Textual Contexts for "Democracy": Using Topic- and Word-Models for Exploring Swedish Government Official Reports

Magnus Ahltorp[1], **Luise Dürlich**[2], **Maria Skeppstedt**[1]
[1]The Institute for Language and Folklore, Sweden
{magnus.ahltorp,maria.skeppstedt}@isof.se
[2]Department of Linguistics and Philology, Uppsala University, Sweden
luise.durlich@lingfil.uu.se

Abstract

We here demonstrate how two types of NLP models – a topic model and a word2vec model – can be combined for exploring the content of a collection of Swedish Government Reports. We investigate if there are topics that frequently occur in paragraphs mentioning the word "democracy". Using the word2vec model, 530 clusters of semantically similar words were created, which were then applied in the pre-processing step when creating a topic model. This model detected 15 reoccurring topics among the paragraphs containing "democracy". Among these topics, 13 had closely associated paragraphs with a coherent content relating to some aspect of democracy.

1 Introduction and background

Methods developed within NLP have been useful additions to the computational social science and humanities toolbox. The classic NLP method of topic modelling is, for instance, widely used (Boyd-Graber et al., 2017). Examples of text genres analysed with topic modelling include news paper text (Blei, 2012), folk legends (Karsdorp and den Bosch, 2013), micro blogs (Surian et al., 2016), student essays (Ferrara et al., 2017) and open-ended survey questions (Baumer et al., 2017).

Topic models are used for discovering reoccurring topics in a collection of documents. The models are based on the co-occurrence of words. That is, words that frequently occur together indicate a recurring topic. Each topic detected is typically represented by (i) a ranked list of the words that have created the topic by frequently co-occurring, and (ii) a ranked list of the documents that are most typical for the topic, i.e., the documents in which the words frequently co-occur.

Since topic models are built on modelling the co-occurrence of words in the same texts, they model the syntagmatic relations between words. There are also NLP methods for building models based on paradigmatic relations of words. That is, models that can detect to what extent words typically occur in similar contexts, e.g., to what extent they are synonyms/near synonyms (Sahlgren, 2006). Although these models have existed for quite some time, interest has exploded in recent years with the re-emergence of neural networks as a popular method for machine learning (Vasilev et al., 2019). Also for these methods, there are many different types of use cases within social science and the humanities (Dahlberg et al., 2017; Loon et al., 2020).

We will here demonstrate how these two types of models can be applied to a text collection consisting of Swedish Government Official Reports, and how the output of the models can be combined for finding reoccurring content in the text collection. The aim for the demonstration task will be to investigate if there are topics which frequently occur in texts that mention the word "democracy".

2 The Topics2Themes tool

Despite the shown usefulness of NLP models, previous research has also demonstrated the importance of performing a manual analysis of their output (Grimmer and Stewart, 2013; Baumer et al., 2017). For instance, to read topic-typical documents extracted by a topic modelling tool, in order to avoid misinterpreting the words representing the topics (Baumer et al., 2017; Lee et al., 2017). We, therefore, here use a tool for topic modelling, Topics2Themes (Skeppstedt et al., 2018), which has a graphical user interface meant to encourage the user to read and further analyse the documents extracted by the topic modelling algorithm.

This tool has previously been applied to other types of text collections (Skeppstedt et al., 2020a,b, 2021). Information from paradigmatic models has also been incorporated previously, in the form of

word2vec models pre-trained on large corpora other than the text collection analysed. We here (i) apply the tool to the text genre of political texts, and (ii) use a word2vec model that has been trained on – and thereby is more specific to – the text type that is to be explored with topic modelling.

3 Swedish Government Official Reports

Committees or special investigators are often appointed by the Swedish Government to investigate a particular issue before a legislative proposal is presented. The results are compiled in reports, which are published in the official report series "the Swedish Government Official Reports" (or "Statens offentliga utredningar", SOU, in Swedish).[1]

The report series is made available as PDF documents and automatically extracted HTML pages at the open data site of the Swedish Parliament.[2] This HTML extraction has not preserved the logical structure of the PDF, e.g. headings and regular text are assigned the same HTML tags, and any distinction by font or type-face is encoded explicitly in style attributes which vary across different reports. Reports between the years 1994 – 2020 (3,558 reports) have, however, also recently been made available in a further processed version.[3] This version includes (i) a separation of summaries from the full texts of the reports, (ii) HTML markup that indicates titles, section headings and paragraphs in the body text, and (iii) removal of tables, lists, diagrams and non-Swedish texts. We here used the full texts of the reports from this further processed version, as well as title and heading markup.

