to assess the presence of these values. Lastly, this set of values was used as input in CorEx topic modelling using which is described in the next section of topic modelling.

This dataset contains 433 values and their associated keywords which are generally used to identify values in organisations and other applications for the purpose of text processing. The core values are used for identifying and making values explicit in organisations. Corporate culture is ”the basic pattern of shared assumptions, values and beliefs considered to be the correct way of thinking about and acting on problems and opportunities facing the organisation” (du Plessis, 2011). Corporate ideology is the combination of an organisation's strategy and the organisation's culture (du Plessis, 2011). Several authors suggest that organisational values are not only reflecting the corporate culture but the corporate ideology since they are products of management philosophy (du Plessis, 2011). Core values are therefore an indication of ”corporate ideology”.

Since core values are subjected to scientific literature using corporate ideology is sub-optimal and it is preferred to call it organisational ideology. Documents reflective of corporate ideology are ”communicative events”, typically corporate documents such as annual reports or internal documents (Fox, 2006). It is on the other hand justified to subject scientific literature to core values, because this is typically produced in universities and corporate environments, which reflect the ”organisational ideology”.

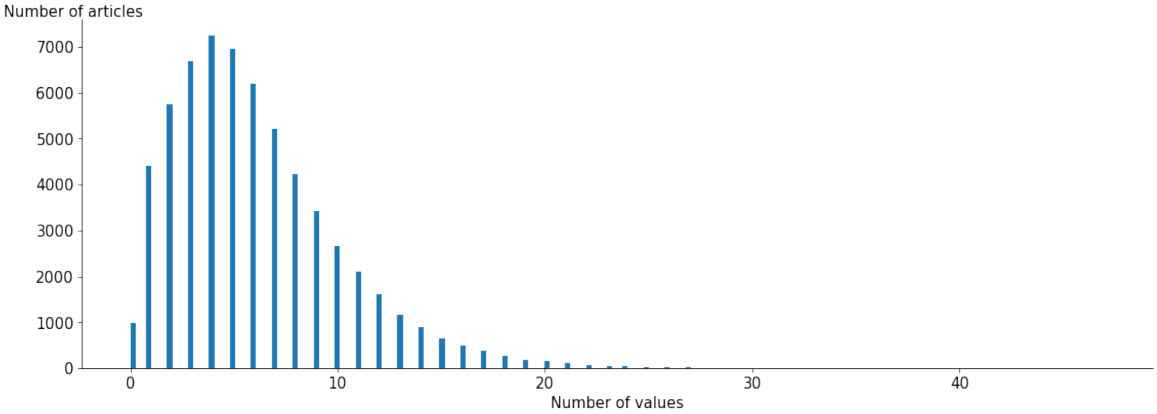


Figure 3.1: Document value histogram. Each bar represents the amount of documents that contain the number of values specified by the x-axis.

3.1.1. Distribution of values among articles

Core values are used to find values in the hydrogen corpus using topic modelling and data-analysis. Data-analysis is discussed first, the first step is creating a value histogram. This illustration was made by counting the number of unique values in each document for each document. A histogram was made from this with the y-axis showing the amount of documents that have a specific number of unique value occurrences indicated by the x-axis in figure 3.1.

Observing the histogram allows a description of the curvature, bins, ”tail” and a description of the relation between dataset and corpus. First is the form of the curve which represents some unknown distribution. The steep curves give the impression of a power law distribution, but the observed bell / hyperbola is absent in a typical power law distribution. The bell curve gives the impression of a truncated normal distribution, which is not the case since the curve is asymmetrical. The curve might conform gamma or lognormal distributions.

Related to this distribution characterisation is the qualification of the tail. It can be classified as a heavy tailed distribution since the tail overextends its expected length (number of values extends up to 50) and thickness (bin count is relatively high at higher number of values) given the expectation based on the observed standard deviation. It is thought that this could reflect writing style.

Lastly is the observation that around a thousand documents contain zero unique values. Scanning these documents shows that these articles are related to short and dense articles in the applied domains such as petrol engineering and chemistry with an abundance of jargon. In this set an article was found on the noise production of hydrogen installations. The dataset used in this context indicates that this article is value neutral, which is not necessarily the case in a context where noise pollution is in consideration where it would affect the quality of life.

This example is a starting point of the interpretation of the graph, which first and foremost shows that this dataset has its limitations and that the inclusion of context is necessary for improving the results. It is thought that the value dataset is biased and that some of the zero value bins, the heavy tail and shape of the curve reflect differences in audience, research objectives and writing style. These three factors would affect choice of include or exclusion of topics, results and certain groups of words.

The added value of this dataset is brought in question. First it is unknown what number of values mean and to what it refers. Secondly it is not thought that a higher number of values is necessarily related to informativeness of that article. Lower number of values clearly indicates that one cannot directly observe values or that values are indeed absent. In some of these cases values can be observed with the use of interpretation or instinct. This was seen as enough reason to filter out articles with zero values in the topic modelling phase.

3.1.2. Frequency of values

Giving meaning to the number of values is done through plotting the frequency of each value. In figure 3.2, the value frequency diagram, it is shown how frequent (y-axis) the 30 most frequent values (x-axis) appear as a ratio of all articles. More values were not plotted, the 30th to 60th most common values have values between 3 and 5 percent and 60th to 90th around 2 percent. A boundary at the 28th

Output	Future	Management
Results	Growth	Order
Impact	Impact	Availability
Performance	Potential	Activity
Stability	Sustainability	Stability
Resolution	Development	Invention
Exploration	Innovation	Firm
Effectiveness	Environment	Structure
Efficiency	Change	Change
Best	Prepared	Variety
		Control

Table 3.1: Classification of the top 28 values. Note that some values are listed in multiple classes.

most common value is identified where a relatively sharp drop signifying the range where all values are appearing at a similar frequency. After this mark frequency steadily decreases and this decrease persists after 90 values. Beyond this point it becomes too large to structure, infer relations or illustrate. Related to figure 3.1, the value histogram, and the slow decrease in the value frequency shows that relatively speaking most documents likely have a few different values apart from the most common ones, some have no values and a few documents more than 10 different values.

Values beyond the 28th most common value are called uncommon values and appear in less than 5% of all the documents. Up to the 90th value these values appear in more than 2% of all the documents and some of the values in this range stand out. Examples of values that stand out are economic viability, cleanliness, resilience, anticipation, local, utility, dependability, power, experience, accountability, diversity, flexibility and spirituality. Even when using a tiny set of documents in the topic model (1000) that these values will appear 20 to 50 times on average. The documents part of this value are expected to communicate aspects on hydrogen that are not arbitrary contrary to first 28 ”general” values. It is therefore thought that a focus on these values is more useful since some do indicate relevant features.

The first 28 values are common since each individual value occurs in 10 percent or more documents. It is interpreted that these values characterise the corpus because of this high frequency. Some context was expected, but the absence of context makes sense since core values is a general too used for reflection of the corpus. The values itself did not contain values that were expected. It is interpreted that the first 28 values can be classified into three groups. This interpretation is shown in table 3.1 on interpreted values. The first reflects ”output”, which corresponds to research output, new implementations or improving production processes. ”The future” is related to anything not in the here and now, specifically future movements of hydrogen processes. Last is ”management”, which could also be labeled as delegation controlling and directing processes. The values and processes probably relate to hydrogen, but this is not clear without structuring the context in the corpus. After performing this task in section 3.2 it becomes clear that processes all these values relate to hydrogen projects, and hydrogen production and storage processes.

3.1.3. Anchoring core values in CorEx

Several topic model runs were performed using the core values as anchors in each run. In these runs the corpus was filtered by removing articles that did not contain any core values. The first run was without anchors and adding of anchors occurred iteratively. This means that the topic model was executed several times and anchors are added depending on the results of previous runs.

Distinct topics were unable to be constructed using core values as anchors. Topic model run had a thousand to ten thousand documents and a hundred to six hundred topics. Increasing the number of topics allows the model to disentangle otherwise aggregated topics. Increasing the number of documents increases the quantity of words part of the vocabulary used to construct the topics and increases the consistency of topics appearing every run. Anchoring of topics using core values did not result in the generation of topics related to these words or values. If topics made sense they instead focused on the context in which articles were discussed. If these anchor words would survive they resembled artifacts in comparison to the other keywords in that topic.

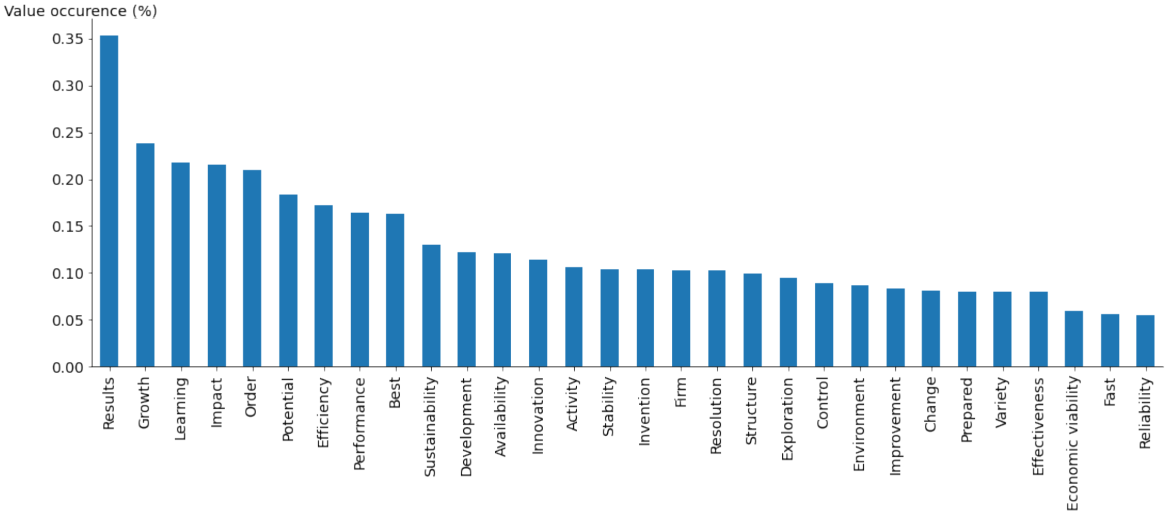


Figure 3.2: Value frequency. Y-axis represents the chance of a value being part of a document. The thirty most common values are shown. This chance could be translated as frequency, hence value frequency.

Context was found to be dominant in the topic modelling process and an attempt was made to counteract this. This was done by filtering all words apart from the words part of the value dataset. This completely removed all context. Topic modelling results now showed a drastic reduction in the amount of topics and keywords per topic although none of the topics made sense. The topics attempt to group keywords together, which more often than not failed as seemingly unrelated words were grouped together. It was hoped that topics would report on one and the same value although this did not happen.

The values and their keywords are seen to be artifacts or noise in the greater picture created by the topic model if these keywords would appear at all. More often than not these anchored values would disappear from the results. In both runs the informativeness of words part of the value dataset, piecewise total correlation (pwtc) is low. However the sum of pwtc, total correlation, (tc) of the model with unfiltered corpus is around 23 nat while the total correlation using only the words in the value dataset is lower than 1 nat. That value keywords are artifacts in the topic model with context indicates that the informativeness of value keywords is very low in this corpus. This means that they don't play an important role in the construction of topics in CorEx. Completely filtering out context is not a solution, therefore some context has to be included. Context is dominating the results of the topic modelling, therefore structuring this context seems to be the next logical step. Therefore the goal in the next section is to identify the context in the hydrogen corpus.

3.1.4. Automatic value identification

An algorithm was developed to identify which anchors are represented in the latent topic. Deciding if an anchor is present in a latent topic goes in two steps, the first step is keyword inspection and the second document inspection. The results of this are shown quantitatively in figure 3.3 and qualitatively in table 3.2. Document inspection is done on a small number of latent topics to determine if an anchor is present in this latent topic. This smaller number of latent topics is an optimisation requirement and obtained through keyword inspection. Keyword inspection is done to filter the anchors that are not present in latent topics. This can be done manually by going through all latent topics and deciding if the keywords show resemblance to the anchor. Doing this is not realistic however, since 433 anchors are used. The advantage of using an algorithm shows in this step as an algorithm is a consistent and efficient way to tackle this problem. The algorithm calculates the mutual information contribution of each anchor relative to all other n-grams (non-anchor words) in each latent topic.

Results are classified in for a run with 450 topics and 3500 documents and shown in table 3.2. Mutual information (mi) contribution for all anchor- latent topic (j) pair are categorised into three classes: $anchor \geq 75\% mi_j$, $50\% mi_j \leq anchor < 75\% mi_j$ or $anchor < 50\% mi_j$. Results contain all anchors with a mutual information contribution higher than 50% a double horizontal line delineates the topics 75% mi contribution with the highest contributing batch on top. Quality of fit varies, around half of the topics show excellent fit with more than 95% of the member documents containing one or more anchor words. This quality is not seen for other topics where for some it even drops below 50%.

The main limitations on this method are number of documents, number of topics and maximum size of n-grams allowed prior to model construction. First the number of topics of 450 is below full disentanglement occurring at 550 topics. Second the number of documents is low, consequently most anchor words were not found leaving 153 out of 433 anchors empty roughly 65%. Lastly the n-gram range gives certain anchors no chance, as some consist only of trigrams and the n-gram range allows unigrams and digrams only.

Anchor	Observation of documents	Interpretation of documents
Best	Discuss optimal conditions, best results and most often ”effectivity” (of some method). A few relate to safety.	Effective methods, best results or safe procedures are not informative to green hydrogen.
Flexibility	Relate to abstract and tangible concepts. In a tangible sense flexibility relates to flexible or elastic materials or objects (reactors). In an abstract sense its use is more varied (e.g. flexible markets).	Topic is is irrelevant for green hydrogen. In an abstract sense it could be useful, but it is not. It could be, because it indicates a resilient- or system property. It is not useful, because it is not tangible. Its use is considered cliché.
Foresight	Small topic using n-gram planning only. Occurs in ”introduction” as motive, or in ”conclusion” relating to activities.	Limited in usefulness as topic, because the anchor corresponds to only one n-gram. Topic is nevertheless informative with excellent anchor topic fit. Motives are a bit informative, tangible action is more informative.
Industry	Documents mention industry or manufacturing. That where things are created. Some articles are selected, because the publisher contains ”industry” in its name	Industry indicates which practices are standard. Industry refers to a collective effort (by companies, universities, etc.). Its tangible and useful application gauges progress relating to, for example, safety of technology and the ability to respond to public concerns. Industry is authoritative. It signals changes in standards due to innovation. However, if it is mentioned in an abstract and non-tangible context, industry is less useful. It becomes responsible for explaining behaviour, change, accepted norms and standards obtaining a mystic and godlike role.
Wonder	Small topic on ”curiosity” In intro- mid-section introducing the research. More often than not mentioned in conclusion where it signals research gaps. Can signal uncertainty or gaps in knowledge	Topic’s usefulness is limited as comprises of only one anchor word. Topic shows excellent fit and its document uncover both motives, uncertainties, and knowledge gaps. This use is somewhat limited, given the age of the documents and the limited number of articles found in this topic.
Accessibility and Availability	Occurs as signalling word as noun. It can relates to availability of resources, tools or services. Availability refers to opportunities, but is used to indicate a lack of something indicating challenges.	Topic has a good fit, but limited in that it uses a single n-gram. Overlaps with industry. Useful for expanding industry as it can indicate opportunities and challenges in this field.
Control	Its n-grams relate to control and operate. It selects documents that talk about control systems, control of substances, the constrains chemical reactions, or where reactions take place. Control sometimes refers to a control sample or control experiment.	Irrelevant to green hydrogen. Documents are relatively technical.

Development	Talks about economic investments, insights for future developments, development of materials and solutions. Development can refer to tools and techniques or something that was done in that research.	Green hydrogen is well represented in this topic. This topic is used to identify development requirements. This can be related to industry as documents are identifying and giving an answer to these challenges.
Effectiveness	N-grams all relate to ’effectiveness’. Contains less n-grams than ”best”, namely only the ones related to effectiveness. Consequently has a smaller amount of content.	See ’best’, but than considering only the part on effectivity. Irrelevant to green hydrogen.
Exploration	Made up of the n-grams related to ’exploring’ and ’research’. Articles on exploring indicate future (explorative) research and research requirements. Research more often than not indicates what this research has done.	Irrelevant to green hydrogen.
Fast	Relates to speed of chemical operations or reactor types. While it relates to speed, it does not address acceleration.	Speed is irrelevant to green hydrogen. Acceleration could be somewhat relevant, but is not assessed.
Global	Usually an adjective referring to global problems, features, solutions, systems. Examples are global warming, oil production, etc. More often than not used in the hydrogen context of renewable energies. Topic has a moderate fit, two third of articles contain keywords. Global in some articles refer to global in a modelling context. Here it means ”overall” instead of ”at one time-step” (local)	Relevant articles describe green hydrogen in a high-level perspective, such as holistic or meta- level. Quoting (Blanchette, 2008) illustrates this best: ”hydrogen is only a metaphor; any change from the current oil economy will entail dramatic changes to the global status quo that must be planned for now”. Other examples are impact of global resource demand (platinum) for the hydrogen economy, viability of hydrogen fuel in the transportation section. Global is more frequently used to indicate the motivation for the research. When this is done the it sometimes signals that the research content refers to general or global processes related to renewable energy systems. Many articles are unrelated to hydrogen and only a small fraction (less than 5% describe global or holistic processes). This means that the described documents are outliers.
Growth	Topic selects documents containing n-grams increase, development and growth. Since these n-grams are general it is no surprise that there is no pattern and articles show to discuss diverse contents. Broad topic with large number of member documents.	Growth in green hydrogen refers to technical developments.
Impact	A broad topic with large number of member documents. This topic selects documents containing n-grams effect, affect, consequence and impact. Since these n-grams are general it is no surprise that there is no pattern and articles are typically technical in diverse contents. Example of titles are ”The rate of an exchange reaction of hydrogen and deuterium in a Mg2Ni bed” and ”Effect of low salinity waterflooding on the chemistry of the produced crude oil”	Impact can refer to any discipline. N-grams are problematic, because it selects articles with these n-grams out of context. Example is the use of impact, effect and consequence in the context ”effect/consequence of”. These articles sometimes indicate a cause instead of discussing the impact.

Innovation and invention	Relies on the n-gram design and in a few cases (≤ 5%) document membership is based on the n-gram innovation. Design and innovation are general terms and occur in technical articles that innovate or design something.	Irrelevant to green hydrogen
Potential	Large and unspecific topic (roughly 20% of the corpus). Contains three n-grams used in different contexts, namely ’potential’, ’possible’ and ’likely’. The n-grams ’potential’ and ’possible’ are both used to indicate some instructions to obtain a specific result, such as (Gopalan & Tyagi, 2020) or to indicate something appears to be possible (potential use). Potential causes dilution as it refers to chemical or electrical potential. Likely indicates something with high certainty, such as ”likely candidates” and ”likely be responsible”.	The value can be expressed in any discipline. If hydrogen is discussed then it is unlikely that it regards a potential information source on relevant future developments. Consider timely articles, discussing potential relevant potential developments that would be relevant. Not only were they not found, if they would be found they will comprise a tiny fraction of the member documents (less than 5%); therefore this value is irrelevant to green hydrogen.
Recognition	Small topic that selects the n-gram identification only. Identification is frequently used in technical articles, such as ”The current model enables identification of conditions” and ”Identification of less-volatile products”. An exception to this is (Collantes, 2008) describing the identification of the main policy issues in the debate on hydrogen fuel as transportation fuel.	Topic does not refer to recognition, but to identification of some objects in technical subjects. Its relevancy to hydrogen exists when it is used in combination with policy identification.
Reliability	Consists of n-grams on reliability (40% of documents), viability (20%), availability (15%) and robustness (10%). When the n-gram reliability is used it often occurs in a context where the value on reliability is justified. Examples of this are, ”to confirm the reliability ... ” and ”new separators were also developed that suggest improved cell reliability”. Viability is used to indicate that a certain method is viable. This is done in text through, for example, ”economic viability”, ”viable route for production” and ”viability of x”. Availability is used in or technical terms, such as ”thermodynamical availability” or to indicate that something is present ”availability of feedstocks”. Robust is used in various contexts ”robust bioconversion”, ”robust measurement”, ”robust catalyst” and ”plan is deemed robust”.	Quality of this anchor depends on the n-gram considered. It is absent in availability. N-grams on reliability explicitly state how this value was achieved indicating that this value plays a major role in these articles. Viability often corroborates reliability. Viability is useful for hydrogen and relate more to potential innovations (see those previous values). Viability is not related to reliability, because it does not say that something is possible or reliable however. Availability does not relate to reliability. Robust is unspecific and rarely relates to a judgement on reliability. Comparing robust to reliable, robust is arguably more complex, but articles fail to specify why. One has to see robust as a given system property, while articles on reliability dedicate their article to this property. It is clear that robustness is not useful, but a loaded term. This is because robustness is used to signal positive characteristics of an object without specifying it, defining it or explaining why.

