DEPARTMENT OF PHILOSOPHY,
LINGUISTICS AND THEORY OF SCIENCE

TOPIC MODELING FOR ANALYSIS OF PUBLIC DISCOURSE

-Enriching topic modeling with linguistic information to analyze Swedish housing policies

Anna Lindahl

Master's Thesis:	30 credits
Programme:	Master's Programme in Language Technology
Level:	Advanced level
Semester and year:	Autumn, 2017
Supervisor	Lars Borin, Love Börjeson
Examiner	Dimitrios Kokkinakis
Keywords	topic modeling, public discourse, housing policies, LDA, semantic coherence measures, part of speech

Abstract

This work investigates how the method of topic modeling can be applied to investigate the public discourse of Swedish housing policies. The data used to represent this discourse is both from the Swedish parliament, the Riksdag, and Swedish newstexts. The lack of housing and current housing crisis in Sweden makes this a relevant area to study. Topic modeling is an unsupervised probabilistic method for finding topics in large collections of data. This is a popular method for examining public discourse, however there is a lack of including linguistic information in the preprocessing steps of it. Therefore, this work also investigates what effect linguistically informed preprocessing has on topic modeling.

Three types of linguistic information are selected and investigated. These are part of speech, dependency relations and lemmatization. Based on these, filters are created for the data. The filters are applied to a test set (a subset of the original data), and a topic model is trained on each filtered version of this test set. The resulting topics from each model are evaluated by both humans and the computational methods perplexity and semantic coherence, and the results from the respective evaluation methods are compared. The semantic coherence named *cv* is found to have a higher correlation with human ratings than the *npm*i coherence. Perplexity is found to not correlate well with human ratings.

Filtering the data based on part of speech is found to most improve the topic quality. Non-lemmatized topics are found to be rated higher than lemmatized topics. Topics from the filters based on dependency relations are found to have low ratings.

Based on the human ratings, an optimum model for respective data set is chosen. The selected topic models are applied to the data, and the results are used for to exemplify how one can use them for analysis. Topic modeling is found to be a suitable method for the intended analysis.

Acknowledgements

This work has been supported in part by a framework grant for the project *Towards a knowledge-based culturomics*[1], awarded by the Swedish Research Council (contract 2012-5738).

This work has also been carried out with the support from Hyresgästföreningen[2], which has provided part of the data used here.

I would like to thank both of my supervisors for all the input and help, the participants who helped with the evaluation and I would also like to thank everyone who has ever shared their knowledge by answering a question on the internet. I also want to thank Colle for support and great discussions.

For the interactive plots please contact the author at *annanlindahl@gmail.com*

[1] https://spraakbanken.gu.se/eng/culturomics
[2] https://www.hyresgastforeningen.se/

Contents

1 Introduction

Arguably everyone in Sweden has an opinion about the housing market and policies. It has been a source of debate for a long time, and considering there has been a housing crisis ongoing since the 1990's, this is not surprising. Lack of housing is still becoming more widespread, and with only a small rise in newly built houses in 2015–2016 (Höjer, 2017) one can wonder what this debate has contributed. With this in mind, the starting point of the present work was a call from the Swedish union of tenants (Hyresgästföreningen) asking for language technology methods for analyzing the housing policies discourse.

In recent years we have seen the emergence of many new techniques for analyzing the constantly increasing flow of information and data. Methods for automatically extracting topics, topic modeling, machine learning for sentiment analysis and information extraction are a few of them.

In the field of humanities and social sciences the use of computational methods has been argued for by many. Sometimes referred to as "Digital Humanities", the importance of tools for investigation of both digital and printed texts is undeniable. However, as Viklund & Borin (2016) argue, these techniques still need refinement and development to be both accessible and more useful. Often, the linguistic information is disregarded, and there is a need to explore what incorporating this can do for the field. This issue is also raised by Tahmasebi et al. (2015), where the concept of culturomics is discussed, and the need for good linguistic preprocessing to make this a successful field. This is especially important with respect to language specificity. Most tools are developed and tested on English, and this doesn't always translate well to other languages.

Therefore, the aim of the present work is twofold. The first is to exemplify and explore how one can apply language technology to investigate the public discourse of the Swedish housing policies. More specifically, the method of topic modeling is chosen to explore the discourse. The second is to investigate how one can adapt and enrich this method with linguistic information, especially with Swedish in mind.

To explore the discourse, the questions asked are:

- How have people been talking about housing policies?
- How has this varied over time?

To represent this discourse, data from both the Swedish parliament and newstexts is used. These areas are two important parts of the public discourse of housing policies. The documents and records from the Swedish parliament is a natural place to look for information on housing policies, and are available over a long span of time. The newstexts were selected on the basis of them containing one or more words from a list of keywords relating to Swedish housing policies. This was done in order to find the newstexts concerning housing policies.

To answer the above questions about the public discourse, the method of topic modeling is used, as mentioned above. Topic modeling has proved a successful method in a wide range of areas for finding structure and topics in large quantities of text. For example, Hall et al. (2008) use it to study ideas within the computational semantics field over time, DiMaggio et al. (2013) investigate the news coverage of U.S. arts funding and Jacobi et al. (2016) use it for analysis of journalistic text.

However, as McFarland et al. (2013), in their study of language differentiation in academe, point out: "topic models require careful thought and revision if they are to be successfully applied to social science research questions." Mohr & Bogdanov (2013) also discuss the need for both knowledge of the method

and the field in question, in order for the topic modeling to have a meaningful outcome. They also mentioned the need for good linguistic preprocessing.

This brings us to the second aim and research question:

- How is topic modeling affected by different kinds of preprocessing steps, based on linguistic information?

Here, preprocessing refers to the steps before the topic modeling, where one processes the data before training a model on it. Preprocessing includes both formatting of the data, such as removing punctuation, but it can also include removing all words of a certain part of speech. The latter is thus based on linguistic information about the data.

Preprocessing is necessary when doing topic modeling, but there is no standard recipe for it and many studies differ in their use of it. The effect of different preprocessing choices has not been studied systematically. There is also a lack of using linguistic information in the preprocessing. Furthermore, most of the topic modeling studies employ models which are trained on English. Swedish, which has a richer morphology than English, can be assumed to be more affected by lemmatizing, for example.

There are many kinds of linguistic information to incorporate and this work cannot examine them all. The aim here is to examine the effect of three selected types of linguistic information and investigate if there is a general effect to be expected from them. The selected linguistic information is part of speech, dependency relations and lemmatization. This information is used to create different preprocessing filters. These filters are applied to a test set of the data, coming from the parliaments documents, which results in different versions of the original data. A topic model is trained on each version of this data.

The resulting models are evaluated, to find how the filters affect the interpretability of the model and which filters result in good models. The evaluation is done by human participants. Their evaluation is compared to computational methods, to ensure the best model is chosen. The highest rated model from the test set is then used to explore the rest of the data from the parliament. To explore the newstexts, the filters from the five highest rated models from the parliaments documents are reused on a test set from the newstexts, resulting in five versions of the test set. Five new models are trained and evaluated on these test sets. The highest rated model is used to explore the rest of the newstexts.

In both data sets, with the help of the selected model for each set, documents concerning housing policies are identified. The occurrence of these over time answer the question of how the housing policies has varied. To answer the question of how people have been talking about housing policies, the cooccurring topics with the housing topics are identified.

