What's on the agenda?

Topic modelling parliamentary debates before and during the COVID-19 pandemic

# Introduction

In democratic countries, a parliament is a central representative and legislative institution. It is composed of elected representatives through which the citizens have a voice in shaping and enacting laws and thus participate in governing all areas of life and social activities. In addition, the parliament often controls the executive branch (Norton, 2002). Due to the parliament’s crucial role in the development of society, its activity has always been an important topic of research in the humanities and social sciences.

In the last two decades, the progress of technology, the increased interest of the media and the citizens in the work of the parliament, and the desire for greater transparency have made the data about the parliamentary activity – including the records of parliamentary debates – more accessible (Norton, 2002). The records are a unique research source as the parliamentary debates reflect the political, societal, and cultural atmosphere of a certain period (Ilie, 2010). Since parliamentary discourse is highly regulated and parliamentary records are often available in digital form, they are a convenient source for building parliamentary corpora. These are temporally limited and structured collections of debate records with added metadata on the speakers and speeches and linguistic annotations (Truan and Romary, 2021).

Usually, parliamentary corpora include large amounts of data that cannot be analysed by hand within a reasonable time frame. Concordancers are popular tools to analyse corpora. You can familiarize yourself with them in a [related tutorial](https://sidih.github.io/voices/index.html) (Fišer and Pahor de Maiti, 2021). Other tools, such as [Orange](https://orangedatamining.com/) (Demšar et al., 2013), used in this tutorial, enable text mining approaches which take large amounts of data to extract patterns and information that are not obvious from the text at first glance (Wiedemann, 2016).

Among other things, text mining techniques have been used for sentiment analysis of parliamentary debates (Rheault et al., 2016; Rudkowsky et al., 2017), for modelling policy conflict between the cabinet parties (Bergmann et al., 2018), for opinion mining (Abercrombie and Batista-Navarro, 2020), modelling argumentation (Petukhova et al., 2015), etc. Among these, topic modelling (Meeks and Weingart, 2012) is one of the most often used text mining techniques in the digital humanities and the one that will be the focus of this tutorial.

This tutorial introduces researchers in the humanities and social sciences to text mining and shows the value of such approaches for research in these scientific fields. The tutorial breaks down the particularities of parliamentary discourse and topic modelling by answering concrete research questions. The analysis is based on the freely accessible corpus of British parliamentary debates [ParlaMint](http://hdl.handle.net/11356/1432) (Erjavec et al., 2021) and the [Orange](https://orangedatamining.com/) tool (Demšar et al., 2013), which enables the use of advanced text mining techniques without any programming knowledge.

# Tutorial overview and instructions

The tutorial is divided into the theoretical and empirical parts. In the theoretical part, the characteristics of parliamentary debates and the ParlaMint corpora are presented in Chapter 3 and the topic modelling method in Chapter 4. The empirical part begins with Chapter 5, guiding the reader through analysis preparation and explaining how to set up the Orange software, import and check the data, and prepare a data sample for the analysis. Chapter 6 moves on to the central empirical part of the tutorial, composed of three related tasks. The tasks use topic modelling and various visualisations to explore the topics of the debates and the prominence of different topics in general and during the COVID-19 pandemic.

All the resources and tools used in this tutorial are freely accessible online. You can find the detailed instructions on downloading the [ParlaMint](http://hdl.handle.net/11356/1432) data and setting up the [Orange](https://orangedatamining.com/) software in Chapter 5. If you are mainly interested in analysing texts in Orange, you can begin with Chapter 5. However, we recommend reading the introductory theoretical chapters containing key information on understanding the data and the topic modelling method. Reading the entire tutorial will reduce the possibility of non-critical use of the method and inappropriate interpretation of the results.

Alongside the descriptions of the procedures, the materials include numerous screenshots that show widget[[1]](#footnote-1) settings and results. At the side of screen, you will find the workflow with the sequence of widgets used for each chapter. The complete workflow is available for download in Chapter 5.1. However, we recommend that you create your own workflow by following the instructions in the tutorial in order to better understand the individual steps of the analysis. The **widget** names are typeset in bold, while *widget settings*, *variable* *names, search queries* and *discussed words* are italicized.

Certain steps of the tutorial might be quite laborious for some computers which results in the process in Orange being stuck or aborted. In this case, you can opt for *Option 2* instructions which will be provided in the relevant parts of the tutorial.

The orange *Try-it-yourself* frames include instructions for independent research and help you consolidate the acquired knowledge.

# Parliamentary debates

The tutorial analyses MPs’ speeches during parliamentary sessions. This chapter focuses on certain general characteristics of parliamentary debates since knowing the data well is crucial for developing research questions and interpreting results.

## Characteristics of parliamentary debates

Parliament is a central political institution. Its institutional nature dictates a clearly defined structure and complex rules of its activities,[[2]](#footnote-2) while numerous informal conventions have developed through history, too (Norton, 2002). These rules differ from parliament to parliament, and they change over time (Sieberer et al., 2011). Therefore, they must be known to the researcher(s) for appropriate analysis design and data interpretation. The research also has to consider local and global political contexts, power relations among MPs and their various public and private roles, and the different audiences present at the debate (e.g., other MPs, guests, and the public) (Ilie, 2010).

Parliamentary sessions follow a clear structure; they have a defined agenda, a designated person leads them, and the floor is passed from one person to another following clear rules (see Proksch and Slapin, 2010). Special rules also apply to specific items on the agenda or the types of debates, e.g., to MPs’ questions and initiatives or interpellations. The structure bears great importance in shaping and limiting parliamentary debates, i.e., the acts of communication in the specific parliamentary environment.

A part of the broader concept of political discourse, parliamentary discourse is its most institutionalised and formal subtype strictly governed by rules (Bayley, 2004). It is a key characteristic of a parliament, which is the central space for the political debate of a community. Here, not only the contents of the debate matter but also the style of speaking, i.e., the discursive strategies that the speakers use in their speeches, and other, non-linguistic circumstances. Therefore, research of parliamentary activities uses an increasingly interdisciplinary approach to the material, which enables comprehensive interpretation of events and processes of causes and consequences (Bayley, 2004; Ilie, 2010).

## Parliamentary corpora

The primary source for the research of parliamentary discourse are the records of parliamentary debates. For most parliaments, they are transcribed and publicly accessible in digital form. Digital form is important both for the public and the research community, yet the actual usefulness of the records depends on the focus of the research (see Mollin, 2007). Much like parliamentary discourse, the records have their particularities which originate from the nature of their source (spoken texts) and from different transcribing traditions of individual parliaments (transcription guidelines differ from parliament to parliament and are generally not made public). As formal written sources, parliamentary records are undoubtedly credible in terms of their content – but not necessarily so when compared to the actual spoken text. The records are [not exact transcriptions of the speeches](https://sidih.github.io/voices/ch4-sl.html#ch4.2-sl) and, therefore, usually lack some or all elements of spoken language (e.g., fillers, false starts) and the information on the non-verbal communication (e.g., interruptions, gestures) (Bayley, 2004). However, they often include additional information or metadata, such as the list of speakers, voting results, the material discussed, etc.

Parliamentary records in digital form are a convenient source for [parliamentary corpora](https://www.clarin.eu/resource-families/parliamentary-corpora), i.e., structured collections of texts enriched with various data. Parliamentary corpora usually include rich metadata that contains varied information on the session (e.g., date, type, agenda), speeches and speakers (e.g., name, date of birth, party affiliation). They are generally abundant in linguistic annotations (e.g., part of speech, basic word form, named entity). Researchers can use these annotations and metadata to perform various kinds of analyses apart from the simple content analysis of the textual data (see Pančur and Šorn, 2016).

Due to their rich metadata and continuity, [parliamentary corpora](https://www.clarin.eu/resource-families/parliamentary-corpora) are invaluable for various research areas, which have grown increasingly interconnected. They include linguistics (Bayley, 2004), history (Piersma et al., 2014), political science (Rheault and Cochrane, 2020), demographics (Kilroy, 2021), etc. Researchers can access parliamentary corpora through [concordancers](https://sidih.github.io/voices/ch3-sl.html) (i.e., web tools for researching and analysing corpora) or [repositories of language data resources](https://www.clarin.si/repository/xmlui/discover?query=parl*&submit=I%C5%A1%C4%8Di&filtertype_2=title&filter_relational_operator_2=contains&filter_2=&query=parl*), which provide access to entire corpora in various formats to be analysed with different tools. The latter option will be chosen for this tutorial (see Chapter 5.2).

## The ParlaMint corpus

The tutorial will use data from the family of [ParlaMint](http://hdl.handle.net/11356/1432) corpora (Erjavec et al., 2021), which contains parliamentary debate records from 17 countries: Belgium, Bulgaria, Croatia, the Czech Republic, Denmark, France, Great Britain, Hungary, Iceland, Italy, Latvia, Lithuania, the Netherlands, Poland, Romania, Slovenia, and Turkey. Most ParlaMint corpora cover the period from 2015 to mid-2020 or more. Designed by the research infrastructure for language resources and technologies [CLARIN ERIC](https://www.clarin.eu/), this corpus family contains 500 million words in 5 million speeches produced by around 11 thousand speakers. The ParlaMint corpora are divided into two sub-corpora: *Reference* (i.e., the reference period) and *COVID,* which mark the periods before and during the COVID-19 pandemic (i.e., *Reference* before November 2019; *COVID* from November 2019 onwards).