Success	Uses n-gram 'success' and indicates when 'success' is achieved. Examples are: "The success of", "Success in this task primarily depends on" and "has a profound affect on the success of". This topics describes constrains of success. Other uses are "was tested with success for" and "is considered a complete success" indicating that favourable results were obtained.	In the first example success is used to indicate limitations and motivation for research. It indicates in detail what constrains success or when success is obtained, such as "in particular in term of manufacturing or handling intense heat loads" and "One of the key figures for the success of proton exchange membrane fuel cells (PEMFCs) in automotive applications is lifetime." Examples of the latter case, when success is reached are, for example, "The Pr2NiO4/SnO2 heterojunction with a mass ratio equal to unity was tested with success for the hydrogen production under visible light irradiation." The latter is most interesting for tracking successful recent developments in hydrogen technology. The former is useful for identifying constrains to hydrogen technology. It is closely related to potential and overcoming these constrains equals obtaining a "win condition".
Worldwide	Consists of n-grams 'global', 'worldwide' and 'universal'. Universal occurs in contexts, such as universal model or universal strategy. Universal rarely refers to a global scale, such as universal right, universal access. Two thirds of all documents contain the n-gram "global". It is used in diverse contexts unrelated to worldwide, such as "global warming", "global optimal solution", "global features" (of a substance) and company names. Like universal it incidentally (10%) refers to worldwide (such as global warming). The n-gram worldwide is unambiguously used to indicate something on a worldwide scale.	The n-gram worldwide is the only one that properly reflects the anchor's value. Similar to the value global. Global also contains the n-gram universal.

Table 3.2: Showing the core values found in the hydrogen corpus. Given percentages are based on reading member documents in each topic. Fifty to a hundred documents were typically read for each topic.

3.1.5. Discussion on values over time

The prevalence of each core value over time is shown in figure 3.3. Two y-axes exist in the figure, the left y-axis shows the prevalence of each core value in that year as a percentage of the entire corpus. This was done for all documents in the corpus by interpolating the results of one sample. Prevalent means that a document is a member of this core value. The right axis shows the number of documents in the corpus for that year. The corpus itself contains older documents, even ones before the second world war. The quantity of documents before 1980 is too low and is therefore not included in this figure. Lower document quantity reduces the reliability of the value signature. Before 1980 the document quantity quickly drops to dozens of documents which was considered to not be worthwhile presenting.

All three axes hold important information and consider the following assumptions when observing the figure. First assume that a low document count reduces the importance or reliability of the value signature. Secondly consider that more recent articles and values are more important. More recent articles are more important in informing what is important right now resulting in depreciation. Lastly the higher the value signature is, the more important the value is. Using these assumptions, observing the figure yields several insights.

Value hierarchy

The first insight is that value importance changes over time. Four values stick out, namely reliability, conformity, potential and best. Best is distinctively unimportant despite showing a peak in the eighties. During this peak it is coupled with a low amount of documents and the time period proceeding this shows a distinctively low signature. Conformity shows a peak after 2001 and declines around 2008, but remains the second most frequent value in the set. Potential is the most frequent value prior to 2001 where its value frequency is double of other values. Reliability takes over this role where its relative frequency is even higher.

There two ways to look at this insight, one is using a contemporary practical purposes and the other historical empathy. The former one is useful for applying the results and the latter one is useful when doing topic modelling. In the former perspective we look at the results of the past five years or so and how we can use that today. Here it can be concluded that reliability is most widely discussed value in hydrogen. All other values, but ”best” are relevant too. The latter perspective describes what values are important when. It acknowledges that these trends change over time. This perspective wants to explain why this comes to be. This process of giving meaning or explaining why things are is thoroughly discussed in the next insight.

Meaning of values over time

The second point is what these values over time reflect considering several perspectives to this meaning. The first perspective is that it reflects nothing. Values are an interpretation of the CorEx optimisation criteria which is based on the tfidf vectorization scheme of a specific set of documents. Note that there are four dependent variables: the corpus, tfidf, CorEx and the assumptions required to make the interpretation (which are the n-grams used to label anchors and the values). Change any of these variables and so does the graph. Let us proceed to the second perspective with this disclaimer in the back of the head. It reflects everything and there will be no examples, because this implies that some entries in this infinite enumeration are more important than others. The third perspective is obtained by taking any arbitrary subset of this infinite enumeration.

The first perspective is the truth and extracting meaning from results requires improvement of the technique. Ranked in importance, the first hurdle is using multiple assumptions instead of one to obtain results. This is because the model just propagates the assumptions (anchored n-grams). Values from multiple runs should be compared to each other see B.1 for more on this. Second order of importance is using different topic modelling and word vectorization schemes. As it stand for now the quality of the results is similar to asking a random person. The difference is that the model devoured a much higher quantity of information to obtain its interpretation.

When referring to the following allegory 2.1.2 to explain how the model works, it becomes clear that this model has zero insight. An individual can have insight, an export opinion is better than the model’s or a random individual’s, because it has the most insight. There is furthermore a risk when using the model. One of the most important things about humans is giving meaning to things (3rd perspective) which happens automatically. This influences ”artefacts” such as norms, values, religion, language, ideology and identity. A dominant believe in contemporary times is the one in numbers and machines. Specifically referring to the believe in the implied objective truth of numbers, and more recent the future together with or in the hands of AI. This quasi-religious concept will become more important as predictive qualities of AI improve. This idea fits in the progressive development from nomadic to sedentary societies (with organised religion) to nation-states (with secular and literate masses) into potentially globalised digital constituents (algocracy). It is a progression from God, human to AI who are authoritative in informing and decision making. This believe is the ”pre-work” for this new social contract which is not achievable yet, because the predictive qualities of AI aren’t that good yet to compete with individual decision making. The believe is there, a notable example is the Dutch childcare benefit scandal where algorithms were used to identify ”high risk” individuals whose benefit money was withdrawn by the government. In case of topic modelling, the topic model would empower the individual. While desire is there; it is unable to do so. The expert opinion (prophet or theologian) remains the most authoritative and the model can not be used to compete with this. The point of topic modelling is to eventually replace the expert opinion in informing what the contents and values are in a specific domain.

Accepting the desire and necessity to add value to results and assuming that our dependent variables suffice allows the second and third perspective, attributing some explanation to the results. The second perspective states that any given example pretending to be a sufficing explanation is false. An

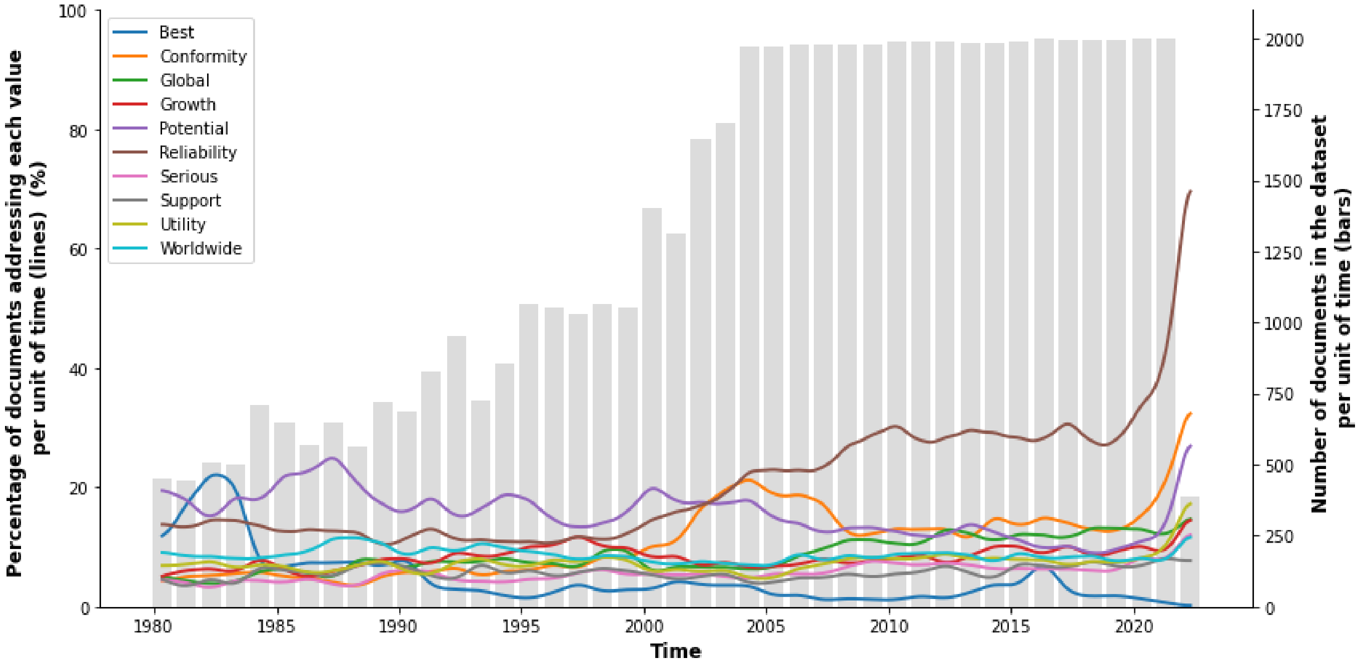


Figure 3.3: Core value prevalence over time. Showing the prevalence of core values where anchors contribute more than 50% of all mutual information.

explanation is an if-so story and the choice of explanation is arbitrary. Explicating any explanation passively inhibits others by priming. Now the first and second perspectives are mentioned it will promptly be ignored and an explanation is given. Some values are more important at different points in time, because this reflects developments in literature. It also reflects cultural conceptions of what is important or standards in a discipline. A good example of this is shown in figure 3.4. This figure looks closer at the value topic of reliability. It shows the prevalence of n-grams in each document belonging to this topic. It shows a distinct development of word choice over time. Robustness starts to be mentioned only after 1985 developing into the most common n-gram by 2015. The graphs reflect trends in n-grams, words and values. Choice between words (with different meanings) might not be so important as it appears to communicate the same thing. Topic modelling allows the bundling of these words which allows the discovery of a larger document set who communicate the same idea.

Corpus growth
Last point is the growth in scientific output in hydrogen. Document quantity over time steadily increases and flats out at around 2000 documents. (Bornmann et al., 2021) found a growth rate of 4.1 percent per year averaged across many disciplines. The hydrogen corpus follows this trend until the year 2000 when a 5 year period rapid growth beings followed by flatting out at 2000 documents in 2005. The growth rate between 1980 and 2004 is around 6%, which includes this period of rapid growth. The stagnation occurring between 2004 and 2022 is a strange observation. Since the values don't exhibit the growth or the deviation in output seen in previous years. It can be implied that this is most likely caused by a technical issue during the corpus retrieval. If the 6% growth rate would persist after 2004 then 5000 documents per year are expected by 2020.

3.1.6. Discussion on obtained values

A set of values was obtained by using the methodology described in section 3.1.4 which has several constraints. First reproducibility is not very good. Rerunning the model gives a different set of results, half of the values appear consequently, the other half does not. And second, if a value is not discussed does not mean that is not relevant. Not only because it might not be found by the model, it may also not the subject of attention in scientific literature. Making these underexposed values that play an role in the public perhaps even more interesting.

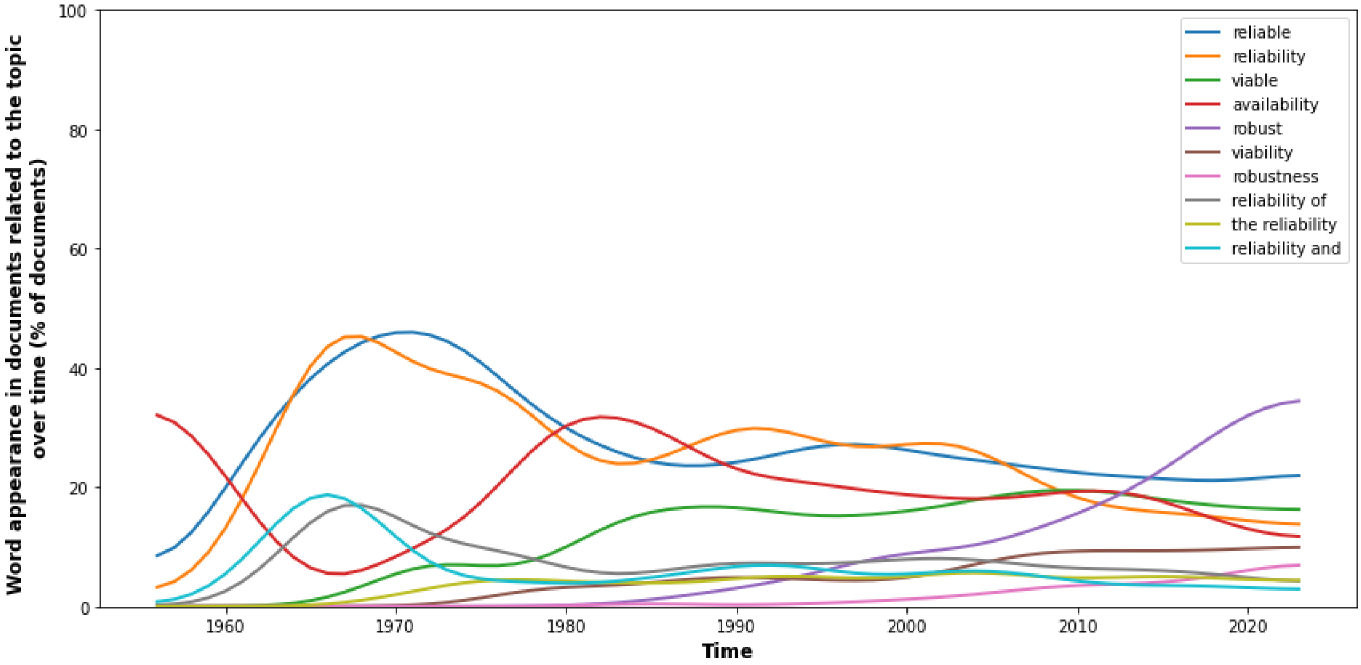


Figure 3.4: N-gram prevalence over time within one value. Showing the prevalence of n-grams over time within the core value reliability

Limitations of quantitative results

Observed problems are “dilution of topics”, “problematic synonyms” and “low anchor quality”. These problems are observed as follows. A diluted topics contains a significant portion of documents that are unrelated to green hydrogen or the topic’s intended meaning. Take as an example the topic global. The n-gram ”global” is used in the context of global (solution) in simulation, therefore the topics starts to include n-grams and documents related to simulation. The second problem are ”Problematic synonyms” which causes articles to be included that are unrelated to the intended meaning of the anchor. An n-gram is unrelated to the meaning of the anchor and the document is nevertheless selected to be a member of the topic. An example of this is the n-gram consequence, which selects articles using the n-gram ”consequence of” relating to cause and not to impact. Both problems are caused by taking n-grams ”out of the intended context” and a set of articles in the corpus that are unrelated to what one is looking for (green hydrogen and values). One should see topic dilution as a pattern emphasising the latter problem, a poorly filtered corpus. In this pattern some sub-topic can be identified. Problematic synonyms should be seen as emphasising the former, namely recurring incidences depending on the use of n-grams in language. Low anchor quality means that the topic fails to appear or has a set of member documents that poorly fits the anchor.

An intuitive solution to the first problem is filtering articles unrelated to green hydrogen. For example, in the value global documents (and n-grams) on numerical simulation are included as global frequently occurs in the context of “global (solution)”. Considering the tfidf filtering these documents increases the quality of topic. Since the main n-gram of these filtered documents, global, now occurs in a much smaller fraction of the documents in the corpus, its importance increases. Other irrelevant n-grams, such as “numerical simulation” are now removed from the topic, which has three benefits. First it improves the congruence of documents within the topic. This in turn increases the importance of existing n-grams and lastly allows the inclusion of other n-grams. It lastly has to be mentioned that filtering is imperfect, not all articles can be filtered and some correct articles will be filtered.

The second problem should be solved by changing the anchor words. It is not really known how to do this, but trial and error. Use of relatively specific n-grams can affect the results. Changing the word filter parameters appears to play a minimal role. Selective filtering of documents affects the values allocated to each n-gram in the tfidf. Document filtering is therefore instrumental, because the algorithm seeks to maximise the informativeness in the topic. Therefore changing anchor words or adding filters will not prevent the model from allocating such documents to topics. If needed the topic model creates topics in the most cumbersome way imaginable. It does so, because it has nothing better to do, but maximising TC. Topics are consequently created with keywords being nothing more but anaphoras, such as ”which” and ”these” and other topics are filled with years. The corpus therefore determines the results.

In the case of low anchor quality, the anchor might not be present or insufficient and poor n-grams are included. It is similar to the second problem, but differs as it spans the anchor dimension. This has two implications, first it is impossible to interpret the results manually. Recall a machine interpreted the location of 433 anchors’ with a total of 2118 n-grams in 500 latent topics. Position of anchors don’t match with latent topics and n-grams can occur multiple times. Optimising anchors is secondly ideally done through an automatic approach. Manually checking and improving existing topics could happen if one would know from which set of n-grams the anchor words have to be chosen. Since this set is larger than 14 thousand n-grams (typical number of n-grams in a sampled corpus) an automatic approach has to be developed.

Four algorithms are proposed to make CorEx fully automated. The first proposition is an algorithm that selects keywords from latent topics and appends these keywords to anchor as anchor word. The second proposition is to use a database to check synonyms for existing anchor words. Then check if synonyms are in corpus, if so append n-gram to anchor. The problem with these methods is that quality control remains manual. A suggestion for quality control algorithm has the goal of identifying irrelevant documents and n-grams. Obtain for each latent topic the document ids Create for this sub-corpus an individual topic model. Interpret from each latent sub-topic’s keywords what is and what is unrelated to the main-topic’s contents. Identify if a sub-topic suffers from “dilution” or “problematic synonyms”. Then add these documents causing dilution or n-grams causing problematic synonyms to an exception list. Conduct in the next iteration the main-topic model without these documents and n-grams.

3.2. Context identification in hydrogen

A second set of topic model iterations were performed with the purpose of finding context in the hydrogen corpus. Anchor words were added for each topic model run using the results of the previous runs. Each run produces latent topics as output containing keywords and documents which can be used for finding coherent and relevant topics. A relevant topic refers to hydrogen and a coherent topic consistent refers to a single subject. Coherency was judged each iteration and a coherent topic should be reproducible every run for the same settings. A topic was deemed reproducible if the topic appeared and the subject did not change. Judging relevance was done during each iteration and after obtaining all topics. The latter is an interpretation of the ”final” results and in the former a topic was deemed relevant if it referred to hydrogen. Each latent topic consists of keywords and documents supporting this decision making process. The iteration started without anchors. Anchors were obtained by using the keywords from a relevant and coherent latent topic. This also means that the first iteration had no anchors. An iteration consists of running the model, interpreting the model and changing the dependent variables of the model. This procedure was continued until no more new topics were found.

Reproducibility was achieved through structuring the result logging, transparency in decision making, and commenting on each iteration. Results of each iteration are interpreted which determine the settings and anchors of the next iteration until the final results are obtained. Final results are obtained when all coherent topics are thought to be found. Furthermore, several improvements were made in this research phase. For every run the model object is saved, visualisations are made and output topics were saved as $.csv$ file. Each of these documents are labelled by the number of topics, sampled documents and unix-timestamp for that iteration. Execution speed of the original code was improved by debugging. Ease of executing an iteration was achieved through compiling the notebook code into a python object. This allowed the code to be executed in a single line and appeared to reduce the memory load allowing higher number of topics and documents on lower-end computers. Commenting on each iteration involved interpreting the results and providing arguments for the settings in the next iteration. These logs can be by contacting the researcher and a short summary is provided below .