The rest of the thesis is structured as follows. In section 2, previous research and topic modeling are presented. In section 3, the data is presented. In section 4 the method is presented and in section 5 the results. In section 6 these results are discussed and in section 7 conclusion and future work is presented.

2 Background

This section is split in four parts. The first gives a short introduction to the housing crisis in Sweden and the second gives an explanation of the linguistic information used. The third explains how topic modeling works and how it can be used. The fourth gives an overview of the use of topic modeling public discourse.

2.1 The housing crisis

According to the latest survey on the housing situation in Sweden, made by Boverket (2017),(the national board of housing, building and planning), 255 out of 290 municipalities report a lack of housing. This is an increase of 72 municipalities in 2 years. At the same time, the population in Sweden is steadily growing (SCB, 2017b), with a prognosis of a historically unusually high growth in the coming years (SCB, 2017a).

More houses are being built in recent years, but there has been a shortage for so long that it will take years to fill the need. Also, it is argued that even though houses are built, they are of the wrong kind. They are too expensive and too few. The need for housing is the largest among socially vulnerable groups, which are also the ones that will have the most difficulties finding the financial strength to buy housing (Höjer, 2017).

The reason and solutions for the crisis are naturally debated, but what is undeniable is the existence of it. The aim here is to investigate the debate, not participate in the debate itself.

2.2 Linguistic information

The term *linguistic information* is very broad and can encompass many different things. As mentioned in the introduction, here the focus lies on three parts. These are:

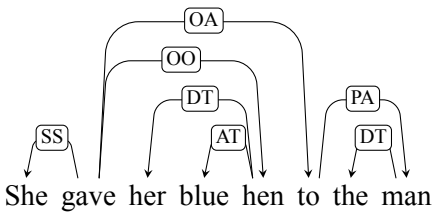
1. Part of Speech
2. Lemma
3. Dependency relations

Part of speech (POS) is the most straight-forward of these, it refers to a word's part of speech category. This can be for example noun, verb or adjective. Function words such as prepositions and pronouns are other examples. Traditionally, nouns and verbs are considered to carry more meaning than function verbs. This is important to consider when selecting words for the topic modeling.

A lemma is a word's dictionary or base form. For example, *cat* is the lemma of *cats* and *walk* is the lemma of *walking*. The process of changing words into lemmas is called lemmatization. To reduce all words in a corpus to lemma can help when ones uses words statistics, because the number of words are reduced, unless one's interest lies in the specific word forms. Lemmatization is presumably more effective in languages with rich morphology.

Dependency relations refer to the syntactic relations between words in a sentence, as specified by a dependency grammar. These relations are called dependencies and they are binary asymmetrical relations. A few of the possible grammatical relations can be seen in table 1 which also shows a parsed sentence. The type of grammatical relations can vary depending on which dependency grammar is used. Advantages of dependency grammars includes a strong predictive power and easy handling of different word

orders. (Jurafsky & Martin, 2009). Here, dependency relations are used as a selection criteria for the data.



SS	Other subject
OO	Direct object
OA	Object Adverbial
DT	Determiner
AT	Nominal (adjectival) pre-modifier
PA	Complement of preposition

Table 1: An example of a dependency parse.

2.3 Topic modeling

Topic modeling is the general name for a group of usually probabilistic models that aim to automatically discover topics in a data collection (Blei, 2012). Often this data is collections of text documents, but there are also examples of topic modeling being used on other forms of discrete data, for example image processing (Blei et al., 2010). However, here only applications in natural language are discussed, that is, topic modeling used for exploring collections of text.

A topic model builds on the assumption that there is a hidden set of topics in the data it models. By observing a collection of documents, it infers a number of topics. To do this, the only information the model needs is the distribution of words in the documents, and for most models, a number of topics to be found. Each topic is then represented as a probability distribution over words.

Table 2 shows an example of topics generated by a topic model trained on Swedish data, with translation. These words have the highest probability in these topics. By observing these words, the topics in the figure can be assumed to be about education and energy, respectively. These labels needs to be provided by a human annotator, the topic model has no real world knowledge of what is a coherent topic. The structure that the topic model infers matches well with the structure of language and this results in topics which we as humans infers as a coherent topic (Blei, 2012).

Topic	Translation	Topic	Translation
skola	'school'	*kärnkraft*	'nuclear energy'
elev	'student'	*energi*	'energy'
lärare	'teacher'	*vindkraft*	'wind power'
kunskap	'knowledge'	*el*	'electricity'
gymnasieskola	'high school'	*företag*	'company'
möjlighet	'opportunity'	*vattenkraft*	'water power'
undervisning	'teaching'	*energipolitik*	'energy politics'
utbildning	'education'	*elproduktion*	'electricity production'

Table 2: Examples of topics from a topic model

Still, topic modeling is a powerful tool for to bringing structure to unstructured data. Once a model has been trained it can be used on new and old documents to find which topics each document contains. This can be used for classification, prediction and information retrieval tasks, among others. Topic modeling is advantageous because there is no need to label the data beforehand, it is a so called unsupervised method. This is useful for exploration and analysis of big corpora or data sets, where one can use topic modeling to to find topics and documents of interest and thus avoid having to manually go through the data. DiMaggio et al. (2013) describe this use of topic modeling as a lens to view a corpus through.

In it simplest form, the distribution of words in documents is the only information a topic model needs, as mentioned above. The model sees the documents as bags of words. This means that the syntactic structure and other information is lost, but topic modeling manages to capture topics despite this. However, there are a number of extended and modified models, many which include more information about the data. See section 2.3.2 for examples on this.

Training a topic model is a trial and error process. A number of parameters is involved, and it is often difficult to know exactly what effect a change in a certain parameter has on the results. It would be a too big scope to evaluate all of them, and therefore most studies settle on the default parameters of the chosen toolkit or own script. This is however something to keep in mind when employing topic modeling.

Perhaps the most important parameter when employing topic modeling is the number of topics. There is no universal number of topics, it depends on the data being modelled. A more diverse set of data will have more topics to be retrieved. The number also depends on what is desired from the topics, a low number usually results in low granularity of the topics and vice versa (Jockers, 2013).

The most commonly used topic model is the probabilistic model Latent Dirichlet allocation (LDA), which is the model employed here. Other topic models include hierarchical Dirichlet process (HDP), which is similar to LDA but there is no need to specify the number of topics beforehand, and LSI, which can be seen as predecessor to LDA.

2.3.1 Latent Dirichlet Allocation

LDA is a generative probabilistic model, and by explaining how it works one will also understand the intuitions behind topic modeling. LDA was developed by Blei et al. (2003), and builds upon ideas from the earlier topic models, such as probabilistic Latent semantic indexing (pLSI), but seeks to improve efficiency and modeling capabilities.

LDA assumes there is a number of hidden (or latent, hence the name) topics in the data, giving rise to the observable data, which are the words in the provided corpus. Each document is seen as a random mixture

of topics, where each word belongs to a topic. A topic is, as mentioned above, a probability distribution over words. A word can only belong to one topic at a time, although it is possible for different occurrences of the same word to belong to different topics. As with most other topics models, the number of topics needs to be provided when one trains an LDA model.

A classic way to illustrate and explain LDA is to describe the generative story of it. This is how the LDA model assumes documents are created, and this story is used to infer the hidden topics. The idea is, in the creation of every document, a random distribution of topics is chosen for that document. This is done by randomly drawing a probability distribution from a Dirichlet distribution. The Dirichlet distribution is a probability distribution over the space of multinomial distributions, here these multinomial distributions are the topic probabilities. The reason for using a Dirichlet distribution for the topics probabilities is that the it makes the inference and approximation of parameters easier. For a more detailed explanation see (Blei et al., 2003).