4 Extracting "democracy" documents

Given the 2021 celebration of 100 years since the first general elections in Sweden, we decided to focus on the word "democracy", and the contexts in which it appears.

Another decision to make when constructing a topic model is how to define a document. When the collection, e.g., is made up of a compilation of short texts, the decision is easy. In this case, we instead have a collection of very long documents, which need to be split up to make them manageable (for a human as well as for the machine). We therefore decided to define a document as a paragraph.

[1] https://www.riksdagen.se/sv/Dokument-Lagar/

[2] https://data.riksdagen.se

[3] At: github.com/UppsalaNLP/SOU-corpus. In the project: datalabb.esv.se/esv-datalabb.html

Word	Occ.
demokratiska ("democratic", plur. & det.)	7,601
demokrati ("democracy")	6,965
demokratin ("the democracy")	6,222
demokratiskt ("democratic")	3,970
demokratisk ("democratic", common gender)	3,033
demokratins ("the democracy's")	2,613
demokratiutredningen ("the democracy-inquiry")	511
demokrativillkor ("democracy-conditioned subsidies")	449
demokratiutredningens ("the democracy-inquiry's")	396
demokrativillkoret ("the democracy-conditioned subsidies")	318

Table 1: Number of occurrences for the most common types containing the string *demokrati* ("democracy")

Following these two decisions, the collection to analyse with topic modelling was constructed as follows: We extracted all paragraphs containing the string *demokrati* ("democracy"). The string was allowed to occur as a sub-string of a word, which led to morphological derivations (e.g., democratic), as well as compound words (e.g., the democracy-inquiry) being captured. A total of 1,174 types were detected (top 10 are shown in Table 1). A manual inspection of the types showed occurrences of (different forms) of five political party names among the types detected, e.g., Social Democratic. Paragraphs containing the string *demokrati*, solely as a part of a party name were therefore excluded from the documents extracted. This resulted in a collection containing 25,988 documents (after making sure there were no exact duplicates), extracted from a total number of 2,965,751 paragraphs.

5 The word2vec model

Standard topic modelling does not take the meaning of the words into account. That is, the algorithm is agnostic to the semantic similarity of word-pairs such as "organisation"/ "organisations", and "states"/"countries". The semantic similarity of the first word-pair can be detected by, e.g., stemming, but for the second pair, there are no morphology-based solutions. Topics2Themes therefore provides a functionality for clustering the words occurring in the texts into groups of semantically close words. These words are then treated as a single

concept by the tool, e.g., the combined concept "states/countries" is created. The clustering algorithm used is called DBSCAN (Ester et al., 1996). As input to the clustering algorithm, the tool needs to be provided with a suitable word2vec (Mikolov et al., 2013) model. That is, the clustering uses the semantic vector that represents each word included in the model. Since, e.g., "states" and "countries" are semantically close, they are likely to also have word2vec-vectors that are close to each other in vector space, and thereby be clustered together in the same concept cluster.

We have previously used pre-trained word2vec models. Here, we instead created a model specific to our collection, by training a word2vec model on the text type that is to be explored with topic modelling. We used the gensim library (Řehůřek and Sojka, 2010), and trained the model on 30% (48,313,487 tokens) of the report collection. We included tokens occurring at least 20 times, and used CBOW with a window size of three.

6 Applying the topic modelling tool and improving its configuration

The Topics2Themes tool was thereafter applied to the "democracy" paragraphs, using the newly created word2vec model for concept clustering.

Despite already having removed exact duplicates, the automatic duplicate detection of Topics2Themes detected 4,841 paragraphs with at least a 15-token overlap with another paragraph. These duplicates were removed, as duplicate content otherwise is interpreted as reoccurring topics.

We configured the Topics2Themes tool to use the topic modelling algorithm non-negative matrix factorization (Lee and Seung, 2001), and to extract 20 topics. However, we also configured the tool to automatically re-run the algorithm 100 times and only retain stably occurring topics. (Due to the algorithm's non-determinism, slightly different results are typically obtained each time it is run.)