Description of the iteration process

A description is given of the settings used to run the model. The set of model executions (run) is called the iteration process, because model settings and anchors are changed every run. The number of topics were varied between 50 and 500 and number of documents between 1000 and 15000. The iteration started with 500 topics and 5000 documents (which took 4114 seconds to complete). More than 400 topics took long to complete, gave results that were difficult to interpret, and yielded empty topics. Topics that were present were tended to change every iteration making them difficult to classify. It was chosen to reduce the number to 100 and 50 which yielded coherent topics. Using less than a hundred topics caused the disappearance of some topics however as these would be aggregated into larger topics, but gave a result that was easier to interpret. Results depending on higher number of topics become unstable for having a hundred topics or more. Unstable means that these topics do not appear consistently regardless of anchoring. Disentangling the existing topics was not seen to improve the clarity in results since these results were already extensive and appeared to reflect all contents of the corpus. It was therefore chosen to keep the number of topics at 50 or 100.

3.2.1. Classification of context in themes

Giving meaning to the results of the topic model was done by interpretation of keywords and documents given with each topic. This interpretation resulted in four different results, namely topic name, topic range, values and theme. The first three are shown in table 3.3. A topic name was created that attempts to represent the document's content in as few words as possible. There is not a single run that can fully represent the results since the results are dependent on the number of topics given to the model. Most topics have some range of number of topics on which it can appear. A lower number of topics will result in some amalgamation or distribution and a higher number will cause splitting and or distribution of the topic across other topics.

The last two results are important for answering the main research objective. First are the relevant values, which are some values that are related to green hydrogen and relevant to society. As an example is the topic on proton exchange membranes which is relevant to green hydrogen, but a general topic dominated by technicalities. Values, even those on results and optimisation, don't play an overwhelming role and it therefore doesn't make sense to add relevant values to this topics. Only when a value clearly pops in a topic is it added to the value column.

Themes bundle topics in hydrogen

Themes are groups of topics with similar characteristics and can be seen as the interpreted structure in the corpus. The first theme, governance, is exceptionally value laden, the second is about hydrogen storage where safety stands out and the last category is about hydrogen production. Themes are created by similarities in topic content and values discussed in that topic. It also reflects various disciplines, journals or companies. Themes firstly make a distinction between green hydrogen and non green-hydrogen topics and secondly indicate what values are consistently discussed in that set of topics. While individual topics are a part of the story they cannot fully make the distinction between what is and what is not related to green hydrogen. This is backed up by the fact that topics overlap, documents can be part of any number of topics and all topics (except combustion and PEM fuel cells) are fluid (recall amalgamation, distribution and splitting when changing the number of topics) and depend on a certain topic range.

Themes above topics

Themes are more useful for determining and presenting which documents are part of green hydrogen than topics. It could always be argued that this is not the case, because they are intrinsically the same. A few advantages of themes is that they are better presentable than topics and are selected based on their relation to green hydrogen and how values are discussed. Themes bundle a set of topics because the topics have strong overlap qualifying their relation to green hydrogen. This means that it becomes redundant to make the distinction and overlap between these topics explicit, which brings us to the second advantage of themes. Themes reduce the amount of boundaries between groups of documents that have to be considered. With that themes reduce the contention to justify discrimination in document membership. A clear distinction between topics doesn't exist in several cases, because topics show strong overlap in content and document membership. This property can be described with the word fluidity and weakens the distinctiveness of that document grouping.

Let us take an example to illustrate the fluidity of topics. The topics space, petroleum engineering and desulfurization are taken as an example. Space and petroleum engineering are closely related through fluid flow simulations and petroleum engineering is closely related to desulfurization as both are important in the petroleum production process. Space has little to do with desulfurization however. It is unsurprising that introducing a topic on simulation would more often than not would compete with

Topic number	Topic name	Topic range given k topics	Relevant values	Theme
2	Investment projects	50	Development, future	Governance
3	Natural gas power plants	50	Supply	Governance
12	Abatement policy	50	Alternative	Governance
13	Environmental impact analysis	100	None	Governance
24	Electrolysis	50	None	Governance
28	Solar hybrid technologies	100	Efficiency, attractive, ideal, improving	Governance
29	Electricity and hydrogen generation schemes	100	Participation	Governance
4	Proton exchange membrane (PEM fuel cell)	Any	None	Hydrogen production
5	Bed reactors	50	None	Hydrogen production
7	Green hydrogen steam reforming	100	Cost-effectiveness, efficiency, feasibility	Hydrogen production
11	Biogas	100	None	Hydrogen production
18	Steam reforming (gasification, pyrolysis and syngas)	100	None	Hydrogen production
19	Gasification	100	None	Hydrogen production
14	Hydrogenation	50	None	Hydrogen storage
20	Hydrogen embrittlement (and nuclear reactors)	50	Safety	Hydrogen storage
26	Hydrogen storage	100	Safety, Interest	Hydrogen storage
27	Alloy storage	100	None	Hydrogen storage
1	Combustion	Any	None	None
6	Hydrogen isotopes	50	None	None
8	Photochemistry	100	none	None
9	Methods for atomic (scale) analysis	50	None	None
10	Simulation	100	None	None
15	Space (and aeronautics)	50	None	None
16	Impedance spectroscopy	100	None	None
17	Gas chromatography	100	None	None
21	(Nuclear) fusion	100	None	None
22	Desulfurization	50	None	None
23	Petroleum engineering	50	None	None
25	Hydrogen peroxide	100	None	None

Table 3.3: List of topics in hydrogen database with their interpreted topic name. Topic number is the order in which the topics were found, lower number indicates that the topic is easier found over runs and it is easier for the model to self-generate this topic. Topic range addresses the number of topics in which this topic typically appears. A single number means that this number of topics is ideal for finding the topic, increasing or decreasing the number causes disentangling or merging changing the meaning of the topic. Relevant Values indicate the values playing a role in this topic indicated by keywords.

space and petroleum engineering causing the disappearance of the latter two. This is the first reason for excluding simulation next to the believe that a methodology does optimally reflect content or context. Similarly, desulfurization is a chemical process important in biogas. Topics have specify context and are not their relationship to green hydrogen making it complicated to find this relation.

Excluded topics

Topics are grouped together into three themes related to green hydrogen. However, topics on combustion, hydrogen isotopes, fusion, space, simulation, photochemistry, petroleum engineering, and methods for atomic scale analysis don't fit any of these three categories and are unrelated to green hydrogen. Hydrogen embrittlement and petroleum engineering do have things in common with hydrogen technology, namely hydrogen transport and underground hydrogen storage. Unlike hydrogen embrittlement, petroleum engineering can safely be ignored as any explicit reference to hydrogen storage will likely cause allocation to the hydrogen storage theme.

Governance theme

The first identified green hydrogen theme was governance and is further divided into two categories. The first category is executive and economical. It fits topics on power plants, investment schemes, solar hybrid technologies and electrolysis. The latter two were its weakest topic were inconsistent and changed every single run. On top of that, electrolysis was seemed to be related to power plants and proton exchange membrane (PEM) fuel cells. Next are environmental impact analysis and abatement policy fitting this theme as it is related to the non technical aspects of hydrogen and is the second category in this theme.

Values in the first category emphasise ”renewables and technology” on the second place are ”economic interests (supply, demand, patents (intellectual rights), costs)” and on the third place ”teamwork, globalisation, and cooperation with universities”. These two topics exhibit values on ”threats, imperatives, prove, danger, health, protection, impact, generations, motivation, ecology and environment”. This could be considered as a topic on legitimacy building, steering decision makers and societal norms. In the second category was the topic on environmental impact analysis found to be the most distinct and value laden topic with a ”normative, collective, philosophical and idealistic touch”. A noteworthy mention is the topic on ”participation”, which is too weak of a topic to be reliably identified. This topic seeks to answer the executive and economical question by stimulating decentralised projects implying that this is more equitous. It deserved to be mentioned as it seeks to serve the same purpose, while completely breaks with the second and third order values in the first category.

Hydrogen storage theme

Second is the hydrogen storage theme consisting of numerous different topics (low temperature, high pressure, reservoir, metal and other chemical processes) which all have their problems in topic modelling. Keywords used in hydrogen storage are common in other topics, for example, reservoir storage of hydrogen is hard to distinguish from petroleum engineering. Other chemical processes related to storage are hydrogenation and another mention is adsorption that sometimes formed a distinct topic. It is observed that many contents related to hydrogen storage continue to float outside the topic. Another issue with storage is the phenomena of hydrogen transport. This topic was not identified, but it is expected that hydrogen embrittlement plays an important role in both storage and transport. Hydrogen embrittlement is inseparable from safety and closely related to nuclear power plants. In hydrogen embrittlement the majority of documents are unrelated to green hydrogen. In hydrogen storage safety and interests of various stakeholders play an important role.

Hydrogen production theme

Introducing the third theme is done by starting with the topic on desulfurization. This is typically a hydrocarbon topic (referring to the chemical petroleum section and not petroleum engineering). As expected, desulfurization also plays an important role in biogas and syngas. Natural sources of hydrogen sulphide is on the other hand a source of hydrogen. Consequently hydrogen production can be introduced as the third major topic. Bed reactors, steam reforming, gasification and electrolysis now join the set. Solar hybrid technologies and electrolysis are hydrogen production technologies, but considered to be stronger members of the first topic on investment schemes. Characteristic for this theme is that its topics and content are the most technical of all three themes. Typical values playing an important role are efficiency, feasibility, cost-effectiveness, results and improving. It could be summed up as ”result-oriented” or evaluation of novel hydrogen production technologies.

3.2.2. Discussion on hydrogen context

Compared to the previous section on anchored core values, the results of context identification yielded distinct topics and themes. The two distinct results, topics and themes, shown in table 3.3 are described and discussed. Topics are the names given to the set of keywords and documents grouped together by the topic model. They relate to specific contexts, although the specification of this context remains coarse grained. Topics are useful for identifying sets of documents on specific subjects.

A theme is the classification given to the set of topics that is useful for human interpretation. Useful because it delineates what is talked about in hydrogen. This delineation is complicated to substantiate for (a multitude of) topics, but doable when assuming that the set of topics are one ”theme”. The classification in themes is useful for making the distinction between green and non green hydrogen and making the distinction between broad categories of documents. Purpose of themes is to identify which document IDs are related to green hydrogen and indicate what values are discussed throughout the theme.

The most important result is not only identifying the contexts, but obtaining the documents of relevant and irrelevant topics allowing the filtering of documents from the corpus. Unrelated topics contain a lot of information that will dilute results, making it harder or even impossible to find the desired results (identifying values). Topics are not specific enough to qualify the contents of a topics. Obtaining a finer grained result that qualifies the contents of topics is not possible using this corpus and these model settings as these topics first fail to emerge consistently and secondly, the contents of these topics doesn’t change. This means that the inputs of the model have to be changed. The purpose of themes and topics is identifying the documents related to green hydrogen or a specific context respectively. This not only allows the analysis of these result, it also allows the running of a topic model over this new set of documents.

3.3. Presence of generic values in Hydrogen

A third approach was taken to find values in the hydrogen corpus using a generic set of anchors. This set of anchors is commonly used by faculty members to find values using topic modelling and here the set refers to ”general purpose use”. These values regard broad societal relevant values and from these, the ones that can not reasonably relate to green hydrogen, such as privacy are filtered from this set. When comparing the anchor set to core values in section 3.1, these anchors are more compounded (using only 22 values), hence more ”general values”. This was a set of anchors commonly used by faculty members to find values using topic modelling. These values regard broad societal relevant values and from these, the ones that can not reasonably relate to green hydrogen, such as privacy are filtered from this set. This anchor set differs from the set used in section 3.1, which was based on ”core values (in organisations)”.

After preparing the anchors and corpus a large set of model runs were performed with 100 to 600 topics and 3000 to 6000 documents. The documents are sampled from the same corpus as in the previous exercises. From these model parameters one representative run is sought. This run is then used to determine if the anchor can be found in these results (latent topics). Several criteria were developed to systematically determine if an anchor is or could be present in the corpus. The results (latent topics) of this run are enumerated below and shown with criteria in table 3.4. A topic selection is made based on the criteria. The criteria are sufficient to judge if a topic can or can not possibly relate to the anchor and its intended value. The documents of the selected topics were read to make a well considered decision whether the anchor is properly represented in the output or not. These values are presented in table 3.5. A sufficient amount of articles were read for each of the latent topics. Based on this reading, a decision is made if the corresponding value is present in the latent topic. This judgement can be extended to the hydrogen corpus.

3.3.1. Topic modelling results

A representative run was sought for obtaining the results, applying the criteria and determining the presence of values. This representative run has distinct model parameters (number of topics and documents). A trade-off exist between the number of topics and documents. Both increase the quality of results and computational costs, but it is not possible to maximize both criteria. Increasing the number of topics increases the disentanglement of the corpus. Increasing the number of documents increases the number of n-grams and documents fed to the model potentially increasing model resolution. It was found that sitting below the level of full disentanglement (between 500 and 550 topics) does not change the results in such a way that it is no longer a representation of the results. All relevant anchors continue to be properly represented at 400 topics showing that lower topics had a negligible influence on the results. However, lowering the number of documents increased the inconsistency in the representation of anchor words in latent topics. Increasing the number of documents increased the quality of the results, but this effect was seemingly no longer noticeable beyond 5000 documents. It was therefore chosen that the representative run (number 1663544953), should have 400 topics and 6000 documents.

Several criteria are developed to classify the results (latent topics) positively or negatively and results in a judgement. To clarify, positive and negative respectively indicate why the anchor is part or is not part of the latent topic. This classification is used to determine which topics are worthwhile investigation of their member documents. Three judgements exist. **Discarded** is the first and occurs when an anchor is completely absent(not a single anchor word is present in the latent topic). Absence of positive and presence of negative criteria is an indicated that the latent topic can be **excluded**. Excluded topics are, but discarded are not discussed in the comments. All other latent topic's are discussed and their documents are analysed. The **present**, **ubiquitous** and **not found** judgements are a results of inspecting the most informative n-grams. Present means that the n-grams correspond to the anchor, while not found means that there is no connection between n-grams and the corresponding anchor. Ubiquitous means that n-grams are selected that correspond to the anchor, but could relate to any other anchor as well. Results of these classifications are found in 3.4 under the column "result" and the present anchors are used in the next step of the analysis.

Negative criteria

"Negative" criteria are first discussed, indicated by capital letters in the result table 3.4. First of the list is, **(D)**, disqualified and different topic. A crisp explanation is that the anchor is not there, it's absent, therefore disqualifying the anchor resulting in a "discarded" judgement. A more detailed description is that a in a disqualified anchor one observes that the results reflect the first subsequent anchor (that is not disqualified). A different topic appears, and this happens because the anchor is too weak to compete with all other topics and self-generated topics. The model believes it is better off without these anchor words. This conviction is so strong that resetting the anchor words back to their original strength during every learning step is not enough. Furthermore an algorithm is used to search these words and if they may appear in any other topic. A disqualified anchor requires its anchor words not to be found elsewhere. Disqualification of the anchor is done with complete certainty. It can be said that the anchor set is not present in the output and cannot be found. Certain, because the model is described at a fully disentangled level for a sufficient number of runs. Therefore further increasing the number of topics or documents does not change the outcome.

Other criteria are: **F**, few anchors, one or more, but no more than three anchors are found in the output keywords. **C**, confounding keywords, means that the anchored keywords found in the output topic are associated with a set of keywords that are unrelated to the topic for which the anchor set was originally designed. For example in accountability the anchor n-gram "responsible for" is associated with chemical reactions ("using the 20 wt% Ni-catalyst might be responsible for the reduction of hydrogen production." (Wang et al., 2021)). A topic is marked as confounding if this trend is clearly seen to dominate the keywords. Lastly is **L**, low informativeness of keywords, where anchor words contribute less than 10% of all mutual information to that topic. Note that is the total mutual information of a topic is the sum of all member n-grams their mutual information.

positive criteria

Opposite to negative are positive criteria of which two exist. **H**, high informativeness of keywords, occurs when anchor n-grams contribute more than 40% of all mutual informativeness to that topic. Last is **M**, multiple n-grams, is reserved for latent topics containing more than five anchor words of their corresponding anchor. Only two positive criteria exist, this is not a problem since negative criteria, confounding and disqualification are the only ones used to dump the anchors before proceeding to read the individual documents.

0. Accountability appears to confound with causal relations x is responsible for y or x accounted

Anchor set name	$(anchor\_id)$ - $(topic\_id)$	Positive observations	Negative observations	Result
Accountability	(0-0)		C F	Excluded
Autonomy	(1-1)	H	C F	Excluded
Comfort	(2-absent)		D	Discarded
Conformity	(3-absent)		D	Discarded
Cooperation	(4-2)		C	Excluded
Democracy	(5-3)		L	Not Found
Equality and economic development	(6-4)		F L C	Excluded
Economic viability and welfare	(7-5)			Present
Efficiency	(8-6)	M H		Ubiquitous
Environmental sustainability	(9-7)	M		Present
Fairness	(10-8)		C F L	Excluded
Freedom	(11-9)		C F L	Excluded
Health, safety and security	(12-10)	M		Present
Inclusiveness	(13-11)		C F L	Excluded
Intergenerational justice	(14-absent)		D	Discarded
Justice	(15-absent)		D	Discarded
Privacy and intellectual property	(16-12)		C F L	Excluded
Reliability	(17-13)	M		Not found
Resilience	(18-14)	H M		Not found
Transparency	(19-15)		C F	Excluded
Trust	(20-16)		C F L	Excluded
Well-being	(21-absent)		D	Discarded

Table 3.4: Quantitative results judging if the value, of the anchors commonly used in topic modelling, is present in the corpus. Three result criteria are used. Discarded topics were absent from the corpus during all runs. Excluded topics means that some of the anchors in the anchor set are part of any of the latent topic's keywords. Although insufficient evidence was found for including them. Included topics are congruent, highly informative and distinct topics that appear to correspond to the intended meaning of the anchor.

for z in y. Topic is heavily entangled with overlapping context of the documents where this was found.

1. Autonomy similarly focuses on ”convention” and ”order” and appear to be drawn out of context. E.g. ”in order to”, ”(preposition) conventional”. It clearly has nothing to do with autonomy and can be safely excluded.
2. Cooperation contains a collection of keywords on ”international”, ”years” (integers), and ”conferences”. The topic seems to be related, but inspecting individual documents shows that there is no relation to cooperation.
3. Democracy topic had insufficient reason to be excluded. Inspection of articles shows that none of the articles were found to relate to democracy whatsoever.
4. Equality and economic development is about power plants and its original meaning relating to distribution of means is absent in this topic. This original meaning can only be seen by devious interpretation of the n-grams ”power”, ”status” and ”power systems” in Marxist context. This is misplaced however as it clearly relates to power plants and (status to) operational conditions.
5. Economic viability and welfare relates to macro level processes on a global level. It discusses technology, energy demand and infrastructure and hydrogen economy. Articles talk about quantitative research on the improvement of renewable technologies, and quality of life improvements, such as reduction in emissions and implementation effects such as utilising residue heat.
6. In the efficiency topic all n-grams in the keywords contain the word efficiency. Inspecting individual documents show that the documents are related in content or objective. Sometimes documents relate to improving a component in a specific process (the subject). Efficiency, improving, etc. is ubiquitous. This topic’s added value is questionable. Furthermore, it selects some and not all articles containing words related to efficiency (see discussion). This brings the topic’s usefulness into question, however its presence is evident.
7. Environmental sustainability discusses renewable energies and the topic of climate change in general. This topic is well delineated in content and defined by keywords. Closer inspection of individual documents show that it contains a significant portion of outliers (technical subject, different field unrelated to the topic such as petroleum or nuclear power). Articles discuss applications of environmental sustainable technologies of green hydrogen, such as hydrogen storage, fuel cells, hydrogen production and electrolysis.
8. Fairness’s anchors consistently appear in the results, but its contents are arbitrary and unrelated. The topic is therefore discarded.
9. Freedom’s anchors consistently appear in the results, but is typically high-jacked by a context topics such as petroleum engineering or gasification. The topic is therefore discarded.
10. Health, safety and security does relate to health, safety and risks, but heavily focuses on accidents and specifically those in nuclear reactors. Documents reinforce this and show that petroleum also plays a significant role. A minor role (lower than 10%) is reserved for topics related to hydrogen topics (biofuel, ammonia, fuel-cells and storage)
11. Inclusiveness is a topic containing incoherent and unrelated content.
12. Privacy and intellectual property similarly contains incoherent and unrelated content.
13. The topic on reliability relates to arbitrary context and reliability, viability and availability. Inspection of documents show that none of the documents seem to relate to reliability.
14. Resilience: consistent set of highly informative keywords. Could be assumed that this topic is present, although inspecting of documents could not find a single article using the term or a meaning related to resilience.
15. Transparency: consistent set of highly informative keywords. Transparency could unfortunately not be identified when searched individual documents.
16. Trust selects the keyword ”open” with a set of arbitrary and incoherent keywords. Trust cannot be identified in this topic nor in the individual documents.