To then produce a word in the document, a topic is randomly drawn from the topic distribution. From the drawn topic a word is drawn, at random. This procedure is assumed to occur for every word in all the documents.

Equation 1 shows the formal description of the generative process describe above, which is a joint probability distribution of the observed and hidden variables. β are the topics and every β_k is a distribution over words from the vocabulary from the observed corpus. θ_d is the topic distribution over document *d*, and $\theta_{d,k}$ is the proportion of topic *k* in document *d*. The topic assignments for document *d* are represented by z_d, and $z_{d,n}$ is the *n*th word in document *d*. The observed words from document *d* are represented by w_d, and w_{d_n} is then the *n*th word in document *d*.

$$p = (\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \right) \tag{1}$$

Using this equation topics can be inferred by observing their posterior distribution. This is done by calculating the conditional probability, which is defined as:

$$p = (\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})} \tag{2}$$

where the numerator is the joint distribution. The denominator represents the marginal probability of the corpus, the probability of the observed words in any possible instantiation of a topic model. This is impossible to compute, and therefore different methods for approximating it is used. One of the most commonly used methods for this is Gibbs sampling, where a Markov chain is used to construct dependent variables, with the posterior distribution used as a limiting distribution. Samples are collected from this distribution to approximate the posterior. Another commonly used method is variational Bayes, where the problem is turned into an optimization problem. This is done by specifying a number of distributions for the hidden structure and then finding the distribution that is most similar to the posterior (Blei, 2012). The implementation here uses a modified version of variational Bayes, made to handle documents in a stream, which makes handling large corpora more effective (Hoffman et al., 2010).

It should be noted that topics from an LDA model will differ slightly every time the model is trained on a data set. That is, different models trained on the same data will differ slightly. This could result in topics from different models not having the exact same composition of words. This is because some of the parameters during inference are initialized randomly, and will thus produce slightly different results.

This should however not be a problem if one has developed a good model or, as is usually the case, intends on using a single model for one's dataset and not multiple models. But of course it limits reproducibility. A possible solution is to set these random parameters to a fixed value. This is however not used here.

2.3.2 Linguistic extensions of LDA and preprocessing

In the original LDA paper, many possibilities for extensions of LDA are discussed (Blei et al., 2003). This includes incorporating more linguistic information from the data, such as relaxing the bag of words assumption in order to include syntactic information. From a linguistic perspective this is certainly desirable. Indeed, since the publication of that paper there has been many extensions and modifications like this, aiming to improve the modeling or find new uses for it.

An early example of incorporating syntactic structure is done by Griffiths et al. (2005). They create a model where words are generated either from syntactic class or semantic theme (topic). That is, a word has either function or content, and the result is both classes and themes (topics). Boyd-Graber & Blei (2009) create the *Syntactic topic model* which adds a syntactic component to the generative process and training the model on parse trees, and the result is topics that has both semantic and syntactic coherence (for example nouns about news), instead of having separate classes for function and content.

There are also many attempts at creating topic models for POS-tagging or dependency parsing. Darling et al. (2012) use a topic model similar to LDA to discover part of speech specific topics, to use for part of speech tagging. Delpisheh & An (2014) use dependency relations to improve the bag of words assumption.

However, above examples extend the original generative model of LDA. Often this is done in order to widen the uses of the topic model, for example for part of speech tagging. In this work the focus is instead on the preprocessing of the data, before LDA is applied to it. The motivation for this is that there are few studies on what effect different linguistic data and preprocessing steps has on the topic modeling and the information sought from it.

In general, the preprocessing steps of topic modeling include removing punctuation, upper-case letters and words from a stop list. This list can be assembled in various ways based on the most frequently occurring words, either from the data or a reference corpus. Contractions can also be expanded. Preprocessing can also include removing certain parts of speech. These preprocessing steps are usually not very thoroughly described, and often not included at all. They also differ between studies. An example of this is that some studies use lemmatized corpora and while others don't. As with the number of topics, it seems that whatever combination that gives the best topic model is chosen, by trial and error. Or, no trial is done and a default set of preprocessing steps are chosen.

There are a few studies reporting on the effect of linguistically informed topics modeling. In the study by Martin & Johnson (2015) they conclude that topic modeling is more informative and effective using only nouns. Following Lau et al. (2014) they also report that lemmatizing improves the results, but it also slows down the topic modeling. They use semantic coherence for evaluation (see the next section) and find that the coherence of the topics improve using nouns only. Jockers (2013) also reports good results for nouns only, but comments that using only nouns can remove some of the information sought after. For example, he argues that if one is looking for sentiment, adjectives are probably necessary to incorporate.

Fang et al. (2012) present a novel cross-perspective topic model which models topics *and* opinions. The topics are also modelled using only nouns from the corpora. The opinions related to the topics are modeled using adjectives, adverbs and verbs. In this study part of speech plays a big role.

Guo (2012) uses dependency parsing relations to filter words as a preprocessing step for LDA, and reports improved result for their specific task of detecting spoilers. This, together with the mentioned studies above, further motivates an investigation of how topic modeling can be improved by filtering the input in different ways, based on linguistic information. It is interesting both from an information retrieval aspect and a purely linguistic aspect, because it can have an effect on both the information and the linguistic data in the resulting topics.

2.3.3 Evaluation of topic models

As with any kind of machine learning, evaluation is an important part of employing topic models. Since there is no gold standard for topic models and the optimum number of topics and other configurations vary depending on the data being modeled, this is especially important. Choosing a suitable model with the right number of topics is not straightforward and many approaches have been suggested, and there is currently no consensus of what the best method is.

There are two approaches one can choose when evaluating a topic model, one can either evaluate the model's predictability or one can evaluate the topics produced by the model. If the model is used for classification the predictability of the model is important, but if the model will be used for corpus exploration, the quality of the topics might be more important. Often it is the case that both of these approaches are used.

To test the predictability of a model, the model is run on a held out test set. The goal is to have a model that has a high likelihood on the test set. In the original LDA paper Blei et al. (2003) use perplexity per word as a measure to evaluate the model. The intuition behind perplexity is to see how many times the model needs to guess before it gets it right. The perplexity is calculated as the inverse of the probability of the held out set, and the per word perplexity is calculated by dividing the perplexity by the number of words in the held out set, as defined in equation 3 where M is the number of documents.

$$perplexity(D_{test}) = exp\left\{-\frac{\sum_{d=1}^{M} log\, p(w_d)}{\sum_{d=1}^{M} N_d}\right\} \tag{3}$$

However, in the often cited paper by Chang et al. (2009) perplexity and predictability are shown not to correlate well with human judgements of the topics produced by a model. Because of this, Chang et al. (2009) argue for evaluation methods focusing on topic quality instead of model predictability. In order to evaluate topic quality, they focus on the interpretability of a topic. To do this, they create two different tasks for humans to consider; word intrusion and topic intrusion. In word intrusion, an outsider word is inserted in the list of words representing a topic. If the intruder word is easily identified as not belonging, the topic can be considered coherent and therefore interpretable. In topic intrusion, the beginning of a document is shown (it can be considered representative for the whole document according to the authors) together with a list of four topics. Three of these are the topics with highest probability for the document, last one is a topic is chosen randomly. Similar to the word intrusion task, if the intruder topic is easily identified the model is good at assigning topics.