Topics2Themes provides functionality for allowing the user to iteratively improve its output, both for improving the core topic modelling functionality and how the word2vec model is integrated. We therefore ran the algorithm 47 times (the first 13 times with another, pre-trained word2vec model), each time adding improvements to the model.

A basic configuration parameter is the maximum euclidean distance for two word2vec vectors to be allowed to be positioned in the same cluster. With a large distance, semantically distant words will be clustered together, whereas a small distance will lead to fewer relevant clusters being created. By manually inspecting the clusters created for different distances, we settled for a distance of 0.62.

Another important configuration improvement consists of adding additional content to four different lists. These lists contain the following (i) stop words (i.e., uninteresting words that are not to be included in the content sent to the topic modelling algorithm, e.g., "therefore", "mainly"), (ii) words to exclude from the automatic clustering since they are assigned to clusters to which they do not belong, e.g., clusters of antonyms and of semantically close words that might nevertheless be relevant to separate (e.g., party, place and person names), (iii) manually constructed clusters, i.e. groups of words that should be treated as the same concept but were not captured by the clustering (e.g. "the parliament"/ "the parliament's"), and (iv) a list of multi-word expressions that should be treated as one word by the algorithm (e.g. "political party").

For stop words, we extended the Swedish stop word list provided by NLTK (Bird, 2002) and a list from a previous Topics2Themes study (Skeppstedt et al., 2020b). By withholding the stop words from the algorithm, it is possible to prevent the creation of uninteresting topics based on these words. For instance, that two documents both contain the word "mainly" is a bad indicator of these two documents discussing the same topic. The clustering facilitates the stop word list expansion. That is, since uninteresting words are often clustered together, the entire cluster can be added as stop words.

In addition to removing stop words, we also removed low-frequency words/clusters, i.e., only the 5,000 most common words/clusters were retained.

7 Final configuration and topics detected

The final configuration resulted in 15 stable topics being detected by the topic modelling algorithm. For this configuration, there were 436 words in the list of words not to cluster, 73 manually constructed clusters, 20 multi-word expressions, and we had added 892 new words to the stop word list. A total of 530 word clusters were automatically detected by the word2vec-vector clustering.

We employed the graphical user interface of Topics2Themes for exploring the output produced by the topic modelling algorithm (Figure 1). The Topics-panel (in the center) contains one element

1: Democracy in municipalities and regions, e.g. its vitality, level of autonomy and responsibilities: municipality*, region council*, regions/responsible authority/regional councils/County Administrative Board, councils, regions, assignments, cooperation, consultation, tasks, municipal law, possibility*, activities*, municipal autonomy/the municipal autonomy
2: Internal school democracy for pupils, the school's commission to teach democracy: pupil*, school*, teacher*/teachers*, common values/values, education/teaching, influence, school, commission, knowledge, schools*, grade/primary school/gymnasium, Agency for Education, children*, curriculum*
3: Democracy-conditioned subsidies for organisations (many from SOU 2019:35): organisation*, conditions*, grants*/the support, activities*, support, civil, authority*, requirements*, ideas, Agency for Youth and Civil Society, fulfils, authority, submitted
4: A non-coherent topic: projects, perspectives, education*, universities, power
5: Political parties and their relations to voters and members (Many from SOU 2016:5): party*, voters/voters, members, internal, representative, candidates, elections, party members, shows/showed, election, role, the voters'
6: EU and democracy, e.g. how EU-democracy works, and its challenges: EC/EU, European*, national, level*, membership*, the Union/the community, Sweden, Swedish, member states*, Sweden's, countries*/states*, the deficit, the cooperation, the parliaments, power
7: Challenges, opportunities and interactions of local and regional democracy: local*/regional*, level*, national*, anchoring, county, strengthen*/improve, experimentation, work, development, development, local autonomy/municipal autonomy, responsibility
8: The importance of a broad political participation: political*, politics*, participation, the system, institutions*, system, engagement, representative, elections, equality, economic, power, forms, decision-making, social
9: Young people's political and societal participation and influence: young people*, children*, influence, commitment, participation, children and young people/children and adolescents, youth councils, to influence, engage, adults, increase, knowledge
10: About basic human and democratic rights: rights, fundamental, human rights*, limitation*/restriction*, law/ordinance rules*/provisions*, the Instrument of Government, protection, universal, common values/values, the right, society*, respect, requirements*, principles*
11: Democracy in municipalities e.g. how to strengthen it: municipal*, elected*, activity*, the audit, municipal autonomy, way of functioning, commission, strengthen*/improve, cooperation
12: Gender equality: women*, men*, gender equality, gender*, violence, power, organisation, equal, female*, women's movement, gender equality policy
13: Decision making in democracies and democratic organisations: decision*, take*, council, decision, opportunity*, influence, the board's/the council's, order, requirements*, majority, responsibility, views, legitimacy/credibility, level*, consultation
14: A wide topic, with texts containing mentions of "the Governmet". Some more specific texts about the relation between the Government and public authorities: the Government*, authority*, administration, state, the administration, commission, administrative policy, public, transparency, activities*, growth, the authority's/the agency's*/ the County Administrative Board, involvement, state administration, the work
15: The state of democracy in different countries, e.g. election participation (Many from SOU 2007:84): United States/China/Japan/Norway/Poles/India/Ireland/Canada/Hungary/Belgium/Denmark/Finland/Italy/Spain/Portugal/Romania/Czech Republic/Germany/Bulgaria/France/Slovakia/Slovenia/South Africa/Austria/Australian/the Netherlands/Great Britain, countries*/states*, Sweden, Estonia/Lithuania/Latvia, Russia, European*, elections, country, election participation, Iraq/Syria/Egypt, Lebanon/Somalia/Turkey/Jordan/Indonesia, Eastern European, Switzerland