Anchor set name	Interpretation of anchor in topic results
Democracy	Not present
Economic viability and welfare	Present
Efficiency	Present
Environmental sustainability	Present
Health, safety and security	Present
Reliability	Not present
Resilience	Not present

Table 3.5: Value presence in Hydrogen corpus. Here the anchor sets reflect values with a possibility of being present in the corpus. Four relevant values were found and three were not present. This was decided after reading articles member of each anchor's latent topic.

3.3.2. Discussion on results

Four values were found in the hydrogen corpus, namely Economic viability and welfare, Efficiency, Environmental sustainability, and health, safety and security. Economic viability and welfare showed the least problems and is most directly related to green hydrogen. The other three values show several issues however.

Impact of corpus filtering

Some articles in environmental sustainability and most articles in health safety and security (90%) are not related to green hydrogen. In both topics the articles relate to hydrocarbon and nuclear fission literature instead. Solution to this problem is filtering the corpus based on the results in section 3.2. If the corpus was filtered these topics might not have and other might have appeared.

Allocation of document-topic membership

Two new problems with the allocation of topics were discovered by analysing the efficiency topic. First is issues in allocation of document membership. Namely not all documents discussing and expressing efficiency are part of the efficiency topic, but of other topics instead. Consider two perspectives on this phenomena, first consider that CorEx's Total correlation maximisation objective is reached through "competition" between topics (Gallagher et al., 2017). Words that have high document frequency, such as the ones in "efficiency", have "low" numerical tf-idf values. This means that their presence conveys little informativeness to that topic making them targets for this competitive behaviour from other topics. Efficiency shows that this process went too far resulting in incorrect allocation for the sake of fulfilling this objective. Fitting the debate of anthropomorphism in AI, this example shows that machine learning is nothing more than an optimisation process (Salles et al., 2020).

Multiplicity of document-topic membership

Further investigation illustrates a second perspective to this observation stemming from the multiplicity of a document's membership. An algorithm was developed to check (verify and/or validate) how well anchor words fit to their corresponding latent topic and all other topics. This was done by counting the occurrences of these anchor words among the corresponding latent topic and comparing it to all other topics. One would expect a high count for the documents, member of the latent topic corresponding to the anchor and a low count for all other latent topics. The algorithm shows that counts are generally three times higher for the corresponding latent topic when compared to all other latent topics. This shows that there is some differentiation between anchors, but that there is a significant overlap between topics. This overlap is high considering both perspectives.

Oscillation in total correlation

The second problem was discovered while seeking for a solution to the first problem. Observing the historic total correlation values shows an oscillatory pattern in initial model runs. This means that when the model is executed, total correlation varies drastically during the first runs and eventually converges to an equilibrium. Fact that it does this oscillation and does so intensely, varying TC by factor ten, indicates that the model learning is too sensitive. This may explain not only the heavy oscillations, but also why anchor words are poorly differentiated among latent topics and why anchors are frequently extinguished in favour for (meaningless) self-generated latent topics repeatedly observed throughout all topic modelling exercises. λ, the sensitivity in learning word-topic membership may be the solution to this problem (lambda is too high), although this variable is not mentioned in code.

Lambda

Shortly turning to the original paper finds that sharpness of the learning function λ plays an important role (Gallagher et al., 2017). λ affects how α, the word topic distribution, is learned through each iteration. α is in its turn responsible for the model optimisation (total correlation). In early iterations the model should be relaxed allowing n-grams to freely flow between topics. A harder criteria is gradually imposed giving shape to topics. While the optimisation criterion, responsible for breaking the model iterations, is heavily relaxed by adjusting ϵ a change in results was not observed.

Observing the total correlation variable over all iterations reveals the issue. Early iterations show oscillatory behaviour in TC after which TC converges to an equilibrium. Lower bound would be twenty to thirty percent lower than the equilibrium and higher bound ten times higher than the equilibrium. Default value of epsilon (allowing a difference between higher and lower bound of one in a million averaged over the last ten runs) cause the equilibrium to never be reached (ϵ was 300 thousand fold increased and this value was used in the last two sections). The upper bound in the oscillation is alarming.

Having a too weak criterion for λ, explains the highly oscillatory behaviour in early iterations as it allows a high flow of words between topics. It also explains the extinction of anchor sets and the weakness of anchors on contents of latent topics. Having a low criterion for λ by default is logical, because this would consistently yield higher values of TC, which is the best and most consistent result that the model can attain for unsupervised (unanchored) models. In a semi-supervised anchored setting where a search is performed, TC is no result criteria in any way.

Total correlation and disentangling

Disentangling maximises the total correlation and is done for choosing a sufficiently high number of topics. This is done to ensure that all possible information from that corpus is retrieved by the model. Maximising total correlation should not be an objective at itself as it does not convey anything about found topics itself. As entanglement can describe a level of richness in a topic, one could argue that having entanglement is beneficial. Lastly the chosen representative run with 400 topics, was not fully disentangled, but was nevertheless considered to be a better representation than the fully disentangled models.

Topics were varied between 100 to 600 topics and it was found that in this range the number of topics has negligible influence on how well the model is able to find anchors and anchor words in topics. Increasing the number of topics does not improve this ability. It typically increases the number of topics that are collections of prepositions or words signalling context.

Conclusion

All in all key giveaways are that four values were found as a result of setting anchors. Topic membership is considered to be too high by the researcher. Total correlation (TC) is a sub-optimal optimisation criteria. Bad for humans, acceptable for machines. This is lower than my expectations. The role of anchors is limited however and their influence on the results is low. So is the role of number of topics. Fully disentangling is not a necessity. Lambda is thought to play an important role, but it is unknown how to influence this parameter. Next to lambda, filtering the corpus is thought to influence the results too.

4 Discussion

In the previous section topic modelling was performed. This section first reflects on these results and identifies important components in the methodology. The second section discusses some aspects of text processing on the results. The third section addresses parameters of the topic model and visualisation of its results. The last section revisits values which forms both input as output.

4.1. Discussion of results

Assumptions is the focal point in topic modelling and therefore this research. Figure 4.1 shows this by conceptualising the topic modelling activities in relation to the operator. Criteria, anchors and (number of) topics are the most important drivers for the results. An overview of these for all three topic modelling exercises (sections, rows) is shown in table 4.1. To summarise, the first variable (second column), topics shows the number that was used to generate the final results (of course it was varied across all possible values). The second section shows a range, because multiple runs instead of a single representative run was chosen for generating the results. The third column, anchors, shows the evolution of anchors during each topic modelling exercise. In the first section the chosen anchors were unaltered throughout the run. In the second section there were no starting anchors. The model was allowed to choose them and the resulting n-grams of one run were fed into the next. The third section used a combination of both, an initial set of anchors were used and updated over time. ”Reinforced” Topics and anchors are fed to the topic modelling algorithm that result in a set of latent topics. Latent topics (statistically inferred abstractions of the topic consisting of θ and α matrices see 2.4) need to be processed.

Section	Topics (n)	Criteria	Anchors	Results
3.1	450	Automatic	Static	Core values (table 3.2 and figure 3.3)
3.2	50-100+	Manual	None: self generated and reinforced	Hydrogen context (table 3.3)
3.3	400	Manual	Generic values and reinforced	Generic values (table: 3.5)

Table 4.1: Differences in anchor inputs and result selection criteria lead to disparate results. Squares are activities and circles concepts. Lines show relations where applicable.

Criteria reflects how latent topics are interpreted which outputs the results (last two columns). Automatic interpretation uses an algorithm (see 3.1.4) to determine which anchors are represented in the latent topics. Automatic interpretation is faster, less prone to error and more consist ant than manual interpretation. It is furthermore the only option for larger anchor sets and anchor sets with similar or overlapping n-grams. Manual interpretation allows the model to reflect on the results. This allows the use of human intuition which is unavailable to machines. Inspection of topic documents and n-grams can give the operator new ideas for updating anchors resulting in better fits. Labels can be attached to topics and overarching ideas represented by sets of topics can be identified, such as the themes in

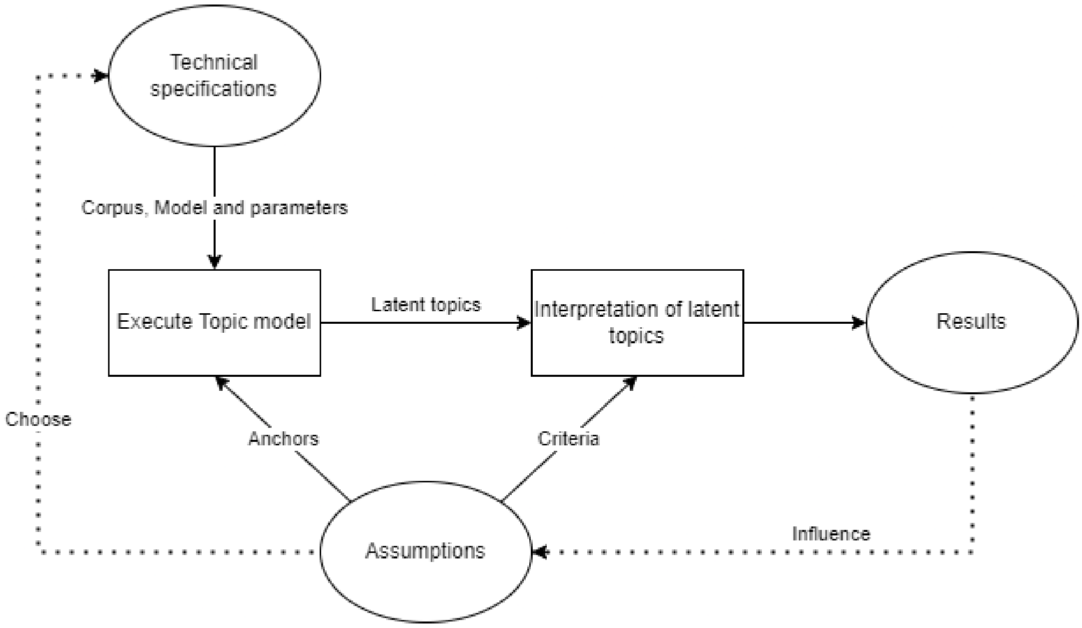


Figure 4.1: Overview of the topic modelling workflow with assumptions being the focal point. The dashed line shows the effects of and on assumptions. Unlike the dashed lines, solid lines reflect the flow of tangible objects. Circles reflect ideas. Squares

table 3.3. Identifying values using topic modelling equals validating the presence of values from a list of anchors in the automatic case. In the manual case this is modulated by bias.

After applying the criteria results are obtained. Results continue to be in the form of latent topics. In the automatic case a subset of the latent topics are obtained. Another property is that every topic is now paired with its corresponding anchor. The resulting latent topics can be visualised or read and interpreted which is presented in tables. In the manual case some selection is made form the latent topic set resulting in some subset. There are no rules for this and in its entirety governed by bias. Presentation of latent topics in figures (backed by numbers) doesn't make sense in the manual case. First because the topic selection procedure was already subject to subjective judgement. Let figures with numerical values not be a distraction. Secondly because the numerical values in a figure can only represent a single run, while the value judgement was based on a set of runs. And thirdly because the numerical values are latent variables, which makes it hard to give meaning to them.

4.1.1. Validity of latent topic interpretation

Let us compare the validity of manual and automatic method in the production of their results. These methods are used to interpret the latent topics to generate a smaller set of latent topics. This smaller set contains the latent topics that we are interested in and makes it possible to manually inspect the documents. We look at validity in their own right without any other considerations. The question is how good are the results of these two methods relative to each other? First consideration is that the most optimal results according to the model converge to one where total correlation is optimised. This has nothing to do with the best solution. A best solution could be one that is held by a relevant expert.

Starting with the manual method, its advantage is that it is possible to strive for this best solution by adjusting the anchors accordingly. By choosing this method there is tacit knowledge of the topic's contents and effectivity of n-grams which is more obscured when compared to the automated method. In any case selection anchors from latent topics is constrained by the optimal solution (TC) and the corpus counteracting tacit knowledge. With some effort manually choosing anchors brings diversity in the results. This prevents the model from reaching its optimal solution. It is therefore unreliable as this is not a new equilibrium position as a small change any settings will cause a change in topics. It is more reliable to improve the quality of the corpus by document filtering. Another consideration is the quality of creating topic inclusion and exclusion criteria, such as in 3.4. This has to be done to make this method reliable and not entirely subjective. It works and allows the inclusion of tacit knowledge. Downsides are that it is prone to mistakes, time consuming, incongruous topic and anchor identifiers and the necessity of cumbersome custom criteria.

The automatic method can simply be considered an improvement. It can always be used independent of topic and anchor size. The method is much faster and consistent when compared to the manual method. Unlike the manual method it doesn't make mistakes or overlook things. Not overlooking things counteracts the advantage the manual method has. The optimisation is based on mutual information and this latent variable doesn't mean a thing. Technically speaking it means these n-grams are the most important in model optimisation, which depends on the corpus and technicalities. Not the real world. Nevertheless, the automatic method is a step beyond. Its advantages are speed and consistency. It also shows that the latent topic n-grams are inconsistent every time the model is run. This is because the automated method yields different results for repeated model instantiations. It shows that confidence percentages of θ are in fact much lower than they appear to be. Fixing this is done by an ensemble run and improves the overall quality of topic modelling. Anchor quality control is a serious limitation in this method. Especially when size of the anchor set increases it is beneficial to develop tools that control anchor quality.

4.1.2. Topic modelling results

Topic modelling results are obtained after subjecting the latent topics to the criteria considered in the previous section. At this point a subset of latent topics is obtained which is small enough to thoroughly inspect and provide with comments. Table 4.2 shows the values found using the assumptions indicated by their columns. Each column represents one of the sections of chapter three. Each entry represents the label given to each value. Labels correspond to anchors for core and generic values. Labels of context topics are instead based on the contents befitting their member documents.

4.1.3. Validity of topic modelling result interpretations

How can topic modelling give us these results? This question seeks to answer the legitimacy of the interpretation of topic modelling results. To reiterate, the topic modelling results are latent topics, their interpretation judges if they are present and gives these latent topics a label (all entries in table 4.2). There are several correct answers to this ranging between illegitimate and sufficing results. This discussion can be reviewed in section 3.1.5. What topic modelling does is identifying the presence of n-gram sets in a corpus. The discussion is about latent topics allegedly representing the labels allocated to them. This discussion is important, because there is faith in science, AI, numbers, machines and whatnot and that doesn't encourage scrutinising its shortcoming. Let us review two perspectives.

First is the perspective that the results suffice and aid the interpretation of the corpus. Second is the perspective where the interpretation of topic modelling results is illegitimate. This is because the nature of the topic model is different from its intended meaning. Recall that topic modelling is an optimisation scheme for the classification of some vectorized text. The machine does not understand the results nor was it made to give purpose to its results. The vectorization does not represent any meaning or understanding. It does not understand different writings for the same word, interpret meanings or relations in texts. Because of this, a topic is very poor in its judgement (worse than a human). In the future a topic modelling may be as good as an average human or even better than the best human in making a judgement (whoever decides that). Even then it continues to be imperfect and bound by assumptions. The more complex the model (adding larger and more technical components and parameters), the harder it is to check these assumptions decreasing transparency. Assumptions propagate into results, meaning it is always impartial. Being transparent about the assumptions is therefore the most important part of topic modelling.

By manually reading the documents of relevant topics a considered judgement can be made. This means that the topics presented in table 4.2 represent their label. Absence of a label from a set means that the topic model was not able to verify its presence. A core value was only included in this result table if the anchor n-grams contributed more than 75% of all mutual information to a topic. A value or label can not be excluded due to the poor quality of language processing.

Core values	Context topics (theme)	Generic values
Best	Investment projects (Governance)	Economic viability and welfare
Flexibility	Natural gas power plants (Governance)	Efficiency
Foresight	Abatement policy (Governance)	Environmental sustainability
Industry	Electrolysis (Governance)	Health, safety and security
Wonder	Solar hybrid technologies (Governance)	
Accessibility and availability	Electricity and hydrogen generation schemes (Governance)	
Control	Proton exchange membrane (PEM fuel cell) (Hydrogen production)	
Development	Bed reactors (Hydrogen production)	
Effectiveness	Green hydrogen steam reforming (Hydrogen production)	
Exploration	Biogas (Hydrogen production)	
Fast	Steam reforming (gasification, pyrolysis and syngas) (Hydrogen production)	
Global	Gasification (Hydrogen production)	
Growth	Hydrogenation (Hydrogen storage)	
	Hydrogen embrittlement (and nuclear reactors) (Hydrogen storage)	
	Hydrogen storage (Hydrogen storage)	
	Alloy storage (Hydrogen storage)	
	Combustion (None)	
	Hydrogen isotopes (None)	
	Photochemistry (None)	
	Methods for atomic (scale) analysis (None)	
	Simulation (None)	
	Space (and aeronautics) (None)	
	Impedance spectroscopy (None)	
	Gas chromatography (None)	
	(Nuclear) fusion (None)	
	Desulfurization (None)	
	Petroleum engineering (None)	
	Hydrogen peroxide (None)	

Table 4.2: Overview of the results of all three sections in a single table. Each column represents a different set of assumptions. These assumptions are reflected through anchor selection. Each column corresponds to the results of each subsection of the results chapter. The first column in this table shows core values, reflecting corporate values. The second column shows the contexts in which hydrogen is discussed. Every context topic is classed in one of the four overarching themes to provide more structure. Lastly, the third column reflects generic values, values that play a role in public discourse.

4.2. Text processing

Text processing occurs before the corpus is fed topic model. This is shown in step 2 of the bottom row in figure 2.2. This regards natural language processing, the process of converting words to numbers.

4.2.1. Filtering stopwords

Manually interpreting topic models involves going through lengthy lists of n-grams where it is observed that most n-grams are irrelevant. An intuitive judgement could tell that the presence of these words could decrease the quality of the topics. It should be noted that most words are already filtered. When a sample of the corpus is taken (of size number of documents). Each word is allowed to occur in no more than (max df) 50% of the documents and each word needs (min df) 10 unique occurrences. Changing this bandwidth does not improve the quality of the results. The only noticeable difference is that the model takes longer to run.

The text vectorization scheme (tfidf) should take care of uninformative n-grams by giving them a low value. Manual lists can be used for filtering domain-specific words before the vectorization. Because the corpus contains multiple disciplines and documents irrelevant to green hydrogen it is recommended to filter these documents instead. Then there is also the risk that stop-word filtering does not change the quality of the results, as documents drive the optimisation mechanic. Filtering a stopword implies the model will use a better n-gram instead. Some disappointment can be expected as the model is likely to use an n-gram of the same or even lower quality to make the model maximise its optimisation criteria instead.

4.2.2. Word usage and evolution

It is important to take into account that the same word can be written in different ways. Various forms of English use different writing for the same word, for example, modelling vs modelling. These anomalies are described and fall within the bounds of prescriptive language. A second observation is that a word is not always written ”correctly”. This goes into the domain of prescriptive linguistics. Incorrectly written words are not interesting as these are most likely filtered out by min_df, the lower bandwidth of the word filter. An interesting case is when a word is commonly written differently (it lacks adhered prescriptions).

A topic model should be able to identify the same word independently of how it is written. This relates to the nature of topic modelling that falls in descriptive and not prescriptive linguistics. Not considering different ways of writing a word effectively excludes articles from a corpus. Understanding how words are used in practice is required to create good anchors and qualitative results. An observation from this exercise is that the same word is written in various ways. Neologisms that happen to be compound words were often observed to fall victim to this.

Consider the term ”stop word” in topic modelling (assuming this is the ”correct” way to write this). A search query in google scholar yields 44 thousand results for ”stop word” and 22 thousand results for ”stopword”. Stop word has twice the hits but also includes ”stop-word” (Google cannot distinguish these two). Therefore it can be assumed that ”stop word”, ”stop-word” and ”stopword” have approximately the same amount of hits. Another 3500 articles use a combination of the three. When this compound word serves as an adjective, such as in ”stop word list” any combination of open, closed and hyphenated compounding will be observed yielding seven unique combinations (which were all found using the aforementioned technique).