Each national corpus has been encoded following the same scheme based on the [Parla-CLARIN](https://clarin-eric.github.io/parla-clarin/) encoding recommendations (Erjavec and Pančur, 2019). Common encoding ensures that the national parts of ParlaMint are comparable which makes ParlaMint a valuable resource for comparative and transnational analyses that have so far been difficult to perform. Furthermore, ParlaMint covers a diverse set of European countries, which increases the possibilities of exploring different non-Western parliamentary democracies and acquiring new knowledge about parliamentary systems. This is crucial given that previous research mostly centred around Western countries and especially because comparative research proved very important in improving our understanding of positive and negative parliamentary practices and advancing the development of parliamentary systems (Norton, 2002).

The tutorial uses the British ParlaMint corpus, ParlaMint-GB, which encompasses the debates from the House of Lords and the House of Commons. The House of Lords currently has 300 members, most of whom are elected, while the rest are appointed. It reviews bills proposed by the House of Commons. The House of Commons consists of 650 elected members and has the primary legislative function. ParlaMint-GB covers four parliamentary terms between January 2015 and March 2021, and holds around 100 million words (Erjavec et al., 2022). The tutorial will use the [linguistically annotated version of the corpus 2.1](https://www.clarin.si/repository/xmlui/handle/11356/1431) (Erjavec et al., 2021b), which includes sentence segmentation (the sentences are delimited), tokenization (tokens, numbers and punctuation marks are defined as the basic analytical unit), lemmatisation, morphosyntactic annotations, and named entities (see Chapter 5.3).

# Topic modelling

To analyse parliamentary debates, we will use topic modelling, one of the text mining techniques used for researching large data sets. Given that several topic modelling methods exist, it is vital to know the advantages and disadvantages of each to choose the one that yields optimal results, which would provide quality results and ensure a critical interpretation of the results (Shadrova, 2021). In this chapter, we present the selected method of topic modelling and some examples of its application to parliamentary discourse.

Topic modelling is a popular technique for automatic text analysis, which extracts the main topics in a corpus. A topic model assigns each document in the corpus to one or more topics. These topics are not actual text topics but rather sets of words that co-occur with high probability and hence form a single topic. The researcher must then manually define or name the topic described with these sets of words.

Various methods (algorithms) can perform topic modelling on a corpus (Vayansky and Kumar, 2020). One of the most frequent ones is LDA or *latent Dirichlet allocation*, which we will use in our analysis. The method was developed by Pritchard et al. (2000) and adapted for text analysis by Blei, Ng, and Jordan (2003). The method is best suited for processing large textual data sets which cannot be analysed manually due to their size.

## The LDA method

The LDA method includes the following steps, performed in iteration:

1. The algorithm first randomly allocates topics to the words in the corpus.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **word**  **document** | **epidemic** | **crisis** | **tax** | **economy** |
| **doc1** | topic 1 | topic 2 | topic 2 | topic 1 |
| **doc2** | topic 1 | topic 1 | topic 2 | topic 1 |
| **doc3** | topic 2 | topic 1 | topic 1 | topic 2 |
| **doc4** | topic 1 | topic 2 | topic 2 | topic 2 |

1. Next, the algorithm counts the number of times a particular topic appears in each document (left table) and the number of times a certain topic is assigned to each word (right table).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  | **topic 1** | **topic 2** | | **doc1** | 2 | 2 | | **doc2** | 3 | 1 | | **doc3** | 2 | 2 | | **doc4** | 1 | 3 | | |  |  |  | | --- | --- | --- | |  | **topic 1** | **topic 2** | | **epidemic** | 3 | 1 | | **crisis** | 2 | 2 | | **tax** | 1 | 3 | | **economy** | 2 | 2 | |

1. The algorithm then assumes it no longer knows the topic of a given word.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **epidemic** | **crisis** | **tax** | **economy** |
| **doc1** | **?** | topic 2 | topic 2 | topic 1 |
| **doc2** | topic 1 | topic 1 | topic 2 | topic 1 |
| **doc3** | topic 2 | topic 1 | topic 1 | topic 2 |
| **doc4** | topic 1 | topic 2 | topic 2 | topic 2 |

1. Then, it updates both tables from step 2 by once again computing the frequency of topics in the corpus (left table) and the frequency of words in the topics (right table).

|  |  |  |
| --- | --- | --- |
|  | **topic 1** | **topic 2** |
| **doc1** | **1** | 2 |
| **doc2** | 3 | 1 |
| **doc3** | 2 | 2 |
| **doc4** | 1 | 3 |

|  |  |  |
| --- | --- | --- |
|  | **topic 1** | **topic 2** |
| **epidemic** | **2** | 1 |
| **crisis** | 2 | 2 |
| **tax** | 1 | 3 |
| **economy** | 2 | 2 |

1. It computes the strength of the **connection between a document and a topic** (the probability of the topic in a document: the blue rectangle) and the **connection between the topic and a given word** (the probability of a word in a topic: the red rectangle).

[Figure0.1\_LDA-eng]

1. The purple rectangle is a product of the red and the blue rectangle and represents the probability of a word in each topic. Based on the computed probability (purple rectangles), the method determines which topic will be assigned to a given document (green star which denotes a random allocation of the topic to the document, based on the computed word-topic probabilities).

[Figure0.2\_LDA-prob-eng]

In short, the algorithm assigns a new topic (green star) based on the probability (purple rectangle). The probability distribution of topics in the document is based on the Dirichlet distribution, which postulates that the probability is never zero. Non-zero probability means that each word has at least a small chance of belonging to a less frequent topic and, concurrently, that even a lesser topic is present in a document. Once the topic is assigned to the word, the documents-words table is updated.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **epidemic** | **crisis** | **tax** | **economy** |
| **doc1** | **topic 2** | topic 2 | topic 2 | topic 1 |
| **doc2** | topic 1 | topic 1 | topic 2 | topic 1 |
| **doc3** | topic 2 | topic 1 | topic 1 | topic 2 |
| **doc4** | topic 1 | topic 2 | topic 2 | topic 2 |

1. Based on the new value from the table in step 6, the algorithm updates both tables: the topic-document and the word-topic table.

|  |  |  |
| --- | --- | --- |
|  | **topic 1** | **topic 2** |
| **doc1** | 1 | **3** |
| **doc2** | 3 | 1 |
| **doc3** | 2 | 2 |
| **doc4** | 1 | 3 |

|  |  |  |
| --- | --- | --- |
|  | **topic 1** | **topic 2** |
| **epidemic** | 2 | **2** |
| **crisis** | 2 | 2 |
| **tax** | 1 | 3 |
| **economy** | 2 | 2 |

1. The procedure is repeated until the topic assignments stop changing. The result is a topic model. The topic is defined by a set of words which frequently co-occur in the text.

Above, we have described the Gibbs sampling version of LDA. Please note that Orange uses variational inference instead of Gibbs sampling. Gibbs sampling is much more precise, while variational inference is faster for larger data sets.

## Characteristics of the LDA method

LDA is based on multiple assumptions. The first one is that the topic is defined by the words that frequently appear together. LDA is a language-independent method since it merges words into groups based on their occurrence in the text and not on their meaning. The same method can thus be used on corpora in different languages. At the same time, each word can be assigned to multiple topics; however, its probability in each topic will vary.

The second assumption is that not all topics in the corpus appear equally often, and that they are unrelated. LDA does not define potential connections between the topics; however, it shows the probability distribution of topics in the text, which defines their importance in each document. Each document contains several topics, with one topic usually standing out (i.e., the document contains more words associated with the main topic compared to other topics).

The third assumption is that the number of topics is predefined. The researcher must first set the number of topics into which the algorithm sorts the documents. The optimal number of topics for a given corpus is the one at which it is easiest to interpret viable topics for given sets of words that must be informative in terms of the research problem. Researchers, therefore, typically apply the topic modelling procedure several times for a given research problem, each time setting a different number of topics, and assessing the informativeness of the word sets that the algorithm extracted from the corpus. Some researchers also use additional statistical tests to adjudicate between results of models with different numbers of topics (Smith and Graham, 2019). Even though a suitable number of topics differs from case to case, the usual number ranges between 5 and 50 topics (Arun et al., 2010) and many papers go with 20 topics (Zhao et al., 2015; Gkoumas et al., 2018; Rosa et al., 2021).

The fourth assumption is that word order in a corpus is not important. LDA works based on the so-called bag of words which does not consider the linguistic structure or specific connections between the words. This assumption is problematic because the word order is one of the key characteristics of language.

The results are importantly affected by document order, as the method randomly assigns topics to documents at the beginning. If the document order changes, the initial assignments will also change. The temporal sequence (i.e., the timestamp) of the documents can also importantly affect the characteristics of the documents (Vayansky and Kumar, 2020).