When it comes to evaluating the interpretability of a topic it is natural to use human judgements, as exemplified above. But, using humans takes time and resources. To make the evaluation more efficient, there have been many attempts to automatically evaluate topic quality, usually by calculating the coherence between the top n words in a topic. In general, the basis for automatically estimating word coherence is using word co-occurrences statistics. There are other knowledge-based methods using Wordnet or similar resources, but they are found not to be as good as the former (Newman et al., 2010).

Röder et al. (2015) evaluate a number of coherence measures based on word co-occurrence statistics by comparing them to human judgements. They create a framework for comparing and combining these measures, where the framework is a pipeline which consists of four dimensions or parameters for calculating the coherence.

The first part of the pipeline is segmentation, how one chooses to segment the words for comparison. It is for example possible to compare every word in a topic, or only compare a word to the preceding and following word, or variations of both. The second part is how to estimate the probabilities of the words one compares. The probability of the words can be defined as dividing the number of documents the words co-occur in with the number of documents in total. Variations of this is also possible, using a sentence or paragraph as a document. Or, the documents can instead be replaced with sliding windows of a specific number of words, where the window slides over the document, one word at a time.

The third step in the pipeline is how one compares how words condition the occurrence of each other. Lastly, the fourth variable is how to aggregate the set of confirmation values for each word-pair or word-subset. An example of this is arithmetic mean. In the end, a numeric value is produced representing the coherence for a given number of words, representing a topic. For more details about the different configurations of the coherence measures see Röder et al. (2015).

Following Newman et al. (2010) and Lau et al. (2014), Röder et al. (2015) compare the correlation between the coherence measures and human judgements. The human judgements are taken from previous research, were participants are asked to assess the usefulness or coherence of a topic by giving it a score on scale between 1 (= useless) and 3 (= useful). As a reference corpus for the word counts both the original corpus and wikipedia is tested. The correlation is then calculated using Pearson's r.

A new combination of the existing coherence measures, named *cv*, is found to be the best with a mean correlation of 0.731. Using wikipedia as a reference corpus is also found to improve the results. Interestingly, these results differ from the results in van der Zwaan et al. (2016) where instead the coherence measure named *npmi* is found to be the best. They also report low correlation for *cv* and lower correlation overall. The topic modeling in van der Zwaan et al. (2016) is on Dutch, and this could be the reason for the different results.

However, despite the performance of computational methods, using human judgment one way or another is still used. It is also needed in many cases. One cannot blindly accept the computational methods (yet). When evaluating topics, as a means to evaluate computational methods or just for evaluation of a model, the coherence or the interpretability of a topic is usually what is asked for. What this means at a higher level is a discussion that has to be left out here, but a few comments are needed.

The opinion of interpretability or coherence of a topic can differ between evaluators, and sometimes the words of a topic can capture something, but it is hard to put a label on it. Different evaluators will certainly have different labels for some topics. As emphasized by DiMaggio et al. (2013) and McFarland et al. (2013) knowledge of the domain is also important when evaluating topics, because otherwise meaningful topics might be lost. However, the need for expert knowledge depends on the domain.

DiMaggio et al. (2013) further discuss the relevance of the topic model and how to select an appropriate one. Rather than settling on one model to describe the data, they describe the process of finding a good model as quoted below.

> Think of the model as a lens for viewing a corpus of documents. Finding the right lens is different than evaluating a statistical model based on a population sample. The point is not to estimate population parameters correctly, but to identify the lens through which one can see

> the data most clearly. Just as different lenses may be more appropriate for long-distance or middle- range vision, different models may be more appropriate depending on the analyst's substantive focus. (DiMaggio et al., 2013, 582)

Thus it is not only about finding an optimal model, it is also important to find a relevant model. This is related to the number of topics, which has an effect on the topics, but it is also a discussion of what a relevant topic is.

Note that when evaluating topics there are usually a couple of topics which are useless. As argued by Jockers (2013), this should not limit the use of the topics that are useful. But of course a model with a larger number of useful topics is to be preferred.

Human evaluation is also important to make sure that the topic model correctly assigns the documents with appropriate topics, something that is currently best done with human judgements. When manually evaluating and exploring a model, visualizing the results beyond word lists can be an effective aid. It is used for example in McFarland et al. (2013). Here different visualization plots are used for exploring the data after classifying it.

Lastly, a note about model stability. As previously mentioned, LDA models produce slightly or more than slightly different topics each run, and sometimes it is not always the same topics that emerge. Wilkerson & Casas (2017) discuss this and criticize the aim to find one good model. To solve this they present a framework which aggregates the result of several runs of the same model. Here results are not aggregated but compared, but the issue of robustness is be adressed.

2.4 Topic modeling of public discourse

Topic modeling has been argued for by many as a means to explore, label and structure data in the social sciences and humanities. Many frameworks have been suggested to enable easier access to topic modeling. However it is not until recent years that the researchers in the mentioned fields have begun to catch on, considering that the original LDA paper came out in 2003.

An early example of topic modeling being used is done by Hall et al. (2008). In their study they model trends in the field of computational linguistics between 1978 - 2001 by using LDA on ACL anthology papers. They successfully use the topics, but there is no mention of preprocessing or evaluation method, except that the topics where chosen manually. Topic modeling has also been used for differentiating language usage (McFarland et al., 2013) and to study sub-corpora to find important passages (Tangherlini & Leonard, 2013). In both of these, the evaluation is done manually and a stop list is used.

Public discourse has been a common subject for topic modeling. Both formal and informal domains have been investigated, these include legal documents or blogs. Topic modeling is used by DiMaggio et al. (2013) for studying newspaper coverage of the US government arts funding. They argue for the usefulness of topic modeling for this purpose, and as discussed above emphasize the need for expert evaluation. More recent examples of topic models being used is the examining of discussions in an internet mental health support group (Carron-Arthur et al., 2016), the analyzing of judgments of the High Court of Australia from 1903 to 2015 (Carter et al., 2016), and topic modeling crime reports (Kuang et al., 2017). Jacobi et al. (2016) also argue for LDA as useful tool for following trends in journalistic papers.

A very relevant study for this work is by Hägglöf (2014). The purpose of her work is to do a pilot study for automatical organization of online Swedish political discourse by testing two different topic modeling methods, LDA and temporal chi-squared(X^2). Data from the Swedish Riksdag is used also used in the

study, however more document types are used in the present work. Two other corpora of social media and newstexts are also used. The 200 most frequent words are removed from the corpus, as are words which only occur once. Punctuation and verbs are also removed. Training the LDA model, different numbers of topics are used and it is concluded that LDA is not a successful model for public discourse, something the results in this work contradict.

The evaluation is done manually, by creating gold standards for one of the corpora, the Riksdag. There is a problem with this, because manually creating gold standards before training the model requires knowledge of what topics there are in the corpus. The result of creating gold standards will rather be which topics one *expects* in to be in the corpus. The granularity of the topics is also an issue. In the examples provided, *labor policy* and *youth* are two examples of gold standard topics. How should one know what granularity one should choose for the topics? The topic model might come up with a topic such as ”youth unemployment”. This topic would be wrong according to the gold standard, but possibly correctly classify a document. This means that one still has to manually do the evaluation, in order to compare the topics and the document being classified.