Table 2: The 15 topics detected and their most closely associated words and concept clusters (translated into English). Concept clusters are indicated by "/" separating the words in the cluster. That the cluster contains different morphological versions of a word is indicated by a "*" following the shortest version.

for each one of the topics detected. To the left, the words/concept clusters associated with the topics detected are shown. Correspondingly, to the right, the documents (i.e., paragraphs in this case) associated with the topics are shown. It is possible to re-sort the texts according to their associated texts and words. The words associated with the topics are also highlighted in the texts. To further support the reading, each paragraph has labels that show (i) the title of the report in which it appears, and (ii) the nearest heading under which it appears.

We read a few of the most closely associated paragraphs (about five) for each one of the 15 topics detected, and added a description of the topic in its text area in the Topics panel. For the paragraphs associated with Topic 4, we were not able to find any common subject. For topic 14, the main connection between its associated paragraphs was that "the Government" was mentioned. However, for the other 13 topics, the associated texts all deal with some aspect of a common topic related to democracy. The topic descriptions and their most closely associated words can be found in Table 2

A potential effect of splitting up reports into paragraphs, and treating them as independent documents, is that the topics detected might correspond to the original reports. That was the case for topics 3, 5 and 15. For these topics, the algorithm had (more or less) detected what corresponded to the content of the reports "Democracy-conditioned subsidies for civil society organisations", "Let more people shape the future!" and "The importance of a high voter turnout", respectively. To avoid this, all paragraphs from the same source report could have been concatenated into one document. However, treating the paragraphs as independent documents also has potential advantages, e.g. making it easier to detect subtopics within a report.

8 Concluding words

It would have been a time-consuming task to manually search for reoccurring topics among the 25,988 paragraphs containing the word "democracy". With the Topics2Themes tool, in contrast, it was possible to very quickly gain an overview of reoccurring content. Additional relevant reoccurring topics might be found by analysing all paragraphs, as the tool is not likely to detect everything relevant to a human. However, in the absence of unlimited resources for manual text analysis, NLP models can be the next best option. For this particular collection – for which summaries and descriptive section headings are provided – there might be other means for gaining a quick overview of the collection. However, in cases where no such meta data exists, automatic models are even more important.

The curated concept clusters based on word2vec-vectors and the extensive stop word lists used here are by no means mandatory additions to the classic out-of-the box topic model. The classic topic model will still be able to produce topics based on the content of the texts. However, by providing the model with this additional data, it is possible for the user to transfer a part of their mental model of what they find relevant or irrelevant. E.g., we decided not to split up the automatically created clusters of foreign countries (see topic 15), as we were only interested in the concept "foreign country", and not *which* foreign country. In contrast, we decided to split up an automatically created cluster of political party names into individual concepts. Given another mental model of relevant/irrelevant, the opposite decision could have been made.