This is because there are no rules for compounding (that determine if a word has to be written open, closed or hyphenated). There are only prescriptions for commonly found compounds in the English language. These prescriptions don’t exist for neologisms found in research such as the one in the aforementioned paragraph. This results in non-standard ways of writing. Somehow this also affects words that are circulating for prolonged times with greater reach such as socioeconomic (portmanteau) and policymaker for which the open, closed or hyphenated forms of writing are commonly found. In all these cases, the choice of writing by an author will influence the evolution of the word in language and give shape to potential writing prescriptions.

In the context of topic modelling, better results are obtained if these observations and implications are taken into account during text processing (part of natural language processing). Identifying the alternative writing forms for words is important and should all refer to the same instead of separate occurrences of a word. This refers to words that lack prescriptions and the ones with significant deviations from their prescriptions. Techniques for performing such quantitative operations on a corpus are part of the field of lexicology. Compound words should be identified during tokenisation and an open compound should be treated as its closed variant, which is a single n-gram.

4.2.3. Quantitative analysis and bias

Document value histogram (figure 3.1) documents with zero values either clutter the corpus or contain n-grams referring to our values of interest that are yet unknown. Documents clutter and should be removed because the lower the amount of values per document is the lower the amount of informativeness these documents provide. It could also mean that a poor input set of values are used and that option has to be ruled out first.

Value frequencies (figure 3.2) provide some information on the degree of ”bias” towards each value. The combination of the corpus and vectorization scheme (given it is the TF-IDF) prefers values that are occurring in fever documents (lower value frequency) but when they occur have a high frequency in that document. If these two quantities (inter-document and intra-document value frequencies) are divided a degree of bias (by corpus and vectorization scheme) is attained. Now there is a quantity for each value. See these quantities as a single distribution. The standard deviation of this distribution now corresponds to the degree of bias by the combination of corpus and word vectorization scheme. This bias holds towards the given list of values and implies not that a single aspect, but that the match between corpus, vectorization and value list is poor.

4.3. Model parameters and visualizations

A topic modelling exercise is performed iteratively. Each iteration consists of feeding the model with a different combination of anchors, number of topics and number of documents. For good quality, a single iteration takes approximately 45 minutes to run.

4.3.1. Number of documents

The number of documents affects the number of words in the corpus. It allows the topic model to use more words to make its optimisation, although after some threshold there are enough words to do this nevertheless (approximately 2000 words). Increasing the number of documents increases the consistency of topics appearing giving a certain number of topics and set of anchors. Increasing the number of documents follows the law of diminishing returns. An improvement is noticeable after 6000 documents but is generally not worth the increased run time costs.

4.3.2. Number of topics

The number of topics determines the level of ”disentanglement” the topic model can accomplish. Disentanglement should be seen as cutting the corpus into smaller pieces (latent topics) and assigning documents as members of each piece (latent topic). At maximum disentanglement the topic model can create no more new latent topics to add documents to. In this corpus, there are 550 latent topics. At this level the model can achieve maximisation of its optimisation criterion (maximising total correlation). Setting a number lower than this increases model performance and also merges topics normally disentangled.

Changing the number of topics can heavily affect the overall idea the topic represents. Section 3.2 shows topic range in its results as some topics fail to appear or no longer represent the topic when changing the number of topics used in the model. The reason this happens is that a different number of topics implies a different equilibrium in document topic distribution. Increasing the number of topics lowers the optimal marginal total correlation level for each latent topic. A competing document topic distribution appears to favour ”reshuffling” the documents among topics. Furthermore increasing the number of topics makes manual interpretation harder and more time-consuming. Changing the number of topics might result in a dissatisfactory set of latent topics.

4.3.3. Visualizing topic models

The raw results of the topic model are interpreted and this interpretation is shown in tables and figures. The raw results are sets of latent topics with each having its own set of n-grams and document members. One interesting question is how these results can be visualised in one graph. Doing this on a 2d grid is most easily accomplished by showing how much one topic relates to the other. An intuitive way to visualise this is using a network graph with node size showing the informativeness of the topics and edges how much informativeness each topic has in common with each other. This network graph was developed and helped with the manual interpretation of results by indicating which topics were ”most important” in a specific run.

A similar tool was developed for latent Dirichlet allocation (LDA) topic models. This tool performs a principal component analysis (PCA) on the document topic distributions which form the x and y axes of the graph. Each node represents the size of document topic membership. Clicking on a node shows its most dominant n-grams. This is very useful for interpreting a topic modelling run. It also introduces a new performance indicator only available in the interactive module, λ which has a value between 0 and 1. Lambda is useful for discovering relevant n-grams in a topic. A lambda of 1 shows the probability of an n-gram being part of the topic. A lambda of zero takes into account how often the n-gram occurs in the corpus, favouring rarer words.

This tool was adjusted so it can be fed with data from CorEx, which is fairly similar. An example is shown in figure 4.2. The problem with this implementation is that the principal components of document

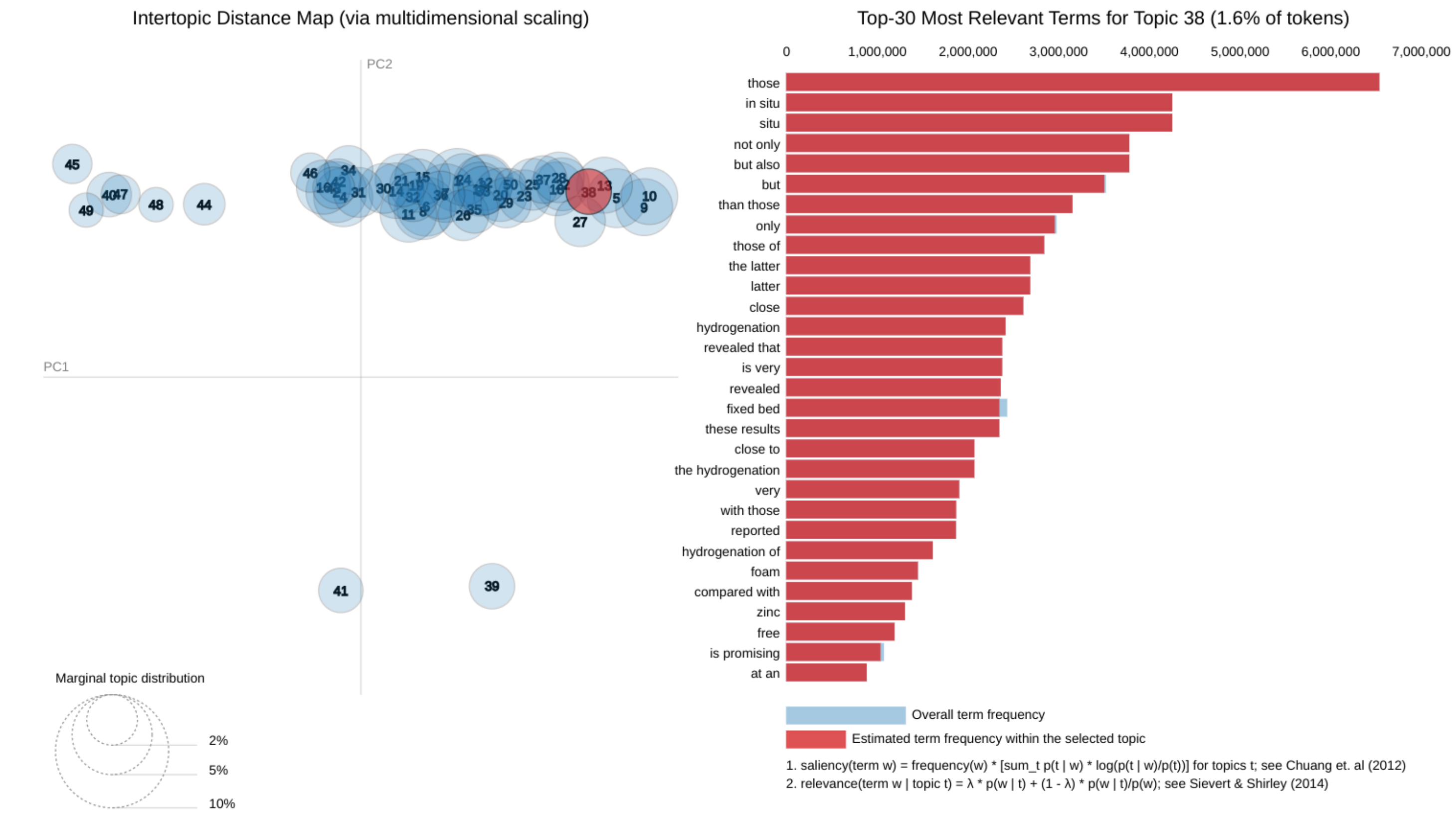


Figure 4.2: Results shown in pyLDAvis interactive visualization adjusted for CorEx

topic memberships tend to allocate all the topics on a single axis. This does not create the same aesthetic satisfaction as in an LDA topic model. This also hints that a PCA is not the desired way of visualisation in CorEx. Another problem arose with the word topic distribution which has unable to show correct values. Getting this module to run with CorEx would improve the quality of the results, makes topic modelling more accessible to a wider audience and improves the workflow of manual topic inspection.

4.4. Values in literature

Values held by stakeholders play a role in hydrogen because they care and have a say in decision-making. Values playing a role in the public debate emerge in this context as a reaction to the proposed policy. Social media and news(papers) propagate (coin, report and spread) values playing a role in the public debate and contentious governance. It is highly unlikely that these values are respected in the policy's design because contention implies that a conflict exists (between values held by stakeholders). The policy is altered and not designed by public discourse. The design of policy is substantiated through best practices and science. Values highlighted by scientific literature, therefore, propel values into initial policy design. This process is done for practical purposes and incorporates the interests of known stakeholders, but doesn't consider objections as plural as in public discourse.

4.4.1. Contentious governance

Governance changes over time and is in some parts of the world, such as the low countries increasingly characterized by contentious governance (see A.2) and stakeholder engagement. Including local communities is considered crucial in the planning of energy projects (van de Grift et al., 2020). Governmental players, such as ministries, municipalities or government officials are interpenetrated with social movements in a regime of contentious governance. This means that the support of social movements is required to legitimise political action. Identifying which values play a role allows governmental players to engage proactively with these communities. Values can represent the interests of actors and can, if left unaddressed, cause the mobilisation of social movements opposing these objectives causing delays, blocks or policy failure. Aspects of the cultural sphere, such as image, perception, norms, values, knowledge and objectives should, contrary to tangible matters, be a major concern in the adoption of green hydrogen.

4.4.2. Public perception of hydrogen

Contentious governance can be observed in European politics through social movements against windmills, nuclear energy, fracking, carbon capture storage, 5G, corona measures and farming policy. These social movements have a significant impact ranging from delaying policy (farming policy) and disturbing economic activities (economic sabotage) to blocking policy (fracking) or discontinuing existing policies (nuclear energy in Germany). Mobilisation and demobilisation of social movements are governed by cognitive processes, such as group identity, morality and perception of language. Rationality is one of the cognitive processes used as a tool in the mobilisation process and therefore plays a minor role. Values perceived to play a role in hydrogen are levers for mobilisation. Stories are used to mobilise values and can have any rationale making rationale arbitrary and complex.

Mobilisation of social movements supporting and opposing green hydrogen will happen as it drastically affects households, communities, industrial sectors and consumption patterns. Hydrogen is already subjected to the sociopolitical organisation, as the green hydrogen backbone is a composite actor of governmental and corporate players. Injecting hydrogen in reservoirs affecting local communities, distribution of economic costs of switching from gas to hydrogen and industries affected by a carbon tax are the most obvious stories used to mobilise social movement opposing green hydrogen.

4.4.3. Core values in literature

In this research two main ”value” sets were used. The one of generic values is a reflection of values held by the public and core values reflect values belonging to corporate ideology. The anchors that were used should be seen as a possible example of these sets and these groups. A problem is that it is not known what values are held by relevant stakeholders. Having a set of anchors where groups of stakeholders agree on is desirable, because then there is something to work with. There are several problems with this. First is that it disregards that values are not static. Values are secondly arbitrary and can be or change into anything. This will results in a huge set of possible value states, this probably relates to linguistics, social cognition and affective cognition.

We have considered the possible value states now let’s address the expressed values. An expressed value is something that is written down or communicated to other people. The problem is that these values not always correspond to the value that is held internally. Expressed values sometimes correspond to goals or aspirations (Lencioni, 2002). In these cases the expressed value, such as cooperation, transparency and efficiency indicates the lack of this value and the desire to attain this. Expressed values do not necessarily reflect internal values or the situation as it is.

In the context of green hydrogen policy, strategic communication comes in to play as there will be opposition to plans and ideas. As an example for strategic communication, take a recent newspaper article reacting with indignation on the actions of some petroleum company that knew about climate change all along, but chose to suppress this information instead. Both the suppression of this information and reaction by media is a form of strategic communication. There is only one relevant value however and that is corporate profit maximisation. It is however not possible to express this and to avoid controversy strategic communication is chosen. Expressed values are in this context a strategic decision part of a desire to satisfying some internal values or objectives.

To return to the question which values play a role now has different dimensions. One the one hand there are values that play a role internally. To clarify an internal value are values and objectives that truly play a role which are not necessarily expressed. Internal values are not exposed by using documents that are publicly distributed. Values found in the documents ideally reflect this, but as discussed is often not the case. There is a tendency to write these in a positive quality. Instead the opposite may be true, such that the value reflects objectives, desires or aspirations. The internal value is opposite of the expressed value. The value may also reflect a strategy of the author. The intention of the author may be a response by, a belief held or knowledge gained by the reader.

Nevertheless, the most challenging part of this research was getting results. Then interpretation plays a secondary role as this would overshadow that which was required to obtain the results in the first place. Interpretation and judgement took place regarding if that value is actually present in the document. These values were presented as they were found. Further interpretation of what they may truly mean is something that should be left open for the reader.

5 Conclusion

5.1. Conclusion

In this research, three different topic models were made to help answer the research objective. To recall this objective was **Which values play a role in scientific literature on green hydrogen?** and will be addressed before turning to the main research question at the end of this section. Results of these topic models are shown in table 4.2. This shows corporate values, the context in which hydrogen is discussed and public values. Context topics have a decent document topic fit. It suffices in its ability to identify and distinguish green from non-green hydrogen. Furthermore, the set of all the context topics appears to represent the corpus. Unlike core and generic values, context topics represent latent topics. Latent topics (or hidden topics) are yet to be discovered topics where the discovery is made by the machine learning algorithm.

Core and generic values are not discovered, but found using a predefined set of anchor words to detect their presence in the corpus. The construction of these topics is, unlike core values, heavily dependent on assumptions made regarding the interpretation criteria and the anchor words. Furthermore, results show a poor document-topic fit. First, this means that the documents on these topics barely pass the bar to be included in that topic. Documents are not primarily concerned with that topic, but relate to it in some way (and thus making it pass the bar). Second, a significant fraction (typically more than 50%) of documents part of a topic have nothing in common with that topic. Last is the instability of the reported results of core values. This instability pertains to running and automatically interpreting the model. This instability is characterised by a different result on each run. This instability means that the α signature is varying each run. This can imply that there is no single representative run for any topic model including the ones that are manually inspected. Instability is exacerbated mostly by increasing the number of topics, decreasing the number of documents and decreasing the number of anchor words per anchor (given that the anchor word is in the corpus). An ensemble run is an attempt to alleviate this problem but no solution for this problem as it produces an average topic based on sets of imperfect topics.

A comparison between data analysis and topic modelling results can be made for core values. There is a clear difference in the results of the data analysis (table 3.1) and the topic model (table 4.2). This difference is caused by the way values are quantified. In data analysis a value counts if it is mentioned at least once in a document. In topic modelling the anchor words and interpretation criteria influence the topic results. The difference in results between data analysis and topic modelling is furthermore influenced by the way ”informativeness” of words is measured, which is dictated by the TF-IDF (see 2.2.4).

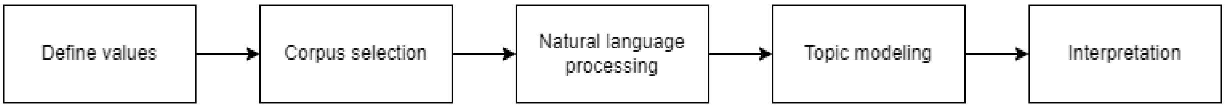


Figure 5.1: Steps in the methodological framework for finding values using topic modelling

A methodological framework for finding values using topic modelling helps with understanding why we get these results. This systematic approach is a possible answer to the main research question which was **How can topic modelling be used to interpret what values are related to hydrogen technologies in scientific articles?**. This framework is shown in figure 5.1 and the subcomponents of this framework are discussed in this conclusion. A clear overview of this framework and each component is in appendix C.

5.1.1. Subquestion: How can it be determined which values are the most important or relevant?

The first step in the topic model framework of figure 5.1 is defining values. This step involves listing the states and behaviours that are desirable according to a subject. This results in a list of values. Each value has to be written down in the form of individual words or groups of words. Closely related terms and synonyms of these words and groups of words should be included in each value. Either an individual, group of individuals or culture can play the role of the subject in this context. When a group is taken the member's degree of entitativity to that group becomes relevant. Values held by stakeholders are the most relevant as they influence contemporary decision making. Other values are also valid as input. Using the temporal dimension is a way to look at different forms of value definitions (see 5.1.1 on time-dependency). In order to find a value in a corpus a comprehensive list of words related to this value is required. These words are called anchors which are used to steer the topic model.

When values and their anchors are formulated it is important that these values and words can be found in the corpus, the second step in the framework. Corpus relates to the nature of documents used as input. There is a different sort of information embedded in scientific literature compared to news articles or tweets. Scientific literature is for example more deprived of emotion and subjective judgement than tweets. Therefore, corpus selection depends on the sort of value that is sought after.

Language processing, the third step, helps with finding these words. It relates to all sets of operations performed on text data and transforming text to numbers. The used language processing in this research are word filtering and word vectorization. For word filter a bandwidth filter was used to filter words based on their absolute and relative frequencies. A word vectorization is converting words into number and the TF-IDf scheme was used. Basic stopword filtering is arguably inefficient, because the word vectorization takes word frequencies into account. If some topics are undesirable in the results, don't use a standard list to filter these words. Use that list to filter documents instead (see section 5.4.2). More sophisticated Language processing was not performed in this research, but is required to properly identify values. Language processing helps the computer with providing information in which context a word is discussed. Since this is not the case many topics (in core values) suffer from the problem that they are taken out of context, consequently not referring to the associated value that was intended in the first place.

All in all value definition precedes and determines corpus selection and language processing steps of the framework (figure 5.1. In this research language processing and corpus were remained constant to illustrate the profound effect of anchoring and the used assumptions to find these values. With the use of semi-supervised topic modelling, the previously defined values can be ”anchored”, this means that the topic model is steered to favour identification of these values. This is explained in section 5.1.1 of topic model parameters. A complex result of the topic model requires systematic interpretation of the results 5.1.3. Manual inspection of documents is required to determine if the document relates to the topic. If that is the case then the topic is present. By utilising these five steps it is possible to identify which values play an important role in a corpus.

Time-dependency of values

Values can be constructed in any desired way. An important dimension of values is time-dependency. Let's illustrate this using three perspectives: the past, present and future. When engaging with values in the past a historic perspective has to be used. What are the conditions that caused people to think this way? A perspective in the present is engaging with stakeholders, because they influence the contemporary decision making and shape the future. The most relevant question is who are the stakeholders and what are their subjective perceptions? This has also been the perspective in this research using core values to simulate corporate ideology and generic values for public perception.

However, the time-scale for policies such as green hydrogen is over several decades and the values held right now do not hold over this time. Figure 3.3 shows that within three decades a newly introduced word (robust) can be come the most dominant word in the most dominant core value topic (Reliability). This figure is backed with too little data to corroborate this claim, but it illustrates the effect of changes in how values are expressed. It's guaranteed that events in the next decades will change the perceptions and opinions on green hydrogen making the used value sets a non-robust solution as it is unable to deal with this shock. While the future is inherently uncertain, it can be estimated what plays a role in the future using scenarios, for example, thriving, stable or degenerating society. In this way future contention can be predicted since topic modelling can help with illustrating what is and what is not being discussed right now.