Table 3 shows a summary of the mentioned studies, to demonstrate the versatility of topic modeling, but also some of the factors that need to be taken into account when doing topic modeling. The prominence of LDA can also be observed. Although this is by no means an exhaustive list over all the studies which uses topic modeling, it justifies using LDA for the purposes of this work. It also show the general lack pf preprocessing, or at least the reporting of preprocessing.

Study	Model	Evaluation	Stop words	POS	Filter	Other	Corpus size	Nr of topics	Software
Carter et al. (2016)	LDA	Tested against new documents	Yes	No	50< >50%	bigrams, unigrams	7476	10, 50	Gensim
Carron-Arthur et al. (2016)	LDA	specificity, coherence (Mallet), domain knowledge	Yes	No		you're etc removed	113004	25	Mallet
McFarland et al. (2013)	LDA	Domain knowledge	Yes	No					Stanford Toolkit
Hall et al. (2008)	Dynamic topic model (Builds on LDA)	Manually		No			12500	100	Self-developed
DiMaggio et al. (2013)	LDA	Manually, Mutual Information	yes	No			ca 8000	12	Self-developed
Kuang et al. (2017)	IDF, NMF		yes	No	5<		805618		Self-developed
Tangherlini & Leonard (2013)	STM (Builds on LDA)	Manually		No			Toolkit	Toolkit	Mallet
Jacobi et al. (2016)	LDA	Perplexity, Manually		VB,NN, JJ, PM	20<, >25%	stemming	51528	10,25	tm package in R
Hägglöf (2014)	LDA & Temporal chi-squared	Manually created gold standard		No VB	200 most frequent		7077, 10 million	100	

Table 3: An overview of the studies mentioned using topic modeling for corpus exploration.

3 Data

The data used here is the Swedish parliament documents and newstexts from Swedish journals and newspapers. The parliament data can be downloaded online, but it is also available through Korp, the corpus infrastructure tool of Språkbanken. There, the data has been enriched with automatic linguistic annotations, see section 3.3. This version of the data is used in this thesis. The newstexts are not available in Korp and are thus not annotated beforehand. To annotate this data, the tool Sparv is used, see section 3.3. This is the same tool that is used to annotate the Korp data.

3.1 The Swedish Riksdag data

All the documents and records from the Riksdag proceedings and correspondence are freely available online, known as *Riksdagens öppna data* (The Parliament open data).[3] They span between 1971 to present day, with the exception of a few document categories missing from the earlier years. From around 1990 the documents are born digital, before that they were digitalized using OCR. This may lead to some inaccuracies and misspellings in the text.

There are 20 different document categories, see Appendix A for the full list. From these, seven were chosen. In order to motivate why these were chosen, the next section gives a brief overview of the work of the Riksdag.

3.1.1 The Swedish Riksdag

Sweden is a constitutional monarchy and a parliamentary democracy. The King of Sweden is the head of state, although his role is only representative. Instead, the Riksdag is the highest decision-making assembly. The role of the Riksdag is to make laws, determine the central budget and control the government. In addition to these main duties, the Riksdag also has other responsibilities. The Riksdag monitors the work of the public sector and the EU-work, follows up on the scientific progress and work with foreign policies (Bäck et al., 2015)(Regeringen, 2017).

The Riksdag is chosen by the Swedish citizens in general elections. In practice, this is done through a party system, where citizens vote for political parties and their candidates to be part of the parliament. General elections are held every forth year on the second Sunday in September. Between 1970 and 1994 the parliamentary period was instead 3 years. On the same day elections for county and municipal council are also held. Everyone who is a Swedish citizen, over 18 and have been registered as a resident in Sweden are eligible to vote (Riksdagen, 2017b) (Riksdagen, 2017d).

After an election, the seats in the Riksdag are distributed. The parliament has 349 seats which are assigned proportionally to representatives of the political parties according to the number of votes. In order to be eligible for a seat in the parliament a party needs to obtain at least 4% of the votes. The Riksdag then votes for a Speaker of the parliament who proposes a new Prime minister. If the Prime minister is accepted he or she is then allowed to form a Government (Riksdagen, 2017c).

The working year of the Riksdag starts in September each year with an opening meeting, the Riksmöte, and ends the following year. The main work of the Riksdag is carried out in the Chamber, where meetings are held and decisions are made related to all the matters listed above. The Chamber is also the arena for political debates, questionings and foreign policies debates. The Chamber is also room for other kinds of

[3] `https://data.riksdagen.se/data/dokument/`

debates, as well as the annual budget debate and debates on EU affairs.

The decisions made primarily concern the changing of laws or the forming of new ones, taxes and how these should be distributed and used. In order to process and prepare for these decisions, the parliament is divided in committees which each are in charge of different policy areas. There are both temporary and permanent committees. The committee are made up by the members of the parliament, proportionally to their respective parties seats in the parliament.

The proposals for new legislation or decisions come in two forms, either submitted by a member of the Riksdag, known as a *motion* or as Government bill, a *proposition*. A proposal can also come from bodies appointed by the Riksdag. The *propositions* are long, thorough proposals. They are usually preceded by an report/investigation by an Government-appointed commission of inquiry. The reports are known as *statens offentliga utredningar (SOU)*. The *motions* are handed in during a special period of time each year, the *allmänna motionstiden* or in relation to an already made proposition. They can be written by a single member of the Riksdag or a group. Each year, about 3000-4000 *motions* are handed in, and although only about 5% of them are realized, they are a way of the members of the Riksdag to work for for interests of their voters (Bäck et al., 2015).

Once a proposal has been handed in, the Chamber is informed and the proposal is sent to an appropriate committee for consideration. Other committees may hand in their opinions on the proposal to the main committee, called *yttranden*. The committee then returns a proposal for a decision to the parliament, a *betänkande*. This proposal is then up for debate in the chamber, and the parliament votes for a decision. The proposals may also be debated during other stages in the process.

All the proceedings in the Chamber are recorded and stored in *protocols*. As part of the work of controlling the government, the members of the parliament may also debate the Government's performance of its official duties. This is done by either writing a question, a *skriftlig fråga*, or an *interpellation* to a member of the Government. The former is answered in text, and not debated in the Chamber. They are intended to influence a decision or gather information. An *interpellation* is answered orally, and it's purpose is to bring up matters for debate and are not followed by a decision. After an *interpellation* is answered, a debate may follow where other members of the parliament may join. Around 400-500 *interpellations* are handed in each year and around 700-1000 *skriftliga frågor* (Riksdagen, 2017a) (Bäck et al., 2015).

3.1.2 Selected documents

The purpose of using the Riksdag's data is to see how the subject of housing policies has varied, and how this subject has been talked about. On this basis, the documents deemed to cover debates, discussions and proposals are chosen. The scope is Swedish housing policies, therefore all documents related to EU matters are disregarded.

An overview of the selected documents can be seen in table 4. As can be seen in the table, the categories *interpellationer* and *skriftliga frågor* don't cover the full time span. These files are broken on the parliament webpage, they are present but empty. Therefore they are not available in Korp either. The timespan ends in February, 2016. This is the timespan available from Korp. Unfortunately there was no time to include data from more recent times. The *proposition* documents from 2006-2009 are corrupted, this also from the Korp data. The documents were split up according to parliamentary periods. This is to able to compare the terms, but also to avoid doing topic modeling over a too big span of time. Topics will have varied over time and this might confuse the topic modeling.