The full potential of Topics2Themes is not shown here, as the tool is also built to support a more thorough reading of the documents associated with the topics. The user interface provides a fourth panel (not shown in the figure), which makes it possible to manually add themes that the user identifies in the texts (Skeppstedt et al., 2020a, 2021). A possible continuation of the work described here would thus be to use this functionality to perform a manual search for fine-grained reoccurring themes in the documents closely associated with the topics. Another possible continuation would be to study the effect on the concept clusters created, when using a word2vec model trained on the entire report collection (instead of on a subset).

The source code for Topics2Themes is freely available[4] for use and expansion, and so are[5] the word lists, scripts, etc. used here. We hope that the demonstration provided here – of how NLP models can be used for finding reoccurring topics in large document collections – will form an inspiration for future work, e.g., work using Topics2Themes.

Acknowledgments

This study was funded by the Swedish Research Council (project number 2017-00626).

[4] github.com/mariask2/topics2themes
[5] github.com/mariask2/democracy-100-years

References

Eric P. S. Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K. Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, 68(6):1397–1410.

Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics.

David M. Blei. 2012. Topic modeling and digital humanities. *Journal of Digital Humanities*.

Jordan Boyd-Graber, Yuening Hu, and David Mimno. 2017. Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296.

Stefan Dahlberg, Sofia Axelsson, and Sören Holmberg. 2017. The meaning of democracy. Using a distributional semantic model for collecting co-occurrence information from online data across languages. Technical report, Department of Political Science, University of Gothenburg.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 226–231, Palo Alto, California, USA. AAAI Press.

Alfio Ferrara, Stefano Montanelli, and Georgios Petasis. 2017. Unsupervised detection of argumentative units though topic modeling techniques. In *Proceedings of the 4th Workshop on Argument Mining*, pages 97–107, Copenhagen, Denmark. Association for Computational Linguistics.

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.

F. Karsdorp and A. V. den Bosch. 2013. Identifying motifs in folktales using topic models. In *Proceedings of the 22 Annual Belgian-Dutch Conference on Machine Learning*.

Daniel D. Lee and H. Sebastian Seung. 2001. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556 – 562.

Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. 2017. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*.

Austin Van Loon, Sheridan Stewart, Brandon Waldon, Shrinidhi K Lakshmikanth, Ishan Shah, Sharath Chandra Guntuku, Garrick Sherman, James Zou, and Johannes Eichstaedt. 2020. Explaining the Trump gap in social distancing using COVID discourse. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Paris, France. European Language Resources Association (ELRA).

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm University.

Maria Skeppstedt, Magnus Ahltorp, Gunnar Eriksson, and Rickard Domeij. 2021. A pipeline for manual annotations of risk factor mentions in the covid-19 open research dataset. In *Selected Papers from the CLARIN Annual Conference 2020*, Linköping Electronic Conference Proceedings 180.

Maria Skeppstedt, Magnus Ahltorp, Kostiantyn Kucher, Andreas Kerren, Rafal Rzepka, and Kenji Araki. 2020a. Topic modelling applied to a second language: A language adaptation and tool evaluation study. In *Selected Papers from the CLARIN Annual Conference 2019*, volume 172:17, pages 145–156. Linköping Electronic Conference Proceedings.

Maria Skeppstedt, Rickard Domeij, and Fredrik Skott. 2020b. Adapting a topic modelling tool to the task of finding recurring themes in folk legends. In *Proceedings of the Digital Humanities in the Nordic Countries*, pages 388–392. CEUR Workshop Proceedings.

Maria Skeppstedt, Kostiantyn Kucher, Manfred Stede, and Andreas Kerren. 2018. Topics2Themes: Computer-assisted argument extraction by visual analysis of important topics. In *Proceedings of the LREC Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources*, pages 9–16.

Didi Surian, Dat Quoc Nguyen, Georgina Kennedy, Mark Johnson, Enrico Coiera, and Adam G Dunn. 2016. Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection. *J Med Internet Res*, 18(8):e232.