## Data preprocessing

Before topic modelling, the data requires preparation which usually includes tokenisation (splitting the text into tokens, usually words, numbers, and punctuation), lemmatisation (assigning the base form to each token), and part-of-speech tagging (assigning the part of speech, e.g., verb, to each token). The procedure enables us to perform topic modelling on a single word form. Research has shown that lemmatisation and limiting the tokens to nouns improve the algorithm's speed and results. The improvement shows in the coherence or the sensible relations between word sets, based on which it is easier to assign a topic (Martin and Johnson, 2015). Differentiating by part-of-speech tags can also be used for answering different research questions. Van der Zwaan et al. (2016) performed topic modelling on nouns to retrieve topics. They then ran the algorithm on verbs, adjectives, and adverbs and used the results to elicit the positions of the MPs. Before using LDA, we usually remove overly common words from the corpus: words that can be too general (e.g., pronouns, prepositions) or too specific for the genre (e.g., words of address, such as *esteemed* in the parliamentary corpus). Depending on the research problem, very rare words, punctuation, capital letters, etc. can also be removed (Smith and Graham, 2019)

## Limitations of LDA

One of the limitations of LDA is that the method requires long texts for good results as it is based on word distributions, which are spurious in shorter texts. LDA is thus not appropriate for topic modelling of tweets, user reviews, or poetry. Even though it could be used to analyse, for example, Facebook posts (see Serrano et al., 2019), it is recommended to use other topic modelling methods for such tasks (Albalawi et al., 2020; Morstatter et al., 2018). Parliamentary speeches are usually long enough to achieve good results with LDA. However, certain speeches can be very short, and it is wise to remove them before running the analysis (Chapter 5.4).

Another limitation is the assumption that words, just like the documents and the topics, are not co-dependent, which is linguistically imprecise on the one hand and does not allow for an analysis of correlations between words or between documents on the other. LDA suffices since these correlations are not the focal point for many research problems. However, if correlation is important (e.g., if we are interested in topic progression over time), there are more suitable methods, such as dynamic topic modelling (Müller-Hansen et al., 2021).

LDA will also underperform if the text does not address the topic coherently but touches upon the topic with a few words only. On the other hand, the method works extremely well for longer, thematically well-defined texts, such as news, academic papers, political speeches, and certain literary genres.

The next limitation is related to the number of topics the researcher has to define autonomously and is usually the result of trial and error. Allen and Murdock (2020) warn about overly specific topics representing only small sections of the text when the number of requested topics is high, making it difficult to establish thematic relations between texts. Conversely, when the number of topics is very limited, they will frequently be too general and thus uninformative to the research.

As a final limitation, we can mention the difficulty of interpreting topic modelling results, including how the results are published. As the result of topic modelling are individual sets of words, there is a danger that the researcher will recognize a pattern in them even when none is present, meaning they will identify the topics they had expected (Shadrova, 2021). It is thus vital to consider the number of inspected words when defining a topic. The results can differ if the researcher assigns topics based on the first ten or thirty words provided by the algorithm (Allen and Murdock, 2020). Qualitative reading and understanding the original text segments in which the top listed words appear are crucial for accurately interpreting word sets and identifying topics. When working in a group, defining the common guidelines for topic identification in advance is also recommended.

Topic modelling with the LDA method is thus not a one-size-fits-all solution that could provide the researcher with robust conclusions without a critical analysis of the results. Understanding the limitations of different topic modelling methods is key to successfully using them for research purposes, since quantitative, automated methods can successfully augment the researcher's analytical abilities, but they cannot replace human interpretation (Grimmer and Stewart, 2013). Nonetheless, the topic modelling technique provides an important advantage over the manual approach, specifically with regard to the processing of large data sets, enabling a more robust data-based generalisation than using a small-sample analysis (Jacobs and Tschötschel, 2019). Moreover, topic modelling enables greater objectivity of the results (Müller-Hansen et al., 2021), even though the technique is not entirely objective due to the aforementioned topic definition process based on word sets.

On the other hand, this is one of the advantages of the technique as the algorithm does not give direct answers but forces the researcher to consider the context when forming the final results (Schmidt, 2012). Topic modelling also enhances systematisation of the analysis and enables a comparatively better reproducibility of the results (Jacobs and Tschötschel, 2019). The popularity and relevance of the technique for the research in the humanities and social sciences are evident from the many publications that use topic modelling as a part of their methodology (see Chapter 4.5).

## Topic modelling of parliamentary debates

In the humanities and social sciences, particularly in political science, topic modelling is increasingly used as an important technique to complement established, qualitative analytical approaches in analysing large data sets. The results of topic modelling may inform the qualitative analysis (e.g., the researcher can identify relevant texts about a specific topic) or can be used as the principal outcome of the analysis. In this chapter, we provide some examples of both from recent applications of the topic modelling technique to parliamentary data.

Topic modelling allows us to **identify the topics** addressed in parliamentary speeches. Schuler (2020), for example, used LDA to analyse the debates in the Vietnamese parliament and compared the results with topics from the news, also extracted with LDA, and with the list of areas under the direct management of the Communist Party (CP). He analysed whether MPs in an authoritarian system, such as the Vietnamese, express their opinions and actively debate important topics. He discovered that they debate only topics outside the CP's direct management, with the party encouraging such debates to pressure the government and blame it for the outcomes of the policies. However, the party does not encourage debates pertaining to areas directly managed by the CP committees. Moreover, debates concerning topics open for discussion do not involve all MPs but mostly those who are not members of the CP and were elected as full-time representatives.

Chizhik and Sergeyev (2021) also used topic modelling to discover topics in parliamentary debates by analysing three decades of Russian MPs' speeches. They researched whether the activity of the parliamentary parties is related to the public's scepticism regarding the multiparty system as a basis for democracy. They established that parties in the Russian parliament focus mostly on foreign affairs, the economy, and the balance of power between different branches of the government. At the same time, other social issues generate much less debate. Furthermore, speeches from all parties, but especially from the long-established ones, show a strong prevalence for ideological and propagandistic discourse.

We saw how topic modelling enables acquiring a general overview of the material, which can suffice if the researcher’s aim is, for example, to observe the frequency of topics under consideration in the parliament. But topic modelling can also be used to retrieve more specific results. As parliamentary corpora are usually rich with metadata, **topics** **can be explored in relation to other variables** (such as gender, age, party affiliation, mandate etc.), which elicits the topics that stand out most when the selected variable changes. We can thus observe how popular a topic was through time or with a certain party. Curran et al. (2018) used LDA and the analysis of complex networks to elicit topics in the New Zealand parliament, which they then related to MPs and the parties. In this way, they not only retrieved popular topics of parliamentary debates for different periods and interpreted them in the context of external events (e.g., the 2011 earthquake) but also defined the interest of a party in each topic. Their results showed that the Labour Party debated the real estate crisis much more ardently than the then-governing National Party. The latter claimed most of the debate, while the contribution of other parties decreased over time. The MPs' specialisations for different topics also decreased, which is evident from the large number of topics addressed by the majority of the MPs.

De Campos et al. (2021) used LDA in combination with the available metadata to create thematic profiles of Spanish MPs which reflect the subject matters they discuss in the parliament. Metadata was also used in research by Høyland and Søyland (2019). In 1919, Norway changed its electoral system to become substantially more dependent on party politics, reducing MPs’ autonomy. Høyland and Søyland investigated whether the change in the political system affected the topics in parliamentary debates. They used a version of LDA called structural topic modelling (STM), which considers both the word distributions and the selected metadata when computing the results. They determined that the topic distribution clearly shows that institutional organisation influences the behaviour of the MPs. After the reform that emphasised party politics, the topics showing clear ideological differences between the parties became more frequent, while MPs gave fewer speeches that directly criticized other MPs. Furthermore, MPs more frequently discussed topics of general interest (e.g., the educational system) and more rarely topics related directly to the issues of their constituents (e.g., improving the infrastructure of a remote town).

Topic modelling enables **exploring the context and selecting topics related to a given concept**. Müller-Hansen et al. (2021) used a version of LDA called dynamic topic modelling (DTM), which enables the analysis of topics through time.[[3]](#footnote-3) They analysed seventy years of German parliamentary debates on coal and explored how they changed through time. The debates from the early years of the corpus show that the MPs considered coal the driver of economic progress and the guarantee of energy safety. Conversely, in recent years MPs primarily talked about energy transition, a general departure from coal, and the flourishing of renewable energy sources. Furthermore, the researchers also established that smaller and younger parties (e.g., the Greens) talk about coal in the context of energy transition and environmental protection more frequently than the other parties.

Topic modelling can **explore the context of a topic or the interplay of topics**. Blätte et al. (2020) aimed to discover how frequently migration is addressed in common European politics. They used LDA to create a topic model of parliamentary debates from Austria, France, Germany, and the Netherlands. Then, they selected the three topics which best represented migration and European matters and retrieved all speeches with a high frequency of the two issues. The results show that the debate on migration was predominantly an internal issue in the larger two countries investigated (France and Germany). Particularly in Germany, the European aspect practically disappeared, while the smaller two countries (Austria and the Netherlands) had a larger share of speeches discussing migration from the European perspective.

In this tutorial, we partially rely on the methodology used by Curran et al. (2018) in the analysis of speeches in New Zealand discussed earlier. However, as our analysis covers a shorter time, we will not split the data into time slices. As seen in the literature, the analysis could be upgraded with structural topic modelling, where we could consider, for example, the party affiliation of the speakers and observe the differences among them. We could also analyse the entire ParlaMint-GB corpus and use dynamic topic modelling to observe the differences in time. Nevertheless, to compare the pre-pandemic and pandemic periods, the use of the LDA method is adequate.