The *protocols* are chosen because they contain all the debates and proceedings in the Chamber. The *motions* are chosen because they reflect the opinions and proposals of the members of the Riksdag. For

the same reason the *propositions* are chosen. *Skriftliga frågor* and *interpellationer* are chosen because they are the questions and inquires of the Riksdag regarding the Government's work. Answers to these questions are also included in this.

The categories *betänkande* and *statens offentliga utredningar(SOU)* are relevant, but these documents are very long and contain thorough investigations. The proposals are also very long. Because of this, and because a single document of these categories can be assumed to only cover one subject, only the first part of these documents is used. This is done by extracting the first 3000 words for the *betänkande* and the first 5000 words for the *propositions* and *SOU*, ignoring the first 500 because they usually contain a list of contents for the *propositions* and *SOU*, with the hope that this part covers the documents topics well enough.

Some of the longer documents do contain summaries, but unfortunately there was no time to solve how to extract only these, since there is no marking in the text indicating where the boundaries of the summaries are. Because not all of the documents contain a summary, the method above was preferred.

Out of the documents that were left out 4 categories are strictly administrative, such as speakerlists. The others are EU-related, long investigations or reports, or committee documents. The committee documents were left out because they concern the work of the committees. The result of this work is a *betänkande*, and the start is a proposal (a *motion* or a *proposition*), so the subjects of the committees are already in the selected categories. There is of course lots information to gather among the rejected documents, but this is unfortunately beyond the limitations of this work.

Downloading the chosen categories from Korp and extracting the files took about 2 hours per document category. All the documents from one category are stored in a single XML-file. Splitting this file into separate documents and then sort them took about 7 hours per document category. In the cases where only the first 3000 or 5000 words are used, this took an additional 2 hours.

Document type	**Description**	**Nr of documents**	**Average document length**	**Period**
Betänkande*	Committee reports with proposals for decisions in the Riksdag.	20 993	2332	1971– 2016
Interpellation	A formal question from a member of the parliament to the government	7384	357	1998–2016
Motion	A formal proposal by a parliament member, submitted once a year.	123 129	680	1971– 2016
Protokoll	Protocols over the daily meetings in the parliament, including all debates.	6392	27866	1971– 2016
Proposition*	Proposals for legislation from the Government.	6030	4906	1971–2016**
Statens offentliga utredningar*	Reports from committees of inquiry appointed by the Government, in preparation for submitting a proposal.	3169	3304	1994–2016
Skriftliga frågor	Shorter, written questions from a member of the parliament to the government.	26 402	228	1998–2016

*Shortened documents are used. **Between the years 2006-2009 most of the documents are corrupted.

Table 4: Overview of the chosen document types.

The time marking of the documents vary. Most of them are marked with date, and in the relevant cases which political party is behind it. But some of them are only marked with year or working year of the Riksdag. Especially the older documents often lack proper date marking. To the extent that it has been possible, the data has been sorted after date and year. For the topic modeling, the data is split up in parliamentary periods. The periods and number of documents can be seen in table 5. This is because this is a natural split in the data and it is interesting to see if the data differs between these peroids. For the topic modeling it is also good to have data that doesn't span over too long time, because the language and topics might differ over time, and thus affect the modeling. All the documents are also not available for the whole timespan.

Parliamentary period	**Nr of tokens**	**Nr of documents**
1970–1973	16979308	5138
1973–1976	19270087	6780
1976–1979	18921011	7022
1979–1982	19770577	8273
1982–1985	21932744	9476
1985–1988	22146856	10835
1988–1991	21950380	12717
1991–1994	21802620	11602
1994–1998	28752522	12856
1998–2002	35523264	23431
2002–2006	43596267	28568
2006–2010	38026084	24600
2010–2014	37224643	21834
2014–2016	13726147	10323
Total:	**359622510**	**193455**

Table 5: Periods for the topic modeling.

It is important to note that the language and structure of the different categories vary. In some of the shorter documents and in the debates, the language is closer to spoken language, and some of the *protocols* are pure transcriptions from debates. However, even these debates are in a more formal style than you would expect from everyday speech, due to the nature of political debates. They are also edited during the transcription in order to standardize syntax and morphology (Dahllöf, 2012).The domain itself also affects the language in all the documents, resulting in lots of domain specific words. The effect this has on the topic modeling will be shown in later sections.

3.2 Newstext

To further analyze the public discourse, another domain of the discourse is chosen. To analyze the media, newspaper and magazine articles have been downloaded from the Media Archive provided by Retriever. [4] The access was provided by Hyresgästföreningen. The Media Archive is the largest archive of daily press in the Nordic countries. Ideally one would use all the published newstexts available, but this was not possible due to limits in data storage and time.

In order to find all the newstexts concerning housing policies a search term list was made together with people from Hyresgästföreningen who are knowledgeable of housing policies. See Appendix D for the

[4] `https://www.retriever.se/product/nordens-storsta-mediaarkiv/`

search terms. Using the Media Archive's search tool, all the newstexts containing the Swedish word for housing, *bostad*, in all its forms, and at least one of the words in the search term list were downloaded. Using the selected search terms captured both relevant and irrelevant newstexts. The topic modeling helps us sort out the relevant ones for further analysis.

All the available newstexts were originally published on the web, no printed media is included.The time span of these newstexts is 2000–2015. Before 2000 there are no newstexts available. For the topic modeling, the data is split up in two 5-year period and one 6-year period, to be able to compare the years and avoid a too long time span. These periods are 2000 to 2004, 2005 to 2009 and 2010 to 2015. In table 6 the number of tokens and documents per period can be seen.

Period	**Nr of tokens**	**Nr of documents**
2000–2004	19 054 870	52 007
2005–2009	59 579 913	122 324
2011–2015	77 340 466	171 903
Total	155 975 249	346 234

Table 6: The different periods for the newstexts data.

In total the newstexts come from 1786 different sources. Most of these sources only contribute with a few newstexts, and there are a few dominant sources. The 100 top most occurring sources can be seen in Appendix B. The 25 most common can be seen in figure 1. The number of newstexts are growing each year, probably because the web has been and is growing as a medium. In general, the number of newstexts drop during the summer months, presumably because of summer holidays.