Intermezzo: Topic model parameters

The third component in the topic modelling framework shown in figure 5.1 relates to its parameter settings. Anchors and number of topics are the most important parameters in the method used in this research. Anchoring is specific to semi-supervised topic models, the method used in this research and allows steering of the model. The number of topics determines the detail and is found unsupervised topic models too. Anchors are used to steer the topic model, thus specifying our bias. To understand what an anchor is, some basic understanding of the topic model is required. A topic puts documents into classes (topics). It does so by learning which word (n-gram) belongs to which topic. Under normal conditions this is expressed by a value between zero and one (unit interval). This value is updated during each ”learning step” of the topic model. When an n-gram is anchored it is fixed to a value of anchor strength after each learning step which is typically higher than 1. In this research a value of three is used. This means that the information contained by the word will count three times more than normal words and that it is not affected by learning. It will always influence the topic formation process.

The amount of detail in the results is determined by the number of topics. This level of detail is the degree of disentanglement (of the corpus). A fully disentangled result is obtained when empty topics appear as all documents are attributed to all possible topics already. Total correlation, the optimisation criterion of the topic model, is maximised in this state. This does not imply that this is the optimal result. Varying the number of topics not only changes the level of detail, but also the content and meaning of the topic. When the number of topics is changed a completely different result is obtained as it splits, amalgamates and merges existing topics. The number of topics not only affect the detail, but also the topic content. Some final choice in number of topics thus reflect desired topic contents.

5.1.2. Subquestion: How is green hydrogen distinguished in the corpus?

Identifying green hydrogen in the corpus was necessary to answer the original research question since the corpus regards hydrogen in general. Recall that a corpus is the set of documents that is used in a topic modelling exercise. Section 3.2 was dedicated to identifying the contexts in which green hydrogen is discussed. Among other things, an unanchored CorEx topic model appears to favour identifying these contexts. The other (irrelevant) topics in that section can be considered noise, while topics that conformed relevant context were fed back into the model as anchors. A topic can be fed back into the model because it consists of member words and documents. The words (n-grams) were fed back. Examples of noisy topics are those that are made up of years, anaphoras or nothing coherent at all.

This method of anchoring worked because the CorEx topic model was able to identify some topics in an unsupervised setting. Anchoring and changing the number of topics are the most relevant parameters that affect the results given an input. The influence of introducing anchors, changing to a semi-supervised setting, is minimal on the results. While it is possible to obtain any result via anchoring, it is unrealistic to have more than 10% of all the n-grams anchored. The model appears to have some ”ideal” state it wants to converge to independent of the appended anchors. The results are thus primarily determined by inputs: the chosen corpus and text processing procedure (See section 5.2).

5.1.3. Subquestion: How can topic modelling results be interpreted?

A systematic approach has to be taken provide a when interpreting the topic modelling results. The results of a topic model are expressed in two distinct matrices: a θ see figure 2.4 and an α see figure 2.5 signatures. Interpretation of these matrices results in something more practical. Two methods, the automatic and manual method, were developed for interpreting these results. A systematic approach is required to ensure reproducibility of the results. Furthermore, without a systematic approach interpretations of different topic model exercises are incomparable, because the logic with which the interpretation was done is inconsistent. Discussions over or observation of different results can be explained by different interpretation criteria, which has nothing to do with the content of the corpus.

Values are verified by the topic model and its interpretation. If a value was not found it is not excluded. While it is possible to check for false positives, it is not possible yet to check for false negatives. There are sufficient alternative reasons for why a value is not identified, such as poor corpus choice, language processing. These two aspects can severely hamper the ability to successfully identify values.

Set of values obtained from different modelling exercises are not comparable to each other. The results of a topic modelling exercise only reflect the corpus that was chosen. This could roughly correspond with what purpose the documents in this corpus were written. Secondly, sets of values relate to the set of prior assumptions, those are the values that we were looking for. It thus does not relate to other sets of values. This means that the results of the three sections cannot be compared to each other.

Intermezzo: Systematic interpretation

The systematic approach is performed by going through the words belonging to each topic, corresponding to the α matrix. If these words correspond in some way to our ”relevant values”, then this topic passes to the next round. If not it is discarded. In the second round the documents belonging to the topic are analysed. This corresponds to the θ matrix. If these documents fit in our ”relevant values” then it can be said that this specific value is present. While it is on paper possible to consistently perform both tasks, it is currently not possible.

We will discuss how this was done, starting with the α matrix. In section 3.2 the strongest and most relevant n-grams were used to construct anchors. While the model started without anchors, each run added and updated existing anchors until a satisfying result was obtained. In section 3.3 criteria were developed for judging topics. This was done manually and is called the ”manual approach” in this research. Lastly section 3.1 used an algorithm to check this. As a criteria the words part of both the topic and their corresponding anchor were required to contribute more than a certain amount (75%) of informativeness compared to all other words in that topic. The problem with this method is that it assumes that the number of other words contributing to the topic is very small. It also assumes that the relevant member words are known prior. Despite this rigidity it is more consistent, faster and less prone to error compared to the manual method. An automated interpretation is favoured over manual, especially when the number of topics and n-grams increase a manual method becomes an unrealistic option.

Inspection of documents, the θ matrix can only be done manual. This step verifies if the presumed topic truly corresponds to our value. This is done by manually reading the documents that are a member of our topic. If the content of the documents corresponds to our presumed ”value” then this value is present in the corpus. In this verification step it is only possible to check for false positives. This step provides insight by providing more detail to the topics. Automating this step reduces the insight of the model and arguably a verification step. It then no longer complements, but leads a human, by telling how things are.

Intermezzo: Cover image

A parable or allegory helps with illustrating the value of a topic model. Comparisons can be made with the parable of the blind men and an elephant or Plato’s allegory of the cave. The latter was chosen as cover image as a warning to not take the topic model too serious as an authority. This comparison was first made in the methodology of this research. In Plato’s allegory a cave is filled with prisoners who haven’t seen daylight, therefore their perception is fundamentally different than ours. A series of objects are passed in front of a flame so that they can see its shadow on the cave wall. The prisoners are similar to a topic model and the objects that are passing by are similar to a processed corpus containing documents and n-grams. Each object can be seen as a document. Then a series of objects is the corpus and the shape of each object is an n-gram. After being shown a series of object the prisoner is tasked to put the objects into a number of categories. The purpose of this story is relativising the value of the numerical values given by the topic model.

5.2. Limitations

There are various possible reasons for why unexpected results are obtained and these constrains are addressed in the framework. Blame can be given to: wrong corpus choice, quality and suitability of language processing technique, choice of topic modelling technique, choice of assumptions and anchors for supervised models, quality of result analysis, bias of researcher and process of determining relevant values. An important asterisks to the results in general is that they only relate to the input values used. A stakeholder therefore has to agree on the input values. The other steps in the framework relate to the ability to identify these values.

Limitations of results through text processing choices

Not only does the topic model not understand the documents, the language processing step that preceded topic modelling is basic and limits the ability to properly identify contents and values. This language processing was done in python using the scipy package and is called term frequency inverse document frequency (tf-idf). The tf-idf may not be the best scheme for a corpus where short texts are used with the purpose of identifying values. Tf-idf favours words that occur frequently in a single document. For documents abstracts and titles were used. The resulting ”article” is high of context as words like ”storage” or ”hydrogenation” are typically present in the title and at least once in the body of the abstract. ”Values” are of much lower frequency in scientific literature. ”Fairness”, ”reliability”, ”impactful” and ”cooperation” are typically not present in a title and occur maybe once in the abstract. Furthermore if the purpose is to find values then it could be considered to shift the focus from topic modelling to language processing. Values could be identified by inferring relations, and the presence of groups of words. Using such an approach is more time consuming but potentially better, because it allows the usage of linguistic theories instead of statistical ones.

Let us consider a second choice that preceded the topic modelling, the corpus choice. Using scientific literature as a corpus is a poor choice for identifying ”values”. In the ”culture” of scientific writing it is discouraged from making ”bold statements”, because they imply other things apart from the factual and that is distracting for the purpose of scientific communication. As an example, when looking for values as was done in section 3.3 then a corpus with titles such as ”comfortable transportation for a fair society in a renewable world” is preferred over ”evaluating applications of PEM-fuel cells”. The former would generate a much higher information signal which is needed to create good topics. It may just as well be the case that both hypothetical articles talk about the same thing, but use a different phrasing.

5.2.1. Limitations in execution

Bias forms the basis for corpus choice, topic model choice and anchor choice. Putting a different person in charge of a topic modelling operation will create a different set of results. Potential differences between researchers are very significant, because assumptions determine the output. To illustrate this, table 4.2 shows the differences if anchors and analysis method are varied. The most important variable are the values or anchors. While leaving these static, it is more realistic to make comparisons while changing other parameters. Lastly on the purpose on modelling its important to emphasise that values are verified and not excluded.

5.3. Recommendations

This section discusses the impact of this work for policy makers and how this technology can be used right now. It also suggests the potential future impact of the technology on society. This discussion is held on a broad level, while the next section suggests improvements for future work on topic modelling.

5.3.1. Discourse surrounding AI

There is discourse relating to ”machine learning” and ”artificial intelligence” (ML&AI) and their effect on society. First it has to be mentioned that ”machine learning” and ”artificial intelligence” are both subsets of algorithms, while formal definitions of machine learning and artificial intelligence are more complicated. Therefore the use of these terms is incorrect or inappropriate at times. Topic modelling is a machine learning algorithm, but it is not artificial intelligence, because a topic model does not mimic intelligence. Iteratively performing a calculation and storing the intermediate values is in my opinion not a form of intelligence. Doing topic modelling makes it clear that the algorithm doesn’t know what it is doing. Here I want to suggest the reader to take a look at the cover image. The results of such a topic model may mimic the results generated by a human. Now we can turn to the point of the discourse surrounding ML&AI which is about the consequences of the dramatic increase in computational capabilities of machines (computers) and the techniques (algorithms) that manipulate this computational power. Assuming this trend continues, then topic modelling will be able to what plays a role in specific subject, where people are talking about and how interests and viewpoints of groups and specific individuals change over time. What are the consequences of these technologies and how does society interact with this?

Regarding ownership of topic models

Let us first consider the owner's side of things, which regards ownership of the algorithm, platform and dataflow. Access to this platform can be constrained to allow access to select individuals (to provide a service) or public access (for collecting data). Access to select individuals is chosen if it provides an exclusive service or to increase competitiveness of an organisation (for example companies or governmental institutions). Wide or open access to a platform is chosen as this generates large amounts of data to be collected. This data is then used to improve the service and it can safely be assumed that it is sold (to make the service possible in the first place). The power gained by ownership of the platform is that it can shape the perspective of its users and therefore their objectives and behaviour. Ownership of the data gives insight into the values and beliefs beliefs held by the people using the algorithm. It also indicates what what people find interesting over time. Not only is this very useful information, it can also be used to proactively shape the results in order to elicit specific behavioural responses. Such operations would not be hindered by privacy regulations, such as the ones based on the European GDPR, that are concerned with individual data subjects because the aforementioned operations use aggregated data and are irreducible to individuals. Such developments do not hint, but screech that (private) ownership of such a platform, algorithm or data flow is undesirable.

Regrading users of topic models

Next is the user side of things. Topic model has the ability to change the perception of users about a specific topic. If topic modelling is used to highlight what values play a role it forms an opportunity to support research that fills this gap. Topic modelling can be used as a tool of reflection regarding disciplines: why and should these values be discussed? On a very broad level, topic modelling has the ability to transform the way in which information is retrieved. If the quality of topic models exceeds that of traditional ways of retrieving information, then topic modelling will replace this traditional task. This can be tasks, such as using search engines, doing literature research or reading the news.

5.3.2. Utilising topic modelling for finding values in a policy context

Two situations are where topic models can be used. The first situation is for policy makers that wish to be more informed regarding a specific subject. A topic model can be used to detect for if values are discussed regarding a specific topic. What the presence of a value means is a question that goes beyond ”finding values”. It has to be noted that absence does not mean that the value is irrelevant but that it is underexposed instead. Values used in this activity relate to groups. It may therefore be wise to invite stakeholders and use the values that matter to them as an input to the topic model. The goal of this activity is to verify whether their values are discussed in literature or not. It also assumes that stakeholders are ”cooperative”: open to reason and make concessions to the other party, because there is a perceived benefit of cooperation. On the opposite side there is a ”defective” stance where stakeholders are bitterly entrenched seeking to exhaust all possible political and Judaical procedures to forward their own interests and obstruct the other. In such a situation topic modelling can be used for mediative, democratising and value clarifying activities.

5.3.3. Other uses of topic models

Topic models can be used for various other purposes. First is its use in society. Trends on social media can be used to identify trends and to what demographics these play a role. This is useful to prepare and timely reacting to future trends. This is especially useful for the youth as they extensively use social media. Example use cases are identifying trends in popularity regarding knives, books, online fraud, drug trade and designer drugs. Another use is in foreign policy where topic modelling can be used for detecting values in a corpus of speeches by a politician. This is useful to indicate a general direction of a nation; more important is identifying the values held by policy makers surrounding (authoritative) decision-makers as this drives policy. Topic modelling can be used in foreign policy to develop roadmaps, identify changes in policy and help answer questions about what will happen if there is a change in leadership or inner circle. Topic modelling can be helpful to identify sentiment in general. This can be used for various things, such as research business or to measure trust in institutions.

Improve understanding of values

Topic modelling was applied in the context of ”identifying values”, but values are poorly understood. An unanswered question is what the values in science are. Having a list of relevant values for specific purposes is beneficial. Values are relevant to a group or an individual and engaging these parties in the process of topic modelling is a solution for listing the relevant values. Knowing how to identify values in a document is also important and this is what is meant by understanding values. Understanding values relates to obtaining all the n-grams that are relevant to this value and are needed to properly identify this value in the corpus. At present, the model uses n-grams from tokenised texts. Some proper language processing allows the inferring of relations between words in text. This is very important when trying to find values. It could even speed up the model by reducing inflections and improving the quality of n-grammisation. This should all lead to better anchor sets. Understanding of values should ideally reach a point where some humanly defined input can be used to call a relevant set of anchors for any given corpus, which falls in the context of automated topic modelling.

5.4. Future work

The development of this topic model should aim at improving the quality of the results and afterwards improving the workflow. This is to increase the reliability and speed at which results can be produced. At some point, workflow improvements increase the accessibility of topic models to a wider audience. Developments can elevate the role of topic modelling where it can lead and elevate individuals. For now, its purpose is limited to identifying the presence of n-grams in a corpus. Results of a topic mode run (an instantiation) should be analysed, n-grams updated and fed back into the model. Ensemble runs should allow the comparison of results between instantiations and different model parameters. Ensemble topic models do exist for the unsupervised model but were not found for semi-supervised topic models such as CorEx.

This section discusses the future improvements to topic modelling, namely the quality of results and the modelling workflow. The first part discusses concrete accessibility, speed and workflow improvements. The idea of improving the workflow is that everyone should be able to do topic modelling. This accessibility is attained by removing the programming skill gap, decreasing the number of steps to obtain results and removing the system requirements. The second part addresses model optimisation and improvements in the quality of the results. The aim of improving the quality is to make the results more reliable in the first place. Enhanced quality also allows the development of new topic modelling implementations. Think about selecting sub-corpus to perform topic modelling, combining sets of assumptions or automating parts of the topic modelling procedure in the future.

5.4.1. Topic modelling workflow

Topic modelling workflow was an issue as it proved to be time-consuming and inefficient. Latent topics in sections 3.2 and 3.3 were all analysed manually. This requires keeping track of all anchors and their n-grams while checking each latent topic and its n-grams. This is okay for smaller anchor sets and latent topics (under a hundred). But it became impossible when the number of anchors and latent topics ramped up in section 3.1. Automation was required to obtain the results of section 3.1.4. This algorithm was able to pick up things I was unable to see. It is arguably the preferred method because it is faster, less prone to error, less biased and more consistent. It uses mutual information as a performance index making the results unpredictable and abstract however when compared to a manual method. Automatic interpretation of the results was nevertheless a huge success.

Other ambitions are increasing the accessibility to topic modelling and improving the quality of (automated) results. It was thought that accessibility is increased through deploying the model on a server. This removes the skill gap as an email with anchors and a number of topics sends you the results. It doesn’t require a python installation and dependencies nor will it cause a memory overload.

New workflow

Topic modelling workflow was adjusted and improved significantly. The initial workflow consists of a Jupyter notebook where each cell was executed manually. Each cell performs some imports, data processing, running the topic model and analysing the results. Frequent kernel death due to memory overload made this process inefficient. It was found that executing this code from the terminal (for the same settings) did not cause kernel death. Increasing settings eventually caused the same issues on the terminal setup. This was caused by memory issues and was the driver for workflow optimisation. Code was rewritten into python objects and functions, results were systematically logged and a server was deployed to run the code. Much later an algorithm was developed for the automatic interpretation of latent topics.

The new workflow was obtained by rewriting the code into objects, making it executable from the terminal and systematically logging results. Executing and logging results now happens in one line of the terminal, the pipeline python script. In this script one only has to specify anchors, the number of documents and topics. It also supports executing multiple models in succession for different settings (a precursor to the ensemble run).

The script was rewritten in an object so that all important parameters can easily be accessed and stored. Each run is stored and contains the important parameters for later analysis in pkl format, some figures characterising the run, and model results in csv format. Each model run is saved under the Unix time stamp so that every run is unique. During model run, the script now shows how long each step took and what the model is doing. It checks for your anchors in which latent topics they occur and the percentage of informativeness they represent. Too many parameters are stored resulting in a huge pkl file. Some optimisation has to be done if to make further ambitions possible.

Server deployment

The topic model was run on two computers and various limitations saw the need for the development of a server. One topic model application was run on windows 10 and the other on a containerised Linux system in Chromebook. The first limitation was performance issues on the Linux system which showed memory overload. The second issue is managing versions across two computers. Package management became an issue due to the use of custom packages, and differences in python versions and operating systems.

An issue arose with respect to available ram as higher numbers of topics, words or documents eventually caused a memory overflow preventing the model from being finished. The Linux used an Intel Celeron N4120 @ 1.10GHz with 4GB of ddr4 ram and the windows used an Intel Core i5-3570K @ 3.40GHz with 32GB of ddr3 ram. The windows system typically runs at twice the speed of the Linux system on both an internal hard drive and solid-state drive. This 4GB of ram would start to overflow with 250 or more topics and 10 thousand documents. Lowering the minimum number of word occurrences for passing it into the vocabulary, min_{df}, from the default value 10 to 4 created memory overflow for the lowest number of topics (50) and documents (1000).

One of the ambitions was to make topic modelling more accessible. The idea is that an email with anchors to the server would result in a reply with topic modelling results. With this tool, everyone with different ideas on what plays a role in the corpus can have their say by feeding their own set of anchors and be given their own results to interpret. The server was deployed using docker on amazon web servers. Eventually, the model was hosted on the server but had problems handling input and output. Doing this without any prior knowledge was a time-consuming endeavour. It was therefore aborted ending up as a wonderful learning experience.

5.4.2. Topic modelling quality

Filtering documents

Filtering documents decreases the number of documents in the corpus. This is much more impactful than filtering stopwords. Filtering stopwords is arguably a poor solution because it doesn't remove documents that are unrelated to the topic influencing the results in an undesirable way. There are various methods for document filtering. Documents can be filtered based on the presence or absence of certain words. Another method is filtering specific documents based on a posteriori knowledge of the corpus. Section 3.2 provides us potentially with documents and their identifiers. This way a small subsection of the corpus can be selected. With this technique, one can create any subcorpus to perform a new topic modelling exercise on. Such results are usually obtained over multiple iterations this requires comparing and performing calculations of results between multiple runs (see section B.1).

Document-topic trade-off

A trade-off exists between choosing the number of documents and the number of topics in a topic model run. During each run a number of documents are drawn from the corpus and only this information is used to create a number of latent topics. In general, the number of documents is typically related to quality and the number of topics to the level of detail of the results. Quality means consistency of the topics across runs. Detail is the ability of the model to extract all information it can find in the corpus. It is assumed that the number of latent topics is ought to ensure disentanglement (maximum level of detail). This means that both topics and documents should be maximised, but both increase model runtime.