# Preparing for the analysis

This chapter begins the empirical part of the tutorial. It will lead us from setting up the software, importing and checking the data, to preparing and preprocessing the sample for analysis.

For the tasks below, you will need about 1 GB of disk space to download and install Orange with Miniconda, 2.3 GB of space for the original ParlaMint files (if you wish to work with the original data), and 1.9 GB for all three versions of pre-formatted pickle files (if you wish to speed up the analysis). Keep in mind that topic modelling is resource-intensive and might get slow on computers with insufficient RAM.

## Orange: setup and use

The analysis will be performed in [Orange](https://orangedatamining.com/) v3.32.0, a Python-based open-source software for data analysis (Demšar et al., 2013). In Orange, the *Text* add-on offers a special tool kit for text mining. Orange is based on visual programming. The analysis is performed through a data analysis workflow, i.e., a series of steps or widgets that the user selects, thereby not requiring any coding knowledge.

First, we download the software from [orangedatamining.com](https://orangedatamining.com/) onto the computer. We open the downloaded file and follow installer instructions. Once we open the program, we install the *Text* add-on (v1.10.0) by clicking the *Options* tab and selecting *Add-ons* in the drop-down menu*.* In the window that opens, we tick the *Text* field, and confirm the add-on installation by clicking the *OK* button (Figure 1).

Figure 1: Installing the Text add-on.

To complete the installation, we must restart Orange.[[4]](#footnote-4) When we do so, the left-hand menu will contain the *Text Mining* tab with various widgets (e.g., *Corpus, Bag of Words*) intended for text analysis (Figure 2). On the right-hand side, there is a white field called *canvas*. We will place the *widgets* on the canvas and connect them in an analytical workflow.

Figure 2: Home screen after restarting Orange, having installed the Text add-on.

The widgets can be added on the canvas by dragging them from the menu on the left and dropping them onto the canvas or by right-clicking on the canvas to open a drop-down menu, typing the widget's name, e.g., **Corpus**, and pressing **Enter**. A double click on the widget will open a settings window. Every widget has an input on the left, an output on the right, or both. They are marked with a dashed line at the side of the widget. In Orange, analysis always runs from left to right, never in the opposite direction.

Using the same steps as when adding the **Corpus** widget, we add the **Corpus Viewer** widget. By double-clicking it, an empty window opens: the widget has not yet received any data to analyse. We can send the data to the widget by connecting the two widgets, specifically by using the mouse to connect the right-hand dashed line of the **Corpus** widget with the left-hand dashed line of the **Corpus Viewer** widget (Figure 3). Then, we double-click **Corpus Viewer** again, and this time, the window will display data.[[5]](#footnote-5)

Figure : A connection between the widgets.

The tutorial will describe how to build a workflow to perform topic modelling of parliamentary debates and explore the topics with additional visualisations. Although you can [download](https://www2.sistory.si/publikacije/material/parlamint/tutorial-eng.ows) the entire workflow, we recommend you follow the individual tutorial steps and create the sequence of widgets by yourselves; this is how you will best understand the separate analysis phases.

## Loading data into Orange

The ParlaMint-GB corpus holds parliamentary speeches from the 2015 to 2021 period. Since the empirical part of the tutorial will compare the speeches made before and during the COVID-19 pandemic, the data must first be limited to two comparably long periods before and during the pandemic. The pandemic period included in [the ParlaMint-GB corpus](https://www2.sistory.si/publikacije/material/parlamint/ParlaMint-GB.conllu.zip) lasts from November 2019 onwards (see Erjavec et al., 2022). We wish to choose similarly long periods; hence we will select 2019 for the pre-pandemic and 2020 for the pandemic period.[[6]](#footnote-6)

The analysis will use the British part of the linguistically annotated corpus of parliamentary data ParlaMint 2.1 (see Chapter 3.3) which is available in the CoNLL-U format. To load the data, follow Option 1 below. If you are experiencing problems, try Option 2 instead.

|  |  |
| --- | --- |
| Option 1: follow the tutorial | Option 2: speed up the analysis |
| [Figure0.3] | [Figure0.4] |
| You will get the most comprehensive understanding of the entire process (from the preparation of data to the final results) if you follow along the tutorial. Certain steps might take your computer some time to process, so please be patient.  If you decide for this option, download [the CoNLL-U files](https://www2.sistory.si/publikacije/material/parlamint/ParlaMint-GB.conllu.zip)[[7]](#footnote-7) and continue below.[[8]](#footnote-8) | If you are unable to load the CoNLL-U files, download [the .pkl data](https://www2.sistory.si/publikacije/material/parlamint/ParlaMint-GB.pkl), load the data to Orange with the **Corpus** widget and continue with chapter 5.3. |

If you are using the CoNLL-U files, add the **Import Documents** widget to the canvas to import data into Orange. First, we open the widget with a double click, and in the first line, select the folder in which we have stored the data (Figure 4). It is not necessary to confirm the import; it happens automatically when we select a folder. Below, we tick the *Lemma* and *POS tags* options, which will import the lemmas and the parts of speech with the speeches. A speech by an individual MP at a specific session will be presented as an individual document. At the bottom of the window, the software will inform us that we have imported 180,565 documents or files.

Figure : Import data window.

For a better understanding of the data structure, here are a few characteristics of the CoNLL-U format. CoNLL-U is a type of TSV format in which tab characters separate the values. In natural language processing, it is used to represent linguistically annotated texts as its distribution of texts and annotations in columns allows for a straight-forward computer processing. Each sentence is considered one segment in this format. The text is in a vertical or long format, i.e., one word per line, which enables a clear overview of added linguistic annotations. There are metadata at the beginning of every sentence (e.g., speech ID, sentence ID, and text) (Figure 5).

Figure : Data in the CoNLL-U format: the first sentence in the file ParlaMint-GB\_2019-01-07-commons.conllu.

## Data overview

[Workflow1]

Before we begin the analysis, let us make sure that the data we have uploaded is correct. We can do this in the **Corpus Viewer** widget. First, we add it to the canvas and connect it from left to right with the previous widget. Then, we double click **Corpus Viewer** to open it and display a list of documents. In our case, these are the individual speeches (Figure 6). By clicking on the list, we can see different speeches. Holding the *Shift* key while clicking will display several speeches simultaneously.

Figure : Data overview in the Corpus Viewer widget.

In the top left-hand corner, we can see the basic information on the corpus: the number of *tokens, types*[[9]](#footnote-9) and the number ofdocuments that match the *regexp filter* if we use one (*matching documents*). As the filter is now empty, all the documents are displayed (180565/180565). The last information, *matches*, is the number of documents matching the search query that we can enter in the *RegExp Filter*.[[10]](#footnote-10)

The viewer on the right displays numerous metadata[[11]](#footnote-11) on the speeches and the speakers.[[12]](#footnote-12) Every speech has data on the *name* of the session it belongs to and a unique *utterance* designation, the last number of which marks the consecutive number of the speech in the given session. We can read the entire speech under the *content* variable. Another important piece of data is the *subcorpus* variable; it marks the time the speech was given (*reference* stands for the speeches delivered before November 2019, while *COVID* stands for those given since).

Next is the data on the speaker: their *role* (chairperson or regular speaker), their *type* (MP or guest), their affiliation (*Speaker party*), their *party* *status* (opposition or coalition), their *name*, gender, and year of *birth*.

## Preparing and preprocessing the subcorpus

As shown in Chapter 3, parliamentary discourse is marked by a clear structure with numerous typical phrases such as “*Ms/Mr … has the floor”, “Thank you for the floor, honourable member”,* or *“The agenda is approved”*.Certain expressions that guide the discussion are especially typical of the chairpersons; others are merely parts of polite communication. Due to their nature, such phrases are very common in parliamentary corpora but not of interest for topic analysis. As they would only represent noise in the results, they should be removed from the corpus. Although it is impossible to remove them automatically in their entirety, their number can still be significantly reduced. We can do this by preparing a subcorpus and removing every chairperson’s speech and speeches shorter than 50 words (Figure 8).

“It is very good to welcome the hon. Member for North West Durham (Laura Pidcock) back to the House.” (John Simon Bercow, House of Commons, 7 January 2019).

The decision is based on a manual overview of the corpus, proving that such speeches mainly interject or express thanks. Other related research (Curran et al., 2018) has done similar filtering. Although these speeches are not necessarily of a purely procedural nature, their brevity still makes them less appropriate for topic modelling following the LDA method, which requires longer texts to achieve good results (see Chapter 4). Along with the too-short speeches, the sample will also exclude speeches from guests of the parliament.

### **Removing unwanted speeches**

[Workflow2]

|  |  |
| --- | --- |
| Option 1: follow the tutorial | Option 2: speed up the analysis |
| [Figure0.5] | [Figure0.6] |
| Continue with the steps below. | If the sampling process is slow, first open a new session in Orange. Then download [ParlaMint-GB-sample.pkl file](https://www2.sistory.si/publikacije/material/parlamint/ParlaMint-GB-sample.pkl) and load it with the **Corpus** widget. The file contains the sample we will create in the next step. Continue with Chapter 5.4.2. |

We will need the **Statistics** and **Select Rows** widgets to create the sample. We place the **Statistics** widget on the canvas and connect it with the **Import Documents** widget – we do not change the settings; they will instruct the **Statistics** widget to perform a *word* and *character count* in the documents. The **Data Table** widget, which we connect to the **Statistics** widget, allows us to see the two columns at the far right of the window with the number of words and characters in each speech (Figure 7). Now that the data on the speech length is known, we can use the **Select Rows** widget to select only the speeches that match the desired length.