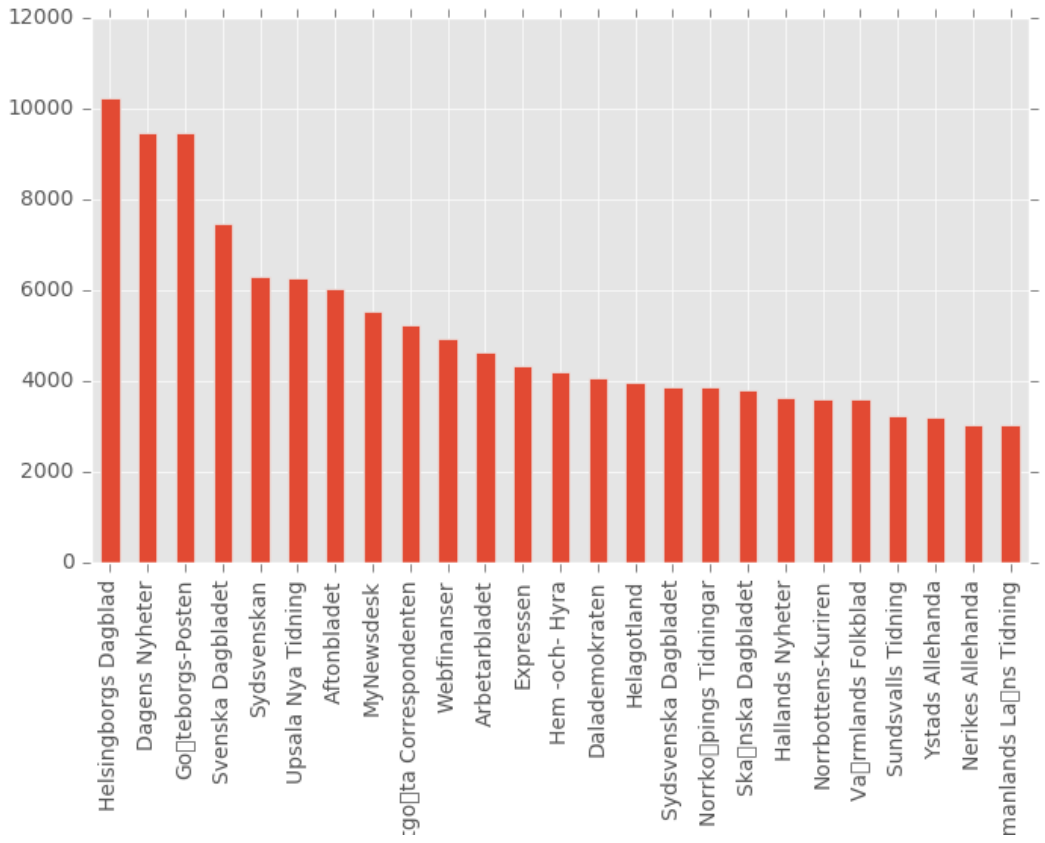


Figure 1: The 25 most frequent sources with their frequency over the whole time span.

As with the Riksdag, it is important to note that the language in the newstexts differ from the language in the Riksdag. The language in the newstexts is more similar to spoken language, and the vocabulary is closer to everyday Swedish.

3.3 The corpus infrastructure Korp

To be able to choose and filter the data based on linguistic criteria, linguistic annotations are needed. For this the corpus tool Korp is used.[5] Korp is a corpus infrastructure tool developed by Språkbanken (The Swedish Language Bank). The aim of Korp was originally to provide a tool for linguistic research, but it has since then grown and now contains a considerable amount of modern and historical Swedish corpora. Korp gathers many resources for corpus exploration and statistics, and is continually being updated and extended (Borin et al., 2012).

Korp consists of three parts. The Korp backend which has web services for searching and exploring corpora and the Korp front end with a graphical user interface. The third component is the Korp pipeline for the annotation and exporting of corpora, which has grown into an independent tool, now developed as the Sparv pipeline (Borin et al., 2016).[6] This is the part used here.

[5] https://spraakbanken.gu.se/swe/node/1535
[6] https://spraakbanken.gu.se/swe/node/19799

The pipeline can handle both partly annotated corpora and corpora without annotations. The input format is plain XML and the pipeline can tokenize, split sentences and paragraphs of the input data. The annotation process provides a number of annotations, among them compound analysis, lemmatization, part of speech (POS) - taggning, named entity recognition and tagging of dependency relations. As previously mentioned, only a subset of these are used, namely the POS tags, the dependency relations and the lemmatization. Lemmatization is done based on a morphological lexicon available in the form of SALDO, a Swedish semantic lexicon and lexical resource (Borin et al., 2013). This lexicon also provides word senses and compound analysis. The POS-tagging is done with a trigram tagger trained on the SUC 3.0 corpus, HunPos. The tagging of dependency relations is done with the MaltParser, trained on Talbanken (Borin et al., 2016).

The data from the Riksdag is already available for download through Korp, and thus has all the annotations needed. In order to have the same linguistic annotations for the newstexts Sparv was used. Sparv is available through a web interface, but due to the large quantity of data this was not used. The newstexts were instead annotated using a downloaded version of the pipeline which took approximately 5 hours per years worth of publications.

4 Method

This section describes how the work was carried out. Different filters based on linguistic information were designed and applied to a test set of the data, which resulted in different versions of the test set. Topic models were trained on each version using the Python library Gensim. The models were evaluated using human judgement, two semantic coherence measures and perplexity. By comparing the correlation between human judgements and the two coherence measures, the automatic methods were evaluated. The highest rated model by the human evaluators was used to explore the rest of the data.

The parliamentary period 2010–2014 from the Riksdag was chosen as the test set. The combination of filters resulting in the highest rated model from this test set was applied to the rest of the periods of the Riksdag data. Then, topic models were trained on these periods. Using the topic models for respective period, the documents from the same period was classified with topics. These classifications were then used for investigation of the data set.

The filters from the five highest rated models from the Riksdag test set was also applied to a test set of the newstext data. The test set for the newstexts was the documents from the period 2010–2015. Models were then trained on the different versions of the newstexts' test set and evaluated using human judgement and the best performing coherence measure from the Riksdag. As with the Riksdag, the highest rated combination of filters for the newstext was applied to the rest of the periods and a model was trained on respective period. The models were then used to classify all the documents and the classifications was used as a basis for investigation of the data set.

As previously stated, the language in the two data sets differ, and this is the reason why the highest rated combination of filters for the Riksdag data is not used directly on the newstexts. Instead, the top five highest rated combination of filters are tested on the newstexts, with the hope that the positive effects of these filters are general enough to be useful in this new domain. Also, after initial inspection, the filters from the Riksdag that used a stop list was changed to instead use a new frequency filter. Ideally one would test and evaluate all the filters that were tested on the Riksdag, however this was not done due to time limits. The same holds for the choice to use the best performing coherence measure from the Riksdag on the newstexts, ideally both coherence measures would be tested on both sets. However, here the hope is also that the performance of the coherence measures generalizes to other domains.

The process of deciding filters and combinations of filters, and evaluating them, has been a continuous back and forth process more than the linear process described here. Initially, the filters were tried and the models were manually evaluated, because many of the models were easily dismissed due to apparent useless topics. Sometimes the effect of a filter needed to be evaluated before it was tried on a large scale. This process was repeated until the different filter combinations below were determined. For simplicity's sake a description of this process is left out of this work.

The first part of this section describes the different linguistic filters, the second describes the topic modeling and the last is about the evaluation.

4.1 Preprocessing and linguistic filters

Before the topic modeling with Gensim all the documents from the original corpora were processed and rewritten into new files with filters applied. There were several filters which were combined in different ways.

By using Korp-annotated data, the data is already tokenized and with white-space is removed. The data is in the form of XML-files where every token is an element with attributes. The attributes contain all the linguistic information available for a token, such as lemma, POS or compound analysis.

Three groups of part of speech are tested. The first contains all parts of speech, as is the norm in topic modeling. The second group is nouns, verbs, adjectives and participles, from here on called *POS2*. The third, following Martin & Johnson (2015) is only nouns. See table 7 for an overview. Initially, proper nouns were also included. However, after careful consideration they were removed. Proper nouns such as names for countries or cities were very informative when they showed up in the topics, but names for persons easily took over many topics, making them useless.

Group	Meaning
All POS	all parts of speech
POS2	verbs, nouns, adjectives, participles
NN	only nouns

Table 7: Overview of the POS groups.

A separate filter was created based on dependency relations. The relations were chosen with the aim to find the meaningful parts of the sentence. The chosen relations are: Agent, object adverbial, direct object, predicative attribute, place adverbial, subject predicative complement and other subjects. When filtering part of speech and dependency relations, a tag is added to each word with the POS/dependency relation. This is to avoid ambiguity between words.