Ivan Vasilev, Daniel Slater, Gianmario Spacagna, Peter Roelants, and Valentino Zocca. 2019. Python deep learning : Exploring deep learning techniques and neural network architectures with PyTorch, Keras and TensorFlow. Birmingham.

A Supplemental Material

Figure 1 shows the graphical user interface of the Topics2Themes tool after 15 topics have been automatically detected, and thereafter provided with a manually authored description.

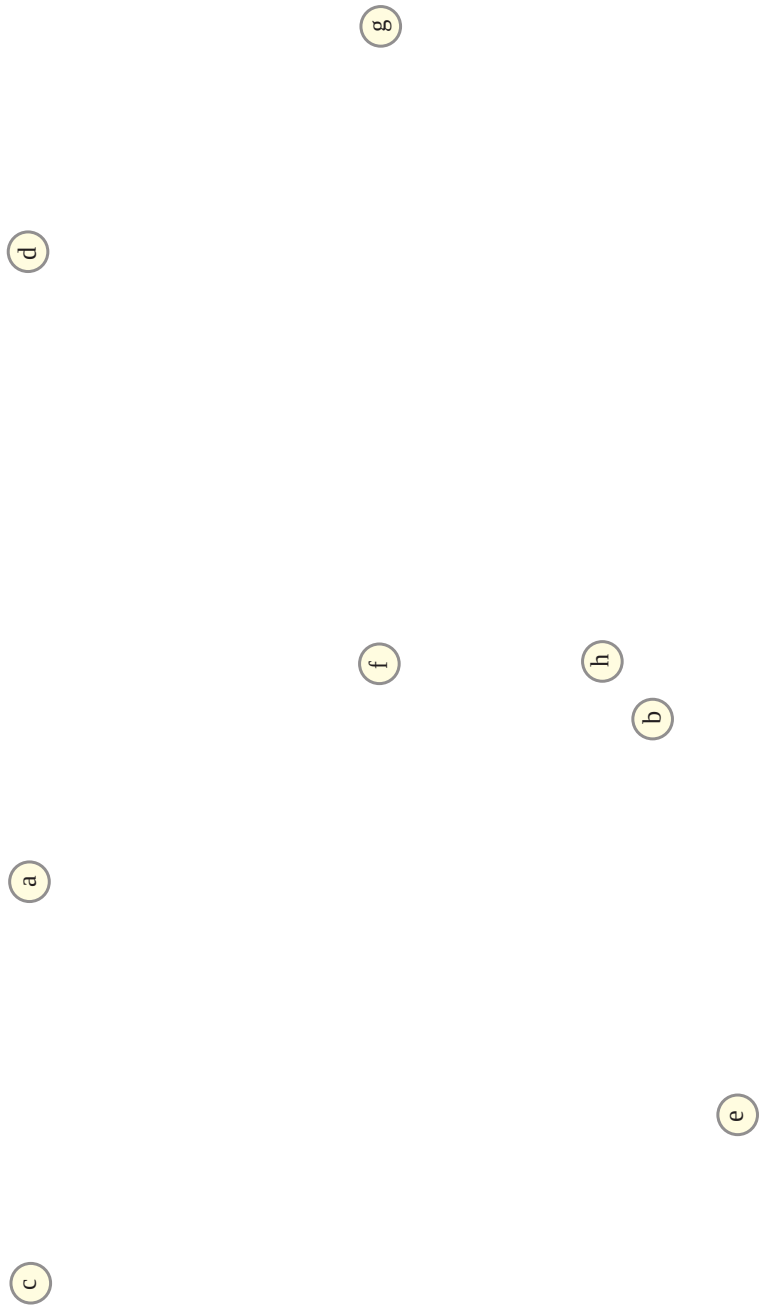

Figure 1: Topics2Themes applied to texts containing the string “demokrati” (“democracy”). The Topics panel (a) shows the 15 topics detected. The topic selected by the user (b) is shown with a blue background, and this topic’s most closely associated words (in the Terms panel, c) and most closely associated texts (in the Texts panel, d) have been positioned as the top-ranked elements in their panels. The Terms panel shows examples of concept clusters, e.g., a large cluster of country names (e). To each text, two labels are attached: The name of the report in which the text appears (f), and the name of the nearest heading under which the text appears (g). There is also a link to the full PDF version of the report in which the text appears (h).