Figure : Word count and character count columns are added to the data (far right).

First, we add the **Select Rows** widget (Figure 8), connect it to the **Statistics** widget, open it, and set three conditions (by clicking the *Add condition* button):

* The first criterion will set the speech length threshold – we limit the *Word count* variable with the *is greater than* option and enter the desired minimal length, in our case 50, which will limit the sample to speeches with 51 words or more;
* The second criterion will exclude the session chairpersons – we set the *Speaker role* variable by selecting the *is* and the *Regular* parameter;
* The third criterion will only keep MP speeches in the sample – we set the *Speaker type* variable by selecting the *is* and the *MP* variable, which will exclude the speeches given by the guests.

Figure 8: A selection of speeches with more than 50 words given by regular MPs.

The data at the bottom of the **Select Rows** widget tells us that the number of speeches has dropped to 130,453 (from the previous 180,565; the complete information on the output can be accessed by clicking the numbers at the bottom).

### **Removing unwanted words**

[Workflow3]

To achieve good topic modelling results, we must also preprocess the data (see Chapter 4.3). We can do this by filtering in the **Preprocess Text** widget, which we insert immediately after the **Select Rows** widget. Before connecting the two widgets, we will first set the parameters to ensure smoother operation. Nonetheless, please note that the preprocessing step may take quite some time to complete. To set the parameters, we first open the **Preprocess Text** widget. In it, we will see the default steps for text preprocessing (their order and settings can be modified). As the data has already been tokenized and the words transformed so that they all begin with a lowercase letter (Chapter 5.3), we can remove the Transformation and Tokenization steps by clicking on the cross in the upper left-hand (Mac OS) or right-hand corner (Windows).[[13]](#footnote-13) We are left with the Filtering step, where a few settings need to be changed (Figure 9):

* Clearing the *Stopwords* option – as we will only focus on nouns, we will not need this option which excludes function words such as pronouns, conjunctions, and prepositions;
* Selecting the *Document frequency* option and the *Absolute* measure, where we set the span from 10 to max[[14]](#footnote-14) (total number of speeches) – the analysis will therefore exclude any words that appear in fewer than ten speeches, e.g., exclude the extremely rare words that do not influence the forming of specific topics;
* Selecting the *POS tags* option because the data contains part-of-speech tagging (see Chapter 3.3). The default setting of this option includes only nouns and verbs in later analyses. However, as nouns proved to be the most useful part of speech in topic modelling (see Martin and Johnson, 2015), we will only keep those and eliminate verbs.

Figure : Setting the Preprocess Text widget.

Once we have set the parameters (as seen in Figure 9), we can connect **Preprocess Text** with **Select Rows**.

We can visualise the words that appear most often in the sample with a word cloud. To do so, we connect **Word Cloud** to **Preprocess Text**. The word cloud will only feature nouns; the size of the word is proportional to its frequency (Figure 10). The displayed words reflect the parliamentary genre of the data. The list on the left shows that the most frequently used word is *people*, which appears 99,103 times. The parliament is voted by the people and works for the people, so such a result is not very surprising.

Figure : The most frequent words in the subset after preprocessing.

Preprocessing is an important part of processing text data, but every step must be clearly defined. Every decision influences the results, which must be considered during interpretation. Please note that filtering in Orange does not modify the original data. Therefore, the preprocessing step does not erase the original corpus from the **Select Rows** widget but only adds information on tokens in the **Preprocess Text** widget.

# Analysis of parliamentary speeches

This chapter is divided into three practical tasks in which we use topic modelling and visualizations to explore the content of parliamentary debates before and during the COVID pandemic. We will answer the following questions:

* Task 1: Which topics are characteristic of the corpus?
* Task 2: Which topics did MPs debate on the most?
* Task 3: Which topics were more frequent before and during the pandemic?

## Topics of parliamentary speeches

In this chapter, we will first prepare a numeric description of the corpus, which is necessary for the LDA methods. Next, we will extract the topics and name them. In the end, we will observe how these topics are distributed in the corpus and how we can find the speech on a given topic.

### **Computing document vectors**

[Workflow4]

Before we can begin topic modelling, we need preprocessed data and a vector representation of speeches. We have already preprocessed the corpus (see Chapter 5.4). To compute the vector representation of the speeches, we will use the **Bag of Words** widget, which constructs a numeric description of the speeches. Using this description, one can compute word distributions for topics or, in other words, perform topic modelling. The numeric description, which we retrieve with bag of words, contains words in columns, with their values representing the number of times a word appears in a given speech. Each speech is thus characterised with a vector, which represents the content of the speech. The more frequent the word, the more prominent the vector of the speech is in the direction of the word.

However, not all words are equal. Some words in the corpus are procedural or genre-specific (see Chapter 5.4), stopwords (i.e., pronouns, articles) or not specific for a given speech. The word *thank*, for example, appears in thematically heterogeneous speeches, as many MPs thank the speaker before them. Hence the word is not thematically informative. We would like to weigh the words so that the words specific to a given speech have a higher weight than those that frequently appear across the entire corpus. This type of weighting is called TF-IDF or *term frequency-inverse document frequency* (Jones, 1972) and can be selected in the **Bag of Words** widget.

We add **Bag of Words** directly to **Preprocess Text** and set the parameters to keep the *Count* under *Term Frequency* and select *IDF* under *Document Frequency* (Figure 11).

Figure 11: Setting the bag-of-words parameters.

### **Topic modelling**

[Workflow5]

Now that we have the vector representation of speeches, we can begin with topic modelling. If the **Bag of Words** process completed successfully, continue with *Option 1.* However, if the computing is taking too long, follow the instructions under *Option 2.*

|  |  |
| --- | --- |
| Option 1: follow the tutorial | Option 2: speed up the analysis |
| [Figure0.7] | [Figure0.8] |
| Add **Topic Modelling** to the canvas and connect it to **Bag of Words.** From here, continue as described below. | If the computation is taking too long, first open a new session in Orange. Then download [ParlaMint-GB-bow.pkl](https://www2.sistory.si/publikacije/material/parlamint/ParlaMint-GB-bow.pkl) and load the data into Orange with the **Corpus** widget. The file contains a pre-constructed bag-of-words matrix.  Now add the **Topic Modelling** widget and connect it directly to the **Corpus.** Continue as described below. |

Open the **Topic Modelling** widget and select the LDA method. Let us set the *Number of topics* to 20. The number of topics is up to the researcher, but related work shows that with larger corpora, 20 is a good choice (see Chapter 4). On the right, we see twenty groups of words that characterize the topics in the corpus (Figure 12). Some topics are easy to name, while others appear a little tricky. In the following chapter, we will further explore the topics which will help us to define the topics better.

Figure 12: LDA results for twenty topics.

It is important to note that LDA is a stochastic method, which means that it returns different results at each run as it is based on a random initial topic assignment. In Orange, we bypass this characteristic by setting a fixed starting point, which enables the reproducibility of the results.

Try it yourself: use a different number of topics, say 5 or 10. Does a smaller number of topics give better results? What about setting a large number of topics, say 50?

### **Topic definition**

[Workflow6]

LDA returns the ten words most related to each topic.[[15]](#footnote-15) However, these words are sometimes not informative enough to enable defining a topic. Hence, we use **LDAvis** (Sievert and Shirley, 2014) to help us define the themes. The main advantage of this visualisation is that it scores words based on how *specific a word is in the topic* vs *in the corpus*. The value of the lambda parameter, which can be set in the widget with the *Relevance* slider, can be between 0 and 1, where 1 displays words based on how specific they are to the topic (as shown in the **Topic Modelling** widget) and 0 displays words based on how frequent they are in the corpus. The exclusivity of the word in the corpus is called *lift*, which represents the ratio between the frequency of the word in the topic and the frequency of the word in the corpus.

Connect **LDAvis** to **Topic Modelling** and ensure that the right data are sent to the input. The LDAvis widget needs a table of word frequency per topic, which is present in the *All Topics* output. The output can be edited by clicking on the connection between the widgets and connecting the *All Topics* signal to *Topics* (Figure 13).

Figure 13: Setting correct input data for the LDAvis widget.

By default, **LDAvis** shows words in balanced order with the same proportion of exclusivity of the word in the topic and the exclusivity of the word in the corpus, which usually gives good results.

The balanced relevance setting gives a different, more informative set of words than seen in the **Topic Modelling** widget (Figure 14). It is evident that Topic 2 talks about furlough and essential workers, Topic 4 about Brexit, and Topic 9 about different tiers of responses to the pandemic.

Figure 14: Visualisation of word frequency in the topic (red) to word frequency in the corpus (grey) for Topic 2.

Try it yourself: move the slider left and right. Which setting gives the best set of words to define the topic?