Runtime can quickly explode (non-linear increases) if these two parameters are not kept in check. The model performs various matrix operations with sizes equal to the number of words (n), documents (i) and topics (j). Words and documents are closely related, increasing the number of documents also increase the number of words. Within the topic model, words and documents are treated separately and both increase the calculation time. Various matrix multiplications are performed involving matrices of sizes $j \times n$ and $i \times j$. It is not known how the model performance is affected by these parameters. Understanding this helps with the optimisation of model parameters and planning the execution of a set of iterations. A set of iterations is desired because it helps understand how the number of topics and documents affect the results. Not only does an iteration provide more insight; but it is also required to ensure sufficient quality of the results.

Poor choice of the number of documents and topics can cause aliasing. An alias is a topic which appears due to poor document draw. In this case, a topic appears because of the combination of documents that were accidentally drawn. Increasing the number of topics also decreases the number of documents members of a topic. Thus increasing the number of topics not only increases the number of aliases but also increases the fraction of aliases. This shows that the probability that documents are members of latent topics is an apparent probability. The real probability, that a document is a member of a topic, can only be obtained by analysing the topics over multiple runs. Section 3.1 used a high number of latent topics and a low number of documents. The observation that results were not reproducible is caused by this aliasing. Aliasing can not be prevented for these settings, thus choosing a single representative run is in this case false.

Ensemble run

A single model produces a single probability that a document is a member of a topic. And while topic membership is only allocated for high θ values of 99% or higher it is only true for a single run. This means that per run (of 6000 documents, 400 topics and 5% average membership) 1200 articles are incorrectly allocated. This is only for a single run among many possible configurations. Increasing the number of observations can reduce this uncertainty. This is useful for reducing uncertainty in the quality of automatic interpretation, it allows reliable classification of documents in sets of topics, creating subsets of the corpus and allows comparisons of multiple sets of assumptions to each other. Think about comparing combining section 3.2 and 3.3, for example exploring values in hydrogen production, when talking about multiple sets of assumptions.

An ensemble run is the combination of the results of a set of topic modelling instances which increases the quality of the results and the speed at which these are attained. Quality is improved by mitigating aliasing and approximating the ”true” α and θ distributions. Speed is improved as several model instances with a lower number of documents are executed at higher speed while providing a higher result quality. Implementations of the ensemble run already exist (for LDA). The main limitation of these implementations is the way in which latent topics between model instances are compared to each other. Designing new algorithms, comparing existing ones and assessing their influence on the final set of latent topics is a possible next step.

Bibliography

Blanchette, S. (2008). A hydrogen economy and its impact on the world as we know it. *Energy Policy*, *36*(2), 522–530. https://doi.org/10.1016/j.enpol.2007.09.029

Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases [Number: 1 Publisher: Palgrave]. *Humanities and Social Sciences Communications*, *8*(1), 1–15. https://doi.org/10.1057/s41599-021-00903-w

Collantes, G. O. (2008). The dimensions of the policy debate over transportation energy: The case of hydrogen in the united states. Retrieved October 19, 2022, from https://escholarship.org/uc/item/91f3d1ns

de Wildt, T. E., van de Poel, I. R., & Chappin, E. J. L. (2022). Tracing long-term value change in (energy) technologies: Opportunities of probabilistic topic models using large data sets [Publisher: SAGE Publications Inc]. *Science, Technology, & Human Values*, *47*(3), 429–458. https://doi.org/10.1177/01622439211054439

de Wildt, T. E., Chappin, E. J. L., van de Kaa, G., & Herder, P. M. (2018). A comprehensive approach to reviewing latent topics addressed by literature across multiple disciplines. *Applied Energy*, *228*, 2111–2128. https://doi.org/10.1016/j.apenergy.2018.06.082

du Plessis, C. (2011). A computational text analysis of the south african banking sector's representation of its core values: A corpus-driven approach [Publisher: Routledge _eprint: https://doi.org/10.1080/02500167.201 *Communicatio*, *37*(3), 422–442. https://doi.org/10.1080/02500167.2011.609994

Fox, R. (2006). USING CORPUS LINGUISTICS TO DESCRIBE CORPORATIONS' IDEOLOGIES. *Tourism and hospitality management*, *12*(2), 15–24. https://doi.org/10.20867/thm.12.2.2

Gallagher, R. J., Reing, K., Kale, D., & Ver Steeg, G. (2017). Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, *5*, 529–542. https://doi.org/10.1162/tacl_a_00078

Gopalan, A., & Tyagi, H. (2020). How reliable are test numbers for revealing the COVID-19 ground truth and applying interventions? *Journal of the Indian Institute of Science*, *100*(4), 863–884. https://doi.org/10.1007/s41745-020-00210-4

Jens, J., Wang, A., van der Leun, K., Daan, P., & Buseman, M. (2021, April). *Extending the european hydrogen backbone*. Retrieved March 2, 2022, from https://gasforclimate2050.eu/wp-content/uploads/2021/06/European-Hydrogen-Backbone_April-2021_V3.pdf

Koopmans, R. (1999). Political. opportunity. structure. some splitting to balance the lumping [Publisher: [Wiley, Springer]]. *Sociological Forum*, *14*(1), 93–105. Retrieved April 12, 2022, from https://www.jstor.org/stable/685018

Lencioni, P. M. (2002). Make your values mean something. *Harvard Business Review*.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2), 93–118. https://doi.org/10.1080/19312458.2018.1430754

Maltseva, K. (2018). Values, norms, and social cognition. *NaUKMA Research Papers. Sociology*, *1*(0), 3–9. https://doi.org/10.18523/2617-9067.2018.3-9

Peters, D., van der Leun, K., Terlouw, W., van Tilburg, J., Berg, T., Schimmel, M., van der Hoorn, I., Buseman, M., Staats, M., Schenkel, M., & Ur Rehman Mir, G. (2020, April). *Gas decarbonisation pathways 2020–2050*. Retrieved March 3, 2022, from https://gasforclimate2050.eu/wp-content/uploads/2020/04/Gas-for-Climate-Gas-Decarbonisation-Pathways-2020-2050.pdf

Pettinicchio, D. (2012). Institutional activism: Reconsidering the insider/outsider dichotomy [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-9020.2012.00465.x]. *Sociology Compass*, *6*(6), 499–510. https://doi.org/10.1111/j.1751-9020.2012.00465.x

Popper, S. W. (2002). Technological change and the challenges for 21st century governance. *AAAS Colloquium on Science and Technology Policy*, *27*, 83–103.

Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in AI [Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/21507740.2020.1740350]. *AJOB Neuroscience*, *11*(2), 88–95. https://doi.org/10.1080/21507740.2020.1740350

Schimmel, M., Peters, D., & van der Leun, K. (2021, January). *Setting a binding target for 11% renewable gas*. Retrieved March 2, 2022, from https://gasforclimate2050.eu/wp-content/uploads/2021/01/Gas-for-Climate-Setting-a-binding-target-for-11-renewable-gas.pdf

Schwartz, S. H., & Bilsky, W. (1987). Toward a universal psychological structure of human values [Place: US Publisher: American Psychological Association]. *Journal of Personality and Social Psychology*, *53*, 550–562. https://doi.org/10.1037/0022-3514.53.3.550

Schwartz, S. H., & Bilsky, W. (1990). Toward a theory of the universal content and structure of values: Extensions and cross-cultural replications [Place: US Publisher: American Psychological Association]. *Journal of Personality and Social Psychology*, *58*, 878–891. https://doi.org/10.1037/0022-3514.58.5.878

Searle, J. R. (1995). *The construction of social reality* [Google-Books-ID: zrLQwJCcoOsC]. Simon; Schuster.

Tarrow, S. (1996). Social movements in contentious politics: A review article (C. F. Andrain, D. E. Apter, J. C. Jenkins, B. Klandermans, H. Kriesi, R. Koopmans, J. W. Duyvendak, M. G. Giugni, E. J. Perry, C. Tilly, & J. W. White, Eds.) [Publisher: [American Political Science Association, Cambridge University Press]]. *The American Political Science Review*, *90*(4), 874–883. https://doi.org/10.2307/2945851

Tilly, C. (1977, March). *From mobilization to revolution*. Center for Research on Social Organization University of Michigan.

Vaessen, J., Gaastra, S., van Hest, F., Peters, R., van Oord, P., & Kockelkoren, T. (2018). *Re-use and decommissioning report*. Retrieved March 1, 2022, from https://www.nexstep.nl/wp-content/uploads/2018/07/Re-use-decommissioning-report-2018-English-Version.pdf

van de Grift, E., Cuppen, E., & Spruit, S. (2020). Co-creation, control or compliance? how dutch community engagement professionals view their work. *Energy Research & Social Science*, *60*, 101323. https://doi.org/10.1016/j.erss.2019.101323

Verhoeven, I., & Bröer, C. (Eds.). (2015, September 11). Contentious governance: Local governmental players as social movement actors. In *Breaking down the state* (pp. 95–110). Amsterdam University Press B.V. https://doi.org/10.5117/9789089647597

Wang, A., Jens, J., Mavins, D., Moultak, M., Schimmel, M., van der Leun, K., Peters, D., & Buseman, M. (2021, June). *Analysing future demand, supply, and transport of hydrogen*. Retrieved March 2, 2022, from https://gasforclimate2050.eu/wp-content/uploads/2021/06/EHB_Analysing-the-future-demand-supply-and-transport-of-hydrogen_June-2021_v3.pdf

Weeda, M., & Segers, R. (2020, June 24). *The dutch hydrogen balance, and the current and future representation of hydrogen in the energy statistics* [Centraal bureau voor de statistiek] [Last Modified: 2020-07-03T11:30:00+00:00]. Retrieved March 2, 2022, from https://www.cbs.nl/nl-nl/achtergrond/2020/27/waterstofbalans-in-relatie-tot-energiestatistiek

A

Theory

A.1. Topic model

Correlation explanation (CorEx) is a topic model used in this research that uses n-grams x part of a set of n-grams X fitting them to a number of topics Y. An n-gram is a word or group of words and can be seen as how a machine perceives a word. In this section of the appendix "word" is used to indicate n-gram, because this makes it somewhat easier to read and understand. Documents are indicated by the symbol l and the presence of a word i in document l is indicated by x_i^l. In the model, x and y are instances of discrete random variables indicated by their capital letter. The algorithm optimises the fit of a group of words X_{G_j} to a topic Y_j where G indicates a set of words. The topic model's output is a set of topics indicated by $j = 1, ..., m$ where j, each topic Y_j contains a set of keywords, X_{G_j}. Both the topic and its keywords are important for understanding a value. The name of a topic, Y_j, corresponds to a certain value. Keywords X_{G_j} give context to the topic and this is quantified by three variables. Namely the degree of informativeness, the direction of informativeness and the that the learned associated by the topic model between word and topic.

A.1.1. Assumptions and specifications of topic model

Various variants of CorEx exist and in this research, the variant of (Gallagher et al., 2017) is adopted. In general, a topic model is an optimisation problem for fitting a set of words X to a set of topics Y by maximising the total correlation of X_G. Various methods exist for solving this problem. This variant of CorEx solves this problem iteratively using a point-wise approach, where each document l represents a point. This gives rise to three important characteristics which are explained using the equations (1) and (2) and (3).

$$(1) a_{i,j}^t = \text{exp}(\lambda^t(I(X_i : Y_j) - \max_j I(X_i : Y_{\bar{j}}))$$

$$(2)\ \text{log}\, p_{t+1}(y_j|x^l) = \text{log}\, p_t(y_j) \sum_{i=1}^{n} a_{i,j}^t \frac{\text{log}\, p_t(x_i^l|y_j)}{p(x_i^l)} - \text{log}\, Z_j(x^l)$$

$$(3)\ \text{log}\, \frac{p_t(x_i^l|y_j)}{p(x_i^l)} = \text{log}\, \frac{p_t(X_i = 0|y_j)}{p(X_i = 0)} + x_i^l * \text{log}\, \frac{p_t(x_i^l = 1|y_j)p(X_i = 0)}{p_t(X_i = 0|y_j)p(x_i^l = 1)}$$

In the calculation of total correlation an indicator variable $a_{i,j}^t$ is used to represent the groups of words G. The model is parameterised in such a way that the model's discriminatory ability, λ seen in the first equation, is 0 in the first iteration and increases slowly over time. This means that all topics learn the same words and then compete for words among each other. The second term indicates that this competition only occurs between the other topic $Y_{\bar{j}}$ sharing the highest mutual information with word i. Competition, normally corresponding to "learning", is represented through the parameter $a_{i,j} \in [0, 1]$ and convergence is reached when this competition stops.

Equation 2 shows the probabilistic labels and is continuously updated using equation three. Total correlation is calculated per point and represented by $logZ_j(x^l)$. This factor balances the equation and ensures that the marginal probability, $p_{t+1}(y_j|x^l)$ is normalised (takes on values of one or zero). This process continues until it converges (no longer changes). This process depends on equation one (discrimination speed) and three (labelling of the documents).

In equation three the variables x_i and y_j are binary resulting in four possible outputs. By default, the model assumes that a word is assumed not part of a document l. This apparently reduces the solution of equation three to $log(1/1) + 0$ when $x_i^l =)$ and to $logP(x_i^l|y_j) - logP(X_i^l = 1)$ when $x_i^l = 1$. The optimisation of this bottleneck improves the overall calculation cost of CorEx from $O(nN)$ to $O(N) + O(n) + O(\rho)$, with n being the variables, N samples and ρ the nonzero data entries. This is a huge improvement compared to latent tree models that run at $O(n^2)$ or worse.

A.1.2. Semi-supervised learning

While the model typically runs unsupervised, supervision is made possible through the anchoring of words. Parameter $b_{i,j}$, the anchoring strength of a word i to a label Y_j, does this by effectively raising $a_{i,j} \geq 1$. β modulates the first term in equation 1. This conserves the information of a word to its topic and reduces the information of that word to other topics. Anchoring changes in competition between topics can give the model a new convergence equilibrium.

This method is semi-supervised machine learning since only a tiny fraction of all the words (labels) will be anchored to topics (outputs). A pitfall in semi-supervised machine learning is assigning too many labels. Too many labels make it difficult to judge the label quality and how it changes the model equilibrium. The problem with that is that the labels can propagate the researcher's bias or that the new equilibrium obfuscates some desired outputs.

A.1.3. Topic model strategy

It is in the process of word anchoring important to list prior biases to increase the chances of obtaining desired outputs. In value exploration, the desired outputs are unknown yet and these outputs are therefore in direct conflict with bias. For each added label it is unknown if it is independent of other relevant topics (values). This can only be done through qualitative analysis of the results.

CorEx adds documents to a single topic based on the highest mutual information with that topic. This process depends on all the words that are part of the CorEx lexicon, X. Some words are highly dependent, meaning that they are used in various relevant values, for example, equality, sustainability and justice. If, for example, sustainability is anchored to environment, then documents talking about economic sustainability, sustainable communities, production chains, etc. will more likely be classified as environmental. These words are relatively noisy and should not be anchored and might require filtering.

Filtering is the second tool for value exploration and has a higher priority than anchoring. Documents, the raw data, are stripped from most irrelevant words such as articles, prepositions and nouns to obtain X. This is however not enough filtering as an unsupervised topic model will classify all words. Most words are semantically unrelated (to the desired results), such as methodologies, publisher names or arbitrary words such as further, criteria and process. Such words have no interesting semantics in this context, but occur frequently, which justifies filtering. Therefore x_i should not be part of any Y.

A.2. Contentious governance and social movements

It is not uncommon in science that multiple viewpoints exist on the definitions or operation of a mechanism embedded in a system. When these factors cause a fundamental change in the operation of a ”controversy” or ”dichotomy” arises. Multiple valid, but incommensurable viewpoints exist in this situation which is an inevitable stage in science. The clue of the story is that the dichotomy is false and both viewpoints illuminate what cannot be understood yet. The goal is to understand this although the ”state of the art” is seemingly a stalemate of opposing viewpoints. This is also the case in contentious governance literature.

A.2.1. Dichotomy in contentious governance

A description is given of the dominant and an alternative viewpoint. The dominant viewpoint is ”the political process” inspired by Tilly who coined the term contentious governance. An alternative viewpoint, inspired by the critique of Jasper and is called ”the strategic perspective”. These names are inspired by (Tarrow, 1996) and (Verhoeven & Bröer, 2015) respectively and illustrated in the figure below.

The political process is the first perspective considered in contentious governance. It contains authorities and social movements. Tilly uses a more constrained definition of authorities, the state. State implies a non-level playing field as a state has unrestricted power to repress social movements

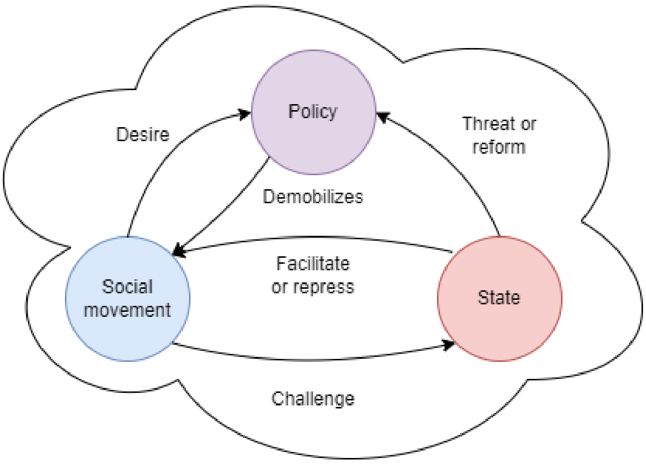


Figure A.1: Arrows indicate the relations between entities in political process theory. Contents in the cloud represent what is understood by ”opportunity”.

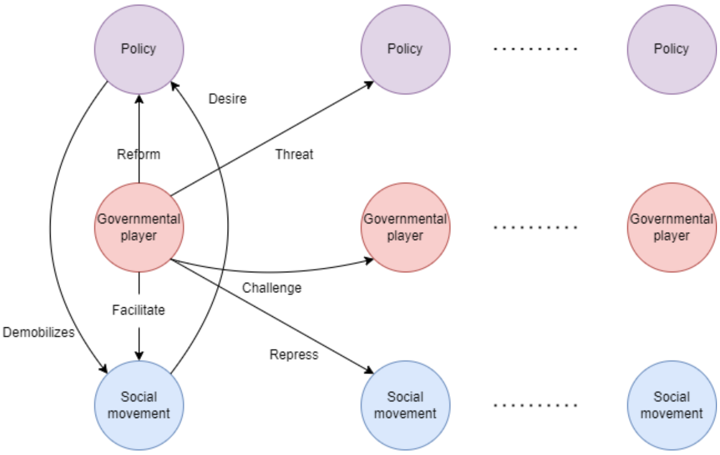


Figure A.2: Contentious governance illustrating a multitude of policy objectives, movements and governmental players.

(Tilly, 1977). Social movements can be defined as ”a sustained challenge to powerholders in the name of a disadvantaged population living under the jurisdiction or influence of those powerholders” (Tarrow, 1996). Authorities and social movements are separate (Verhoeven & Bröer, 2015). Authorities can provide opportunities or constrain social movements by reform (pass favourable policy), threat (pass opposite policy), facilitate (provide resources) or repress (increase costs of social action) (Koopmans, 1999). The study of these opportunities is considered to be important in determining the outcome of a social movement, although this is the most scrutinised term in political process theory (Koopmans, 1999).

The alternative viewpoint frame authorities as a multitude of actors. It is argued that increased autonomy and the dispersion of governmental networks saw governmental players influencing and engaging with social movements (Verhoeven & Bröer, 2015). Governmental players initiating claim making are civil servants seeking to expand their power and advance their ideas of public interest which attract bystanders. Their engagement with social movements gives these players support for their policy and is pro-active (Verhoeven & Bröer, 2015) (Pettinicchio, 2012). Competition between governmental players by means of claim making can be seen as a strategy for gaining legitimacy and breaking political stalemates (Verhoeven & Bröer, 2015).

B

Ensemble run

B.1. Ensemble run

Reliably filtering documents from the corpus requires an ensemble run. An ensemble run is a set of topic model runs that combines all the outputs together. Using many runs increases the reliability of the results since results are subjected to randomness. Furthermore, a single run assigns a probability to a documents belonging to each topic, $p(y|x)$. This probability is volatile which can be reduced by combining the results of many runs.