[Figure0.9]

In this way, we can inspect all the topics. For an easier interpretation of the results, we can replace the generic topic names (i.e., Topic 1) with meaningful labels (i.e., T1: UK & nation). To do this, we connect **Topic Modelling** with **Select Columns** and the latter with **Edit Domain.**

First, let's open **Select Columns**, where we see the entire variable list, including word frequencies, which we created with **Bag of Words** for topic modelling. These variables are no longer needed, so we remove them by selecting all the variables using Ctrl+A in Windows or Cmd+A in MacOS in the *Features* section and drag-and-drop them to the left side, where all the variables we would like to ignore are placed. We enter *Topic* in the filter on the left side, which lists only variables containing the name *Topic* (i.e., Topic 1, Topic 2). Next, we select them and transfer them back to the right side, where all the variables we would like to include in the analysis are placed (Figure 15).

Figure 15: The list of variables in Select Columns.

Then we open **Edit Domain,** in which we will name the topics. From the list on the left, we select the first topic and set its name in the *Name* field on the right, i.e., *T1: UK & nation* (Figure 16). Naming the topics can help interpret the visualizations, which we will use later to explore the topics.

Figure 16: Renaming of topics in the Edit Domain widget.

Once the topics are named, we get a list of topics MPs debated between January 2019 and December 2020. Unsurprisingly, we find some epidemic-specific topics on the list, such as *virus and politics* and *vaccination*. Others, such as *UK & nation* and *media freedom*, might be more difficult to name if we are unfamiliar with the concurrent events. Roughly, the topics cover the ministry areas, such as security, trade, economy, higher education, transport, and crime. At the same time, looking at the list of missing topics, which one would generally expect to see covered, is telling (i.e., health and social care). The fact that certain topics are missing from the list does not mean the MPs did not debate them, but it does show they were not talked about as much, or that they were debated only in combination with other topics (i.e., the pandemic).

Such a list of topics enables a quick overview of the themes that characterised parliamentary debates at a given time. However, these results do not reveal additional information, such as which topic was debated the most, how the topics are related to one another and what the topical differences between different periods are. To answer these questions, we have to analyse the results further, which we will do in the following chapters. However, we will inspect how the topics are distributed in the corpus before doing so. In this way, we can better understand the context of speeches and, if necessary, adjust the names of the topics. At the same time, it is a great way to retrieve the speeches where a certain topic is prevalent.

Try it yourself: name all the topics with a suitable label. Some topics will be harder to define. You can use **Corpus Viewer** to find a word from the topic and explore its context.

### **Distribution of topics in a corpus**

[Workflow7]

Due to the nature of topic modelling (see Chapter 4), the speeches are characterised by more than a single topic, but topics will have different frequencies in different documents. Topic frequency or the likelihood of the topic in the speech is given between 0 and 1, where 1 means the topic fully characterizes the speech and 0 means the topic does not characterize it. Since we deal with values on the same scale and want to compare topic frequency, the most suitable visualization is the heat map. Connect the **Heat Map** widget with **Edit Domain.**

Colour represents the value in the visualisation: high values are displayed in yellow and white (or any other colour on the right side of the colour scale). In contrast, low values are displayed in blue (or any other colour on the left side of the scale). Each column represents a topic, and each row a speech. In the visualization, the speeches are displayed in the same order they were read initially, making the diagram quite difficult to interpret. A couple of settings can fix this.

First, we will join speeches with similar topic distributions. We are dealing with many speeches (130,453), so the visualization is extremely tall. We can represent very similar speeches with a single line and make the visualisation more compact. To do this, use *Merge by k-means*, which uses the k-means method to join similar speeches. The default value is 50, but we will increase it to 500 because we have many speeches and do not want to lose too many details.

Visualization is now more organised, but it would be even more informative if similar rows lay close to each other. Note that each row is no longer a single speech but a group of similar speeches. Rows can be further organised with another clustering technique, which we set in the *Clustering* section. Set the *Rows* option to *Clustering (opt. ordering)*.

Figure 17: A heat map of topic frequency. The selected branch of the dendrogram contains topics with a highly expressed T8: constituency-related issues topic.

We see a much nicer diagram (Figure 17) with a dendrogram on the left side. A dendrogram is a tree-like structure of speech similarity, which shows connections between groups and enables an easy speech selection. In the previous chapter, we learned that the interpretation of certain topics is not quite clear, for example, *constituency-related issues*. The diagram enables selecting speeches for a highly expressed topic, which can be inspected further.

We have selected *T8: constituency-related issues* for further observation. Speeches are selected by clicking on a branch in the dendrogram, where the topic is most expressed (yellow or green colour). Clicking will send the selected subgroup to the output of the widget. Now add **Corpus Viewer**to **Heat Map** to read a couple of speeches.

The speeches deal with various topics concerning MPs’ constituents, from access to health care to discrimination against minorities. In this way, we have clarified the topic label and selected a subset of speeches on a given topic, which we can use for downstream analyses.

When interpreting the results of topic modelling, we need to consider that a speech is not characterised by a single topic but a mixture of them, which we can see for the topic *T4: trade*. If we select the speeches and give them a careful read, we will see that they touch upon various topics, including the new taxation of foreign goods after Brexit. The topic is also evident from the diagram, where certain topics with high values of *T4* also have high values of *T3: legislative*. These two topics thus overlap in certain points as already evident from the heat map. The visualisation is great for investigating topic overlap, which is crucial if we are interested in selecting speeches on a given topic for further analysis.

Try it yourself: select speeches about the virus and politics and explore them.

## Topic map

[Workflow8]

Now we know the distribution of topics by speech. We have learned that several topics characterise a speech, so we would like to know how the topics are related to one another. Besides, we would like to know which topics are the most prevalent in our corpus. To answer these two questions, we will use a topic map, where the topics will be positioned based on their similarity to one another and marked by their frequency.

We will construct the map with **MDS**, which is short for *multidimensional scaling*. The visualisation tries to find a projection in a 2-dimensional space such that related topics lie close to one another and those unrelated are far apart. MDS computes topic relatedness based on the importance of words in the topic. High relatedness means a very similar word distribution, where some words can be even shared among the topics.

We set the connection between **Topic Modelling** and **MDS** by connecting *All Topics* to *Data*. In the beginning, we will see only grey points in space. A point represents a topic. For easier interpretation, we will label the points. We do this by setting *Labels* to *Topics*. Each point will be marked with a topic name – not the one we gave them in **Edit Domain**, but the original names. Thus, we need to use a manually created list of topics for interpretation (Table 1).

Table 1: A list of topics with the original and assigned label.

|  |  |
| --- | --- |
| **Topic name** | **Assigned label** |
| Topic 1 | T1: UK & nation |
| Topic 2 | T2: security |
| Topic 3 | T3: legislative |
| Topic 4 | T4: trade |
| Topic 5 | T5: procedural |
| Topic 6 | T6: business & industry |
| Topic 7 | T7: virus and politics |
| Topic 8 | T8: constituency-related issues |
| Topic 9 | T9: pandemic and energy transition |
| Topic 10 | T10: economy |
| Topic 11 | T11: sports |
| Topic 12 | T12: child well-being |
| Topic 13 | T13: climate change |
| Topic 14 | T14: vaccination |
| Topic 15 | T15: higher education |
| Topic 16 | T16: media freedom |
| Topic 17 | T17: schools in pandemic |
| Topic 18 | T18: transport |
| Topic 19 | T19: crime |
| Topic 20 | T20: housing |

We will also set the size of the points to match the frequency of the topic (which is the sum of topic probability in all the speeches, weighted by the length of the speech). Set the *Size* option to *Marginal Topic Probability*. To make things even clearer, set the *Color* to *Marginal Topic Probability*.

Figure 18: Displaying topic relatedness in MDS.

Thematic map displays topic relatedness with the position of the points, while the size and the colour show how frequent the topic is (Figure 18). When topics are related, but the points lie far apart due to the limitations of the 2-dimensional display, the relatedness is marked with a line between the points. The most frequent topics are *T3: legislative* and *T8: constituency-related issues*, while narrower topics such as *T11: sports*, *T13: climate change* and *T14: vaccination* are the least frequent.

The high frequency of Topic 3 (*legislative*) is unsurprising, as the parliament is the main legislative body of the state. It is placed close to Topic 5 (*procedural*), which means legislative speeches contain a lot of procedural words. It is also close to Topic 7 (*virus and politics*), which shows that the government had to adopt certain legislative measures to combat the pandemic. Topic 7 is also close to Topic 8 (*constituency-related issues*), which could mean there was a debate on translating pandemic measures into a local environment.

[Figure0.10]

In short, topics that lie close to each other or are connected with a line are related. However, sometimes it is not easy to understand how the two topics are related — for example, Topic 1 (*UK & nation*) and Topic 16 (*media freedom*). To better understand this connection, we can review the speeches with a high frequency of both topics.

First, we will create a subcorpus with a strong presence of topics *T1: UK & nation* and *T16: media freedom*. We do this with the **Select Rows** widget, which we will connect to **Topic Modelling**. We have to set two conditions in the **Select Rows** widget: *Topic 1 is greater than 0.4,* and *Topic 16 is greater than 0.4* (Figure 19). We will select the speeches where the two topics are present with over 40-percent probability.[[16]](#footnote-16)

Figure 19: Setting the threshold in the Select Rows widget.

**Select Rows** is then connected to **Corpus Viewer**, which displays the selected 26 speeches at the intersection of nation and media (Figure 20). It is clear from the speeches that they refer to internal matters of organisation of the two Houses of the Parliament and the specific roles given to the MPs.

Figure 20: Overview of the selected speeches in the Corpus Viewer widget.

Try it yourself: Explore the Topics 3 and 5 in the same way.

## Topics before and during the pandemic

[Workflow9]

We identified the topics which stand out the most in our corpus, but now we would like to investigate which topics are the most characteristic of the pre-pandemic and the pandemic periods. The differences between the two periods (already labelled in the data with *Reference* and *COVID*) can be explored with **Box Plot**. The visualisation, also known as a box-and-whisker plot, shows the distribution of the variable and enables an easy comparison of topic probability by [categorical variables](https://en.wikipedia.org/wiki/Categorical_variable) (i.e., gender, date, party).

Connect **Box Plot** to **Edit Domain**, which will keep our assigned topic labels. As we wish to compare two periods, select the *Subcorpus* variable in the lower-left section of the widget. In the upper left section, select *T1: UK & nation*. On the right side, we will see two box plots, the upper for the pandemic period (*COVID*) and the lower for the pre-pandemic period (*Reference*) (Figure 21). The visualisation shows that the debates on the UK nation were more frequent before the pandemic than during the pandemic. At the same time, the test result below the plot shows that the difference is statistically significant, as its p-value is below 0.05. We can conclude that historical topics were more frequent before the pandemic based on this information.

Figure 21: Box plot for topic T1: UK & nation before and during the pandemic.

[Figure0.11]

We could inspect the distribution for every topic separately, but this would be quite laborious. Since we are not focusing on a single topic but would like to understand which topics show the most difference between the two periods, we will use the *Order by relevance to subgroups* option in the *Variable* section. This option will sort the variables based on the results of the statistical test. At the top, we will see those variables that show the greatest difference for the selected categories, which we defined in the *Subgroups* section (the *Subcorpus* variable).

At the top of the list, we will see the variables such as *category*, *From, To,* and *Term.* These variables show the greatest differences between *Reference* and *COVID* subcorpora. The difference is unsurprising since these variables are time-related, which was also a criterion for forming the two subcorpora (see Chapter 5.3). We are more interested in the variables following these four.

At the top are the topics *T7: virus and politics* (Figure 22) and *T17: schools in a pandemic*. Once we select one of them, the visualisation on the right shows that the MPs talked more about the virus and politics in the pre-pandemic period and more about the school in the, unsurprisingly, pandemic period. The Student's t-test (57.184, p<0.05) below the plot shows that the difference between the two periods is significant and that the topic is more prominent in the given period. The result for Topic 17 is unsurprising, but the result for Topic 7 is a little strange. One would expect the speeches on the virus to be more prominent during the pandemic. Looking at the speeches in **Corpus Viewer**, this is indeed the case – all the speeches containing the word »virus« come from the *COVID* period. We see that these speeches talk about the virus in the context of the EU response to the pandemic (making it likely that the virus under consideration is indeed the coronavirus and not some other pathogen). However, the significance of the pre-pandemic period in this case is due to a large portion of speeches belonging to the pre-pandemic period. Skimming through the speeches, we can see that the topic consists of speeches covering a range of EU-related issues linked to Brexit which was in the spotlight before the COVID outbreak.

Figure 22: Box plot for topic T7: virus and politics.

The results show the usefulness of topic modelling, but they also point out how vital it is to understand the corpus and explore the speeches with close reading.

The topics can be further explored with **Select Rows** and **Corpus Viewer**. Connect **Select Rows** to **Edit Domain.** In **Select Rows**, set *T7: virus and politics* *is greater than 0.98*, by which we will output only the speeches with more than 98% likelihood of T7. The selected speeches can be inspected in **Corpus Viewer** or with **Word Cloud.**

**Corpus Viewer** enables us to explore the context of a given word. Let’s say we are interested in learning more about the lemma »vote«, which is characteristic of Topic 7 (see Chapter 6.1.2). We have already selected speeches with a high frequency of Topic 7, which outputs 62 speeches. We would like to see which out of those contain the lemma »vote«. We can enter the lemma in the filter at the top of the widget and press *Enter*. The widget will display the speeches where the lemma »vote« appears – there are 36 such speeches. Indeed, we can see that the speeches refer to the relationship with the EU (Figure 23).

Figure 23: Corpus Viewer with speeches containing the word "vote".

An alternative option is to display the most frequent lemmas for the topic in a **Word Cloud**. Word cloud would give an in-depth look into the concepts discussed in this topic (Figure 24).

Figure 24: Word cloud of the most frequent words in Topic 7.

The speeches mostly refer to *deals, voting, government, people,* and *extension*. Considering that the lemma *referendum* is also quite prominent, these speeches probably refer to Brexit. While almost exclusively present in this topic, it seems like the word virus is generally quite infrequent in these speeches. The results indicate *virus* has to be interpreted in the context of other words characterising Topic 7. It is vital to compare the frequency of characteristic words for the topic and the frequency of words in the corpus (like we did in Chapter 6.1.3) to ensure accurate topic interpretation.

Try it yourself: In the same way we compared the subcorpora *Reference* and *COVID*, compare the distribution of topics in opposition and coalition speeches.

# Conclusion

Parliamentary corpora, which contain the records of parliamentary debates, provide an important source for researching politics and its impact on society. These corpora usually hold rich metadata on the speakers and speeches and include multi-layered linguistic annotations that enable researchers to explore various research questions. Due to the size of such corpora, text mining methods, such as topic modelling, which enables topic extraction, prove to be extremely useful in researching their content. In this tutorial, we present the LDA method, one of the most popular methods for topic modelling. The analysis based on this method is performed in Orange, an open-source software for visual programming, which allows advanced data processing without code. The analysis in this tutorial was made on the ParlaMint-GB corpus that contains British parliamentary records.

Figure 25: The final workflow.

The tutorial is designed for self-study and breaks down the analytical process into simple steps illustrated by numerous screenshots for easy progress (Figure 25). It also includes instructions for additional individual work, which helps consolidate the acquired knowledge and encourages users to use the software independently. While special emphasis is given to the presentation of the key characteristics of the analysed data, the tutorial also describes the specificities and limitations of the method used, thus promoting a critical approach to data analysis.

Although the tutorial bases its analysis on the British parliamentary data, it is easy to extend the research to other text genres and other languages. Since the presented method is not language-specific, it can be used on any of the ParlaMint corpora. The value of the tutorial for students and researchers in the social sciences and humanities, therefore, reaches far beyond the specific research problems explored in this tutorial.

**Acknowledgements**

The work described in this paper was funded by the Slovenian Research Agency research programme P6-0436: Digital Humanities: resources, tools, and methods (2022- 2027), the Social Sciences & Humanities Open Cloud (SSHOC) project (<https://www.sshopencloud.eu/>), the CLARIN ERIC ParlaMint project (<https://www.clarin.eu/parlamint>) and the DARIAH-SI research infrastructure. We would also like to thank Çağrı Çöltekin, Marta Kołczyńska, Jiřina Popelikova, Mladen Zobec and Jure Skubic for their thorough reviews and thoughtful comments.

# References

Abercrombie, G., & Batista-Navarro, R. (2020). Sentiment and position-taking analysis of parliamentary debates: A systematic literature review. *Journal of Computational Social Science*, *3*(1), 245–270.

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, *3*, 42.

Allen, C., & Murdock, J. (2020). *LDA topic modeling: Contexts for the history & philosophy of science*.

Arun, R., Suresh, V., Veni Madhavan, C., & Murthy, N. (2010). *On finding the natural number of topics with latent dirichlet allocation: Some observations*. 391–402.

Bayley, P. (2004). *Cross-cultural perspectives on parliamentary discourse* (Vol. 10). John Benjamins Publishing.

Bergmann, H., Geese, L., Koss, C., & Schwemmer, C. (2018). *Using legislative speech to unveil conflict between coalition parties* [Preprint]. SocArXiv. <https://doi.org/10.31235/osf.io/pgnwa>.

Blätte, A., Gehlhar, S., & Leonhardt, C. (2020). *The Europeanization of Parliamentary Debates on Migration in Austria, France, Germany, and the Netherlands*. 66–74.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Chizhik, A. V., & Sergeyev, D. A. (2021). *Exploring the Parliamentary Discourse of the Russian Federation Using Topic Modeling Approach*. 403–416.

Curran, B., Higham, K., Ortiz, E., & Vasques Filho, D. (2018). Look who’s talking: Two-mode networks as representations of a topic model of New Zealand parliamentary speeches. *PLOS ONE*, *13*(6), e0199072. <https://doi.org/10.1371/journal.pone.0199072>.

de Campos, L. M., Fernandez-Luna, J. M., Huete, J. F., & Redondo-Expósito, L. (2021). LDA-based term profiles for expert finding in a political setting. *Journal of Intelligent Information Systems*, *56*(3), 529–559.

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., & Starič, A. (2013). Orange: Data mining toolbox in Python. *The Journal of Machine Learning Research*, *14*(1), 2349–2353.

Erjavec, T. et al. (2021). *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1* (v2.1) [Computer software]. Slovenian language resource repository CLARIN.SI. <http://hdl.handle.net/11356/1431>

Erjavec, T., Ogrodniczuk, M., Osenova, P., & et al. (2022). The ParlaMint corpora of parliamentary proceedings. *Lang Resources & Evaluation*. <https://doi.org/10.1007/s10579-021-09574-0>.