A proposed measure for merging the results of multiple runs (ensemble run) is suggested. For an ensemble with n runs take the chance that document x_l belongs to topic y_k, $p_{ensemble}(y|x)$ or simply $p_{ensemble}$. Representing the results of the ensemble run can be done in various ways. First is the chance the document belongs to that topic every run. Second the chance the document belongs to that topic next run. Third is the chance the document belongs to that topic every run with some degree of certainty, c. A fourth way could be introduced if one wants to reduce dependency on the number of ensemble runs. The fourth is the chance a document belongs to that topic every run relative to the chance a document is assigned to a topic in general. This would add a new discriminant, based on relative chance and this is not further considered. Relying on relative chance obscures the actual chance a topic appears, which is undesired as it can result in topics with $p \leq 50$% appearing or $p \geq 99.9$% disappearing.

$$p_{ensemble} = \prod_{n=0}^{n} p(y|x) \tag{B.1}$$

$$p_{ensemble} = \sum_{n=0}^{n} p(y|x)/n \tag{B.2}$$

$$p_{ensemble} = \begin{cases} 1 & \prod_{n=0}^{n} p(y|x) \geq c \\ 0 & \text{otherwise} \end{cases} \tag{B.3}$$

The third equation, equation B.3, is suggested to be used and its desired parameter settings are discussed. In this equation, the ensemble reflects the chance a document belongs to a topic in all runs given some certainty criteria c. Typical values for c are 95, 99 and 99.9% known from statistical testing. c reflects all runs in the ensemble and is related to an individual run through n, the number of runs in the ensemble. Of interest is the minimum required chance a document belongs to a topic in an individual run, p_x, which is dependent on n and c in the equation ${p_x}^n = c$ which can be expressed in terms p_x, see equation B.4.

$$p_x = 10^{log(c)/n} \tag{B.4}$$

Before going into the choice of number of runs n the number of sampled documents has to be considered. For a lower number of topics, the results appear to be fairly constant, although increasing the number of topics will start to show fluctuations in topics every run. An explanation for this phenomenon

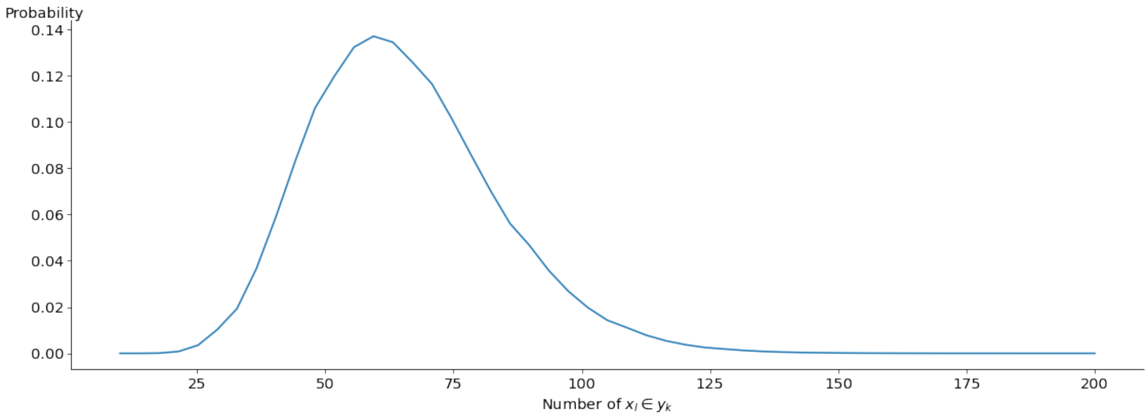


Figure B.1: Chance to draw 10 documents of a topic. Ten thousand samples were drawn from the corpus where the y-axis shows the chance that at least one document is drawn that belongs to a specific topic with the x-axis specifying the amount of documents in that topic.

c (%)	n	p_x
95	20	99.7%
99	50	99.97%
99.9	100	99.999%

Table B.1: Table showing the minimum chance a single document belongs to a certain topic for some given c, certainty and n, number of ensemble runs. While the bump up in p_x seems large, it reduces the number of documents passing condition c by only 20%

is that the sample of documents contains too less words or documents. This can be solved by increasing n_docs or decreasing min_df (in the tf-idf vectorizer). Changing the latter is not recommended though as it exponentially increases the number of words and does not solve the issue of ”low content variety” which an increase in n_docs would solve. Low numbers of documents furthermore decrease the size of documents part of any topic. The chance of successfully drawing documents in a corpus follows a hypergeometric distribution. Drawing k out of K documents when drawing n documents from a corpus of size N is described by equation B.5 where brackets signify the binomial coefficient. An illustration, B.1 is made for drawing ten thousand documents out of a corpus of sixty thousand documents where the figure illustrates the chance for exactly drawing 10 documents part of a random topic that has a total document membership as indicated by the x-axis. Drawing ten thousand documents is large meaning that the number of documents part of a topic consistently exceeds a hundred documents. This figure indicates that poor choice of n_docs will consistently yield low draws that prevent a topic from emerging. It also shows that n_docs doesn't play a role when a low number of topics are chosen. n_{docs} could play an important role when the number of topics is increased. Firstly, this is where n_docs starts to affect run time significantly. How it affects topic formation is unknown. What can be said is that a too low value definitely affects the results completely changing most of the topics. For reasonable values of n_docs it can have some effect on individual runs, the most it can do is unexpectedly change topics and severely strengthen or weaken other topics, although that should not be a problem. Quantifying how n_docs influences the results is a future research option.

$$p_k = \frac{\binom{N}{k} * \binom{K-k}{n-k}}{\binom{N}{n}} \tag{B.5}$$

The example in figure B.1 is generous since a relatively high number of documents were chosen. Even when using several hundreds of topics the effect of drawing documents on the results is negligible. Using lower amounts of documents will have a much more profound effect because the graph effectively moves to the left. Knowing what number of documents to choose to get an acceptable document draw is advantageous to increase both performance and quality.

Another issue with the results of individual topic model runs is their apparent probabilities. Since results vary per run it is not possible to choose one representative run. Such a run would be a combination of existing runs. Using the result criteria from statistics. Given certainty criteria c and n, the number of runs in the ensemble, one can calculate the average chance (a document belongs to a topic in an individual run) that is minimum required to satisfy condition c. For now, this value is called p_x and is expressed by the equation ${p_x}^n = c$. p_x is found by $p_x = 10^{log(c)/n}$. Table B.1 shows that the average certainty per run required to satisfy c is significantly higher. By intuitive judgement, a reasonable c would be 95% and a reasonable n around 20. This means that p_x equals 99.7%. This should not be a problem since in an individual run p_x tends to be close to 0 or 1 (higher than 99.9%).

Ensemble runs would work if the topics are supposed to represent the same thing. The problem is that they don't. Topics between runs are always fundamentally different things. Let's describe several scenarios of what can happen to the topic. Three things can vary, the member documents, n-grams and if the topic appears or not. Topics don't appear consistently at the same place, but can still be identified based on prior n-gram contents. As long as the same number of topics are employed imagining a concept of a representative run is legitimate. Things get more complicated if the number of latent topics is varied as well. In this case, the topic is no longer the same and represents something different making an ensemble run unsuitable.

C

Topic modelling framework: A guide for finding values

The topic modelling framework is illustrated in figure 5.1 and each step is explained in this appendix. This framework structures the procedure in distinctive steps and should help people who want to use or improve topic models. Each step in the framework is affecting the results. Not only is it transparent to communicate how results are obtained, it effectively forms the approach's limitations as it tells what the model can and cannot say. It thus helps with answering to whom the results are relevant. Using the framework should speed up topic modelling exercises as knowledge of these limitations allows efficient problem identification in the overall procedure. This allows operators to improve the quality of their results by working on the proper steps and preventing them from spending time on steps that would yield marginal improvement.

C.1. Introduction to topic modelling

Topic modelling is a technique that classifies groups of text data. These groups can be anything, for example, tweets, scientific literature, books and news articles. The advantage of topic modelling is that it is quick compared to non-computational methods. It was therefore decided to look at the possibility of finding values using topic modelling. Alternatives exist for this purpose. Think of search engines or other quantitative methods. Topic modelling is doing it in a different way seen in both the method and in the results.

General algorithms for topic modelling use the number of topics as the only available dependent variable. This variable depends on the form of the topic, also called a latent topic. It is latent in terms of being derived (mathematically) from the documents. How a topic takes shape is in relation to the number of topics and the entire group of documents. More topics mean more detail and vice versa. The set of all topics is some representation of the entire group of documents (corpus). A corpus relating to a specific subject will yield different results from a generic corpus.

It is possible to steer topic models with semi-supervised models giving specific ”learning” components a heavier weight than others. This influences how topics are shaped. An idea is to use topic models to find values which mean that these values require sufficient presence in the corpus to be detectable in the first place. The question now is what the limitations are to this challenge.

C.2. Which values are we looking for?

The first challenge is formulating the values we are looking for. First is the consideration that not all are represented. Values (see the introduction) relate to an individual or a group of individuals. The results of the value finding exercise only relate to this group. It does not relate to individuals with a different set of values. If you want the results to relate to another group, then a new exercise has to be conducted for them. In this research, this is seen in the first section of the results reflecting corporate values and the last section of the results reflecting general public values.

The formulation of values happens in the form of n-grams. N-grams are the equivalent of a word or word-group for a computer and are used as input for the model. Formulating values comprises listing the relevant values. A list of n-grams has to be assigned to each value. A requirement is that this n-gram has to occur in the corpus in a minimum of min_df documents. This means that value selection is closely related to corpus selection.

C.3. Corpus selection

The choice of included documents has a big effect on the topics that appear in the results. The output topics in an unsupervised model are a direct reflection of this corpus. Irrelevant documents should be filtered because it will cause insensible topics to be generated. Filtering words makes no sense because the word vectorisation scheme (tf-idf) primarily depends on the document and not the words. Word filtering is comparable to whac-a-mole as filtering one irrelevant word will cause another irrelevant word to appear without solving the problem of having a poor corpus.

What are irrelevant documents? Is an important and difficult question to which several answers are given. First, it could be that a document is unrelated to a theme. We are, for example, not interested in nuclear reactors when we are looking for values in green hydrogen. Second, is considering a document irrelevant if it does not contain at least one value as defined by the n-grams in the previous step? If no value is present, then the document could not possibly generate a relevant topic.

The most important reason for filtering documents is that their presence hinders the identification of interesting values. It has to be noted that finding relevant values is challenging because the abundance of irrelevant information dominates the process of anchoring in the results. Note that anchoring is the process of giving an n-gram related to a value a higher learning weight in the hope that it forms a topic.

C.3.1. Corpus selection and latent topics

Corpus selection affects the ability of the model to identify latent topics. Latent topics are latent variables which are mathematically inferred variables. Latent topics can also be ”latent” in the sense that they convey hidden information that was not known during the value definition phase. This can mean that there is new information regarding a value as defined before or that a new topic appears which is a relevant value to the stakeholder, but which was not defined before. Anecdotal advice from this research prescribes that a latent topic usually conveys information that was not known before. The removal of documents conveying no values could therefore constrain this. This furthermore reduces the topic modelling to the confirmation of the presence of prior known values.

Now consider the latent topics in the results that do not correspond to the values as defined in the first step but are relevant to the stakeholder in hindsight. These ”new” latent topics convey a new idea, value or concept. These topics are also latent because they are not only latent variables, but they also convey latent, hidden, information. These topics are most useful when the number of topics is kept low to keep the difficulty and time costs of result interpretation low. A consequence of this is that ”new latent topics”, are becoming less useful when more values are sought as this raises the number of topics. In the context of finding values, such latent topics are irrelevant because these will not appear. Specifically irrelevant to the values as defined in the first step. This has most likely to do with corpus selection and language processing.

C.4. Natural language processing

Natural language processing relates to the activity of translating words in a document into numbers. The aforementioned word filtering and word vectorization scheme are part of this. The goal of this step is to give the machine the ability to detect the presence of a value. It, therefore, includes alternatives to the word vectorization. The word vectorization scheme used in this research, the tf-idf, takes a contiguous word or groups of words. The size of this group can be arbitrary. In this research a size of up to three was taken. The problem with this scheme is that the machine is not able to detect in which context a word is used. This causes documents to be incorrectly attributed to a topic. What solves this problem is using different (and more advanced) ways of processing text data. Language processing ideally happens in such a way that the information relating to the context in which a word was used is retained. Such a possible solution relates to deep linguistic processing models.

Because of this reason, several values are unable to be properly found and other values will incorrectly appear (false positives and false negatives). Many examples of these false positives are found in the results of core values. Examples of values are ”best”, ”control”, ”exploration”, ”flexibility”, ”impact”,

”potential” and ”recognition”. The description of these values emphasises that documents are out of context and do not relate to these values, but to something else instead.

Another question is what the presence of a value truly means. While a value relates to a desirable state or behaviour, it seems that values are used to fulfil a need, scarcity or demand. Take the topics ”best” and ”effectiveness” as an example. These are often used in the context of chemical processes. It seems that these values are communicated because an increase in efficiency is desirable in that field (a norm). The presence of values in scientific literature could therefore indicate what is normatively desirable to talk about. It is a description of the culture in which the scientific discourse takes place. What the presence of a value further means is guesswork and reflects the imagination of the individual who does this interpretation.

C.5. Topic modelling

This step relates to the choice and execution of the topic model. Various topic modelling techniques exist. Each has its advantages and disadvantages. In the context of the goal of finding values, a semi-supervised model was chosen. There are not a lot of variables in the model execution phase. The first variable is feeding the model with the processed corpus in the previous step. The second variable is using the list of values in the first step. Third and last are the number of topics. Depending on the method there may be more variables, but discussing them in detail is irrelevant here. The last variable, the number of topics, determines the level of detail in the results. The values determine what we look for (in the results) and the corpus forms the basis of the results and determines where the results are sought.

C.6. Interpretation of results

The results of the topic model require translation from numbers to words. Interpretation is the process that translates these results from numbers to words. It is a formal process and there is ideally no interpretation of the researcher taking place. The results of the topic model can be condensed into two matrices. These are the word-topic (α) and document-topic (θ) distributions. These respectively tell for each word and document to which topic they belong.

How these distributions look differs per model. The word-topic distribution consists of two distinct components in CorEx, the topic model used in this research. This is the weight for every n-gram towards each topic. The other component is the quantity of information that a word communicates to a topic. There are also two components for topic-document distribution. The first one contains the chance a document belongs to a topic and the second one is the total information contained in a topic.

There is an important relationship between these four components. These are the total information in a topic and the information that a word communicates to a topic. The sum of all information contained by the n-gram to a topic equals the total information in a topic. Through this equality, it is possible to figure out how much information a specific group of n-grams delivers to a topic. Let this group of n-grams be the values defined in the first step, if their contribution is higher than a predefined criterion it is possible to objectively measure if this group of n-grams is present in a topic. This way of result interpretation is called the automatic method. The documents of the chosen topics are required to be manually read, however.

This is the counterpart of the ”manual method”, which means that latent topics are manually interpreted to determine if a topic is interesting or corresponds to one of the values defined in the first step. The manual method requires specification of formal criteria and this is more cumbersome and time-consuming than the automatic method. The advantage of the manual method is that it is possible to discover latent topics that convey some ”hidden” information mentioned earlier. The disadvantage of this method compared to the automatic method is that it is less consistent, more prone to error and that it takes much more time especially when the number of topics is increased.

D
Green hydrogen

By 2050 the European Union aims to have net zero emissions. The majority of emissions originate from industry, transportation and power generation. These sectors are primarily fuelled by fossil sources. Drivers of this changing energy policy are climate change, but also the depletion of native petroleum sources. Such an ambitious policy will lead to rapid adjustments in society. This is a multifaceted problem addressing among others domestic support, technical capabilities and international relations. Hydrogen, the molecule, plays an important role in this transition and will be the focus of this research

D.1. Hydrogen and zero emissions

Hydrogen is already playing an important role, according to CBS hydrogen is primarily produced using fossil fuels (99%) contributing to 180 out of the 3000 PJ yearly energy consumption (Weeda & Segers, 2020). A colour system exists to distinguish the various types of hydrogen. The Hydrogen used today is called grey hydrogen because it is produced by fossil fuels. It is projected to be replaced by blue hydrogen and green hydrogen. Green hydrogen (GH) or renewable hydrogen is hydrogen produced from electrolysis using renewable energy sources only. Blue hydrogen or low-carbon hydrogen is produced from Methane (natural gas) using steam reforming and opens up the opportunity for carbon capture storage (CCS) to further decrease the carbon footprint. Biogas or biomethane is methane produced from non-fossil solids such as waste and biomass. It is despite its high costs a proposed bridge from conventional natural gas towards blue hydrogen (Schimmel et al., 2021). Steam reforming from biogas would net negative emissions when combined with CCS, however, this is not only expensive, CCS is only feasible on a large (centralised) scale and biogas is produced on a decentralised scale (Wang et al., 2021) (Weeda & Segers, 2020).

D.2. Hydrogen as energy carrier

An energy carrier can be defined as a medium in which energy is stored, for example, oil, electricity, coal, natural gas and ammonia. Notable properties are costs of storage, speed of transportation, energy loss during transportation and potential greenhouse gas emissions. Hydrogen can be used as an energy carrier allowing transportation (through pipes) and storage (in aquifers and depleted reservoirs) of electricity. Because of this, ample use of GH becomes advantageous as a society relies more on renewable energies. The development of low-carbon blue hydrogen is a bridge towards a carbon-free energy system. The use of hydrogen can be expanded in industry, transport and energy sectors as well as regular buildings. To illustrate, many existing industrial processes rely on grey hydrogen, this could be replaced by green or blue hydrogen. Blast furnaces using cokes and fossil fuels as oxidisers could replace carbon with hydrogen as oxidiser. In the transport sector vehicles could increasingly use hydrogen fuel cells. The advantage of a hydrogen energy carrier is its wide set of applications and absence of greenhouse gas emissions. Ammonia, with similar properties, but advantages in storage and transportation is sometimes presented as an alternative to hydrogen. Natural gas will be phased out due to its greenhouse gas emissions. Decommissioning of natural gas production infrastructure is an opportunity for these alternatives. This transition requires the adjustment of existing natural gas infrastructure.

D.3. Hydrogen backbone

The leading gas transport companies in Europe founded the European hydrogen backbone in 2017 directing the future gas market in Europe. Ambitious plans are made in favour of hydrogen and liquefied natural gas (LNG) is there to serve a transitional role in the decarbonisation (Wang et al., 2021). This initiative has the goal of rapidly expanding the hydrogen infrastructure and replacing natural gas with hydrogen by 2050. Infrastructure expansions would amount to 11600 km in 2030 (specifically Lower Saxony and the low countries) expanding to 39700km in 2040 connecting most European countries to the hydrogen network. Most of this, 69%, is from recommissioning natural gas pipelines to hydrogen pipelines (Jens et al., 2021). The goal of the hydrogen backbone is to supply 11% of all gas with renewable gas (3% renewable hydrogen and 8% biomethane) by 2030 while the remainder is fossil and low-carbon gas (Schimmel et al., 2021). Fossil hydrogen has to be replaced by low-carbon and renewable gas to reach the net-zero emission goal by 2050 (Wang et al., 2021).

The future of hydrogen is a collective initiative of European gas transport companies conforming to the EU goals and the Paris climate agreements. Accelerating the transition of natural gas infrastructure to GH amplifies its vulnerabilities. Minor concerns are economic means, environmental impact and technological uncertainties. Hydrogen technology is instead an opportunity for the (petroleum) industry to meet (European) climate goals. It should be noted that gas transport companies are typically owned by petrol and state players. Hydrogen is an economic opportunity in the context of gas infrastructure decommissioning (Vaessen et al., 2018). Meeting climate goals is a necessity from the perspective of business viability. The policy pathway requiring action on the national and EU levels for the realisation of the green hydrogen ambition should be of major concern.

D.4. Green hydrogen ambition

Policy requirements for accomplishing the green hydrogen goals are ambitious, to say the least. These goals affect many stakeholders in society. Several examples are given in this section. These goals dictate requirements on the incorporation of the carbon market regarding carbon pricing, the European emission trading system (ETS) and including aviation and shipping into ETS. Not only the carbon market but also the hydrogen market and heat pump market require suitable conditions and regulations on a European level. Lastly, collective national action is required regarding national CCS funds, national green hydrogen funds with mandatory hydrogen goals, phasing out of coal-based power and the refurbishment of gas power plants to hydrogen power (Peters et al., 2020). Failing to implement these policies increases the risk of failing the goal of net-zero emissions by 2050. Numerous pathways exist for blocking or delaying any of these objectives and can occur on European, national or domestic levels.