Erjavec, T., & Pancur, A. (2019). Parla-CLARIN: TEI guidelines for corpora of parliamentary proceedings. *Book of Abstracts of the TEI2019: What Is Text, Really*.

Fišer, D., & Pahor de Maiti, K. (2021). »Prvič, sem političarka in ne politik, drugič pa…«. *Contributions to Contemporary History*, *61*(1). <https://doi.org/10.51663/pnz.61.1.07>.

Gkoumas, D., Pontiki, M., Papanikolaou, K., & Papageorgiou, H. (2018). *Exploring the Political Agenda of the Greek Parliament Plenary Sessions* (D. Fišer, M. Eskevich, & F. de Jong, Eds.).

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

Høyland, B., & Søyland, M. G. (2019). Electoral reform and parliamentary debates. *Legislative Studies Quarterly*, *44*(4), 593–615.

Ilie, C. (2010). *European parliaments under scrutiny: Discourse strategies and interaction practices* (Vol. 38). John Benjamins Publishing.

Jacobs, T., & Tschötschel, R. (2019). Topic models meet discourse analysis: A quantitative tool for a qualitative approach. *International Journal of Social Research Methodology*, *22*(5), 469–485.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*.

Kilroy, D. (2021). All the king’s men? A demographic study of opinion in the first English Parliament of James I, 1604–10. *Parliaments, Estates and Representation*, *41*(1), 1–23.

Martin, F., & Johnson, M. (2015). *More efficient topic modelling through a noun only approach*. 111–115.

Meeks, E., & Weingart, S. B. (2012). The digital humanities contribution to topic modeling. *Journal of Digital Humanities*, *2*(1), 1–6.

Mollin, S. (2007). The Hansard hazard: Gauging the accuracy of British parliamentary transcripts. *Corpora*, *2*(2), 187–210.

Morstatter, F., Shao, Y., Galstyan, A., & Karunasekera, S. (2018). *From alt-right to alt-rechts: Twitter analysis of the 2017 German federal election*. 621–628.

Müller-Hansen, F., Callaghan, M. W., Lee, Y. T., Leipprand, A., Flachsland, C., & Minx, J. C. (2021). Who cares about coal? Analyzing 70 years of German parliamentary debates on coal with dynamic topic modeling. *Energy Research & Social Science*, *72*, 101869.

Norton, P. (2002). *Parliaments and citizens in Western Europe* (Vol. 3). Psychology Press.

Pančur, A., & Šorn, M. (2016). Digitalni pristop k parlamentarni zgodovini: Uporaba gradiva Državnega zbora v digitalni humanistiki. *Četrt stoletja Republike Slovenije - izzivi, dileme, pričakovanja*, 115–126.

Petukhova, V., Malchanau, A., & Bunt, H. (2015). *Modelling argumentation in parliamentary debates* (M. Baldoni & et al., Eds.). Springer.

Piersma, H., Tames, I., Buitinck, L., Van Doornik, J., & Marx, M. (2014). War in parliament: What a digital approach can add to the study of parliamentary history. *Digital Humanities Quarterly*, *8*(1).

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

Proksch, S.-O., & Slapin, J. B. (2010). Position taking in European Parliament speeches. *British Journal of Political Science*, *40*(3), 587–611.

Rheault, L., Beelen, K., Cochrane, C., & Hirst, G. (2016). Measuring emotion in parliamentary debates with automated textual analysis. *PloS One*, *11*(12), e0168843.

Rheault, L., & Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, *28*(1), 112–133.

Rosa, A. B., Gudowsky, N., & Repo, P. (2021). Sensemaking and lens-shaping: Identifying citizen contributions to foresight through comparative topic modelling. *Futures*, *129*, 102733.

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (n.d.). *Supervised Sentiment Analysis of Parliamentary Speeches and News Reports*.

Schmidt, B. M. (2012). Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities*, *2*(1), 49–65.

Schuler, P. (2020). Position taking or position ducking? A theory of public debate in single-party legislatures. *Comparative Political Studies*, *53*(9), 1493–1524.

Serrano, J. C. M., Shahrezaye, M., Papakyriakopoulos, O., & Hegelich, S. (2019). *The rise of Germany’s AfD: A social media analysis*. 214–223.

Shadrova, A. (2021). Topic models do not model topics: Epistemological remarks and steps towards best practices. *Journal of Data Mining & Digital Humanities*, *2021*.

Sieberer, U., Müller, W. C., & Heller, M. I. (2011). Reforming the rules of the parliamentary game: Measuring and explaining changes in parliamentary rules in Austria, Germany, and Switzerland, 1945–2010. *West European Politics*, *34*(5), 948–975.

Sievert, C., & Shirley, K. (2014). *LDAvis: A method for visualizing and interpreting topics*. 63–70.

Smith, N., & Graham, T. (2019). Mapping the anti-vaccination movement on Facebook. *Information, Communication & Society*, *22*(9), 1310–1327.

Truan, N., & Romary, L. (2021). Building, Encoding, and Annotating a Corpus of Parliamentary Debates in XML-TEI: A Cross-Linguistic Account. *Journal of the Text Encoding Initiative*.

van der Zwaan, J. M., Marx, M., & Kamps, J. (2016). *Validating Cross-Perspective Topic Modeling for Extracting Political Parties’ Positions from Parliamentary Proceedings.* 28–36.

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582.

Wiedemann, G. (2016). *Text mining for qualitative data analysis in the social sciences* (Vol. 1). Springer.

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). *A heuristic approach to determine an appropriate number of topics in topic modeling*. *16*(13), 1–10.

1. Building blocks performing different steps of the analysis. [↑](#footnote-ref-1)
2. The key and binding rules of procedure governing parliamentary organisation and work as well as the MPs’ rights and obligations are codified in the rules of procedure of individual parliaments, i.e., [the Rules of Procedure of Slovenian National Assembly](http://www.pisrs.si/Pis.web/pregledPredpisa?id=POSL34), [the Rules of Procedure of the UK Parliament](https://erskinemay.parliament.uk/), [the Rules of Procedure of the German Bundestag](https://www.btg-bestellservice.de/pdf/80060000.pdf), etc. [↑](#footnote-ref-2)
3. Time-based topic analysis can be done with plain LDA, but a separate topic model must be built for each period; the topics must be interpreted and then compared between time periods. Such an approach requires a lot more manual and subjective work, which can negatively affect the results. Another option for temporal topic analysis is *dynamic non-negative matrix factorisation* (*dynamic NMF*) (see Gkoumas et al. 2018), which considers the time periods indirectly (Müller-Hansen et al. 2021). [↑](#footnote-ref-3)
4. If you have used Orange before, please clear widget settings under Options 🡪 Reset widget settings. This will enable you to repeat the analysis as described in the tutorial. [↑](#footnote-ref-4)
5. The data shown come with the widget. [↑](#footnote-ref-5)
6. It has been established that the delimited periods are comparable in terms of the quantity of speeches and sessions that they encompass. [↑](#footnote-ref-6)
7. The folder will contain linguistically annotated parliamentary speeches in the CoNLL-U format and metadata in the TSV format. Keep all the files in the folder for the import. [↑](#footnote-ref-7)
8. This is the file format you get from the CLARIN.SI repository if you search for the ParlaMint data with added linguistic annotations. The provided link will lead you to a selection of files that are relevant for this tutorial (i.e., only the data from 2019 and 2020). The Parla-Mint-GB corpus, however, includes much more data. If, at a later time, you would like to analyse the entire corpus or a different time period, you can get the entire corpus from the [ParlaMint-GB.ana.tgz](http://hdl.handle.net/11356/1431) folder in the CLARIN.SI repository and make a selection of the data according to your needs. [↑](#footnote-ref-8)
9. The token number is the number of every word, number, and punctuation mark in the corpus, while the type is the number of unique tokens in the corpus. [↑](#footnote-ref-9)
10. When specifying a query, you can use [regular expressions](https://www.sketchengine.eu/guide/regular-expressions/), which enable searching for specific words or word forms. The query *epidemic\** will, for example, list the word *epidemic* in all its forms. [↑](#footnote-ref-10)
11. Certain metadata might be missing if the original parliamentary records are imperfect. Linguistic annotations have been added automatically. This means that you should allow for some annotation errors, even though the tools used are pretty accurate: 98–99% for lemmatisation, 94–97% for morphological tagging and 87–94% for syntactic tagging (Erjavec et al., 2022). [↑](#footnote-ref-11)
12. While the entire ParlaMint corpus family has the same set of metadata (see Chapter 3.3), not every corpus includes all metadata. [↑](#footnote-ref-12)
13. If you wish to disable the data updating after every parameter change, uncheck the *Apply Automatically* option in the bottom left corner – then click *Apply* once you have done all the changes. [↑](#footnote-ref-13)
14. Set *max* by removing/deleting the upper threshold. [↑](#footnote-ref-14)
15. It is possible to get a different result with LDA than seen in the tutorial. LDA is a generative model, which initiates randomly. You should always be able to get the same results on your computer, but the results can differ between different versions of Orange and different operating systems. [↑](#footnote-ref-15)
16. The threshold of 40 percent is selected because it is the lowest value which returns at least a couple of documents. One should be aware that the threshold is low, which means the speeches do not have a very high likelihood of the two topics. The value can be adjusted freely. [↑](#footnote-ref-16)