Punctuation and numbers are removed from all corpora, and all words are changed to lower-case. Unless stated otherwise, tokens that occur in 50% or more of the documents are removed. This frequency filter is called **filter 1**. Words that occur in less than 5 documents are also removed. However, the Riksdag consists of up to 7 sub-corpora for each of the document types, and some words are very frequent in some types and rare in others. Because of this, the filter is changed to removing words that occur in 50% of the documents of each sub-corpus.

Ideally, a frequency filter would suffice to remove the unnecessary words. Initially when testing the filters on the Riksdag data the goal was to avoid the use of a stop list. The reasons for this are many. If a suitable frequency filter is found, this filter can be used or at least used as a starting point, when modeling the other data sets. Also, because the Riksdag data set spans over a long period of time a stop list crafted

for one period may not be suitable for other periods, or other data sets. But, initial experimenting with frequency filters showed there was a need for a specific stop list.

At first, a general stop list for Swedish was used. However, this did not have the desired effect because the Riksdag data contains many domain specific words, which flooded the topics and made them useless. These domain specific words did not add any meaning to the topics. To try to solve this, filters with different numbers of the top n-frequent words removed were tested. These filters did not succeed in removing all domain specific words, and therefore these words were instead added to the stop list was manually.

Since every token in the data also is tagged with its lemma, an extra list of lemmas to be excluded was created. By doing so, the writing of the many forms of a word is avoided. This list was added to the stop list. The lemma list and the stop list can be seen in in Appendix G. The lemma list contains about 150 lemmas and the stop list a little less than 500 words and punctuation marks. Note that there might be some overlap between the lists. Punctuation is also removed in previous steps, therefore the list of punctuation marks in the stop list is not exhaustive.

Lemmatization is something that could also have an effect on the topic modeling, and therefore the models here are trained on lemmatized data. There are also models trained on data without lemmatization, so one can compare the difference.

In table 8 all the filters are shown, sorted by POS-group. In total 18 different combinations of filters were tested. If nothing else is stated, all filters had the frequency **filter 1** applied. All groups are tested without frequency filter, with lemmatization, and with lemmatization and stop list. The all POS and the POS2 groups are also tested with filters based on dependency relations. The POS2 group has 5 more filters. The reason for this is that in initial manual inspection by the author it was found that the POS2 group and the all POS group had similar topics, and the effects of the filters were also found to be simliar. Because of this, and the need to limit the scope of this work, only the POS2 group was tested with extra filters. This is also the reason why the NN group is only tested with three types of filters.

Applying filters to the corpus and rewriting the files took ca 1 hour each time.

All POS	**POS2**	**NN**
No frequency filter	No frequency filter	No frequency filter
Lemma	Lemma	Lemma
Lemma, Stop	Lemma, Stop	Lemma, Stop
Lemma, Stop, Deprel	Lemma, Stop, Deprel	-
Lemma,Deprel	Lemma, Deprel	-
-	Stop, Deprel	-
-	Deprel	-
-	Stop	-
-	Deprel, no frequency filter	-
-	Only frequency filter	-

Table 8: Filters for the Riksdag corpus

The linguistic filters applied to the newstext data can be seen in table 9. These filters were chosen based on the results from the topic modeling of the Riksdag data and manual inspection. Through the initial manual inspection using only a frequency filter was found to work better for the newstext data than the Riksdag. The stop list for the Riksdag data is very domain specific, so it couldn't be reused. Because of this, instead of making a new stop list, a new frequency filter was made. The alternative filter, named

Filter 2, removes the 300 most frequent tokens in the data and tokens that occur in 75% of the documents. Applying a filter and rewriting the newstext corpus took 2 hours each time.

All POS	**POS2**	**NN**
No frequency filter	No frequency filter	
Lemma	Lemma	Lemma
Lemma, Deprel	Lemma, Deprel	Lemma, Deprel
Lemma, alternative frequency filter	300 most frequent, 75%, lemma	300 most frequent, 75%, lemma
	Only frequency filter	

Table 9: Filters for the newstext corpus

4.2 Topic modeling with Gensim

The topic modeling was implemented using the python library Gensim.[7] Gensim is easy to use and especially suitable for larger data sets (Rehurek & Sojka, 2010). Evaluation can be carried out with methods in the library, such as perplexity and various coherence measures. Except for the common LDA model there are other topic models available, such as HDP and LSI. These models were initially tried, but didn't perform as well as LDA and were disregarded. In the case of HDP the number of topics returned were much fewer than could be expected to exist in the corpora, and the topics were also of low quality. The topics returned from LSI were also of lower quality compared to the topics returned from the LDA model. The quality of the topics were determined by manual inspection by the author, where a low quality topic is a topic that is not understandable.

When training an LDA model in Gensim there are a number of parameters to consider, as with any topic modeling. First and foremost is finding an optimal number of topics. Guided by previous papers, experiments were run between 50 - 200 topics. After manual inspection the topic numbers 50-75-100 were chosen for further testing and then 75 topics where chosen for the filter tests. Other than this the default configurations were used.

4.3 Evaluation

Evaluation was carried out in three steps. The following was done on all of the models trained on different versions of the test set. First, model perplexity was calculated. During training, 15% of the documents were held out and then used for calculating the perplexity. The 15% were chosen at random from the corpus.

Using the coherence model available in Gensim, the two coherence measures *cv* and *npmi* were calculated. *cv* took 2-3 hours to calculate on the Riksdag data and *npmi* took 1-2 hours, depending on the filter. *cv* was found to be the best measure by Röder et al. (2015), but is contradicted by van der Zwaan et al. (2016) who finds *npmi* to be the best measure, and therefore these are compared. As a reference corpus for the word counts the same corpus as the model is trained on is used.

To assess the performance of the coherence measures, human judgements were collected. Before the human evaluation was carried out, a short manual inspection of the models were done by the author. This resulted in two models being disregarded due to them containing mostly useless topics. These were from the All Pos group, one without any filter at all, and one with lemmatized words and a frequency filter.

[7] `https://radimrehurek.com/gensim/`

The rest of the models were kept, in total 16. These models can be seen in tables in the next section.

Six evaluators each rated 8 models, with three people rating the same 8 and the other three rated 8 other. In total, there are human judgements for 16 models. The evaluators were between the age 20-30, all native Swedish speakers and with an education level of undergraduate or above. There was an equal gender division.

Following Newman et al. (2010) and Lau et al. (2014), the evaluators were asked to assess the understandability of the top 10 words from each topic. The instructions given for the rating can be seen in table 10. The instructions are translated from Swedish.

Rating	Instruction
1	I don't find the words to be belong together, I don't understand the topic.
2	I find about half of the words to belong together, the topic is semi-understandable.
3	I find the topic to be understandable, there is at most one word which doesn't belong.

Table 10: Instructions for the human evaluators.

For each topic, the mean of the human ratings were calculated and the correlation between these ratings and the coherence measures were then calculated using Pearsons r. This was done for every model. The *cv* coherence measure was found to have the better correlation with the human judgements, and was thus used as an evaluation measure for the newstexts. As stated in the previous section, five models corresponding to the five top rated models in the Riksdag test set was chosen for this.

To compare how well the *cv* corresponds to human judgements for the newstext data, this was also collected for the 5 newstext models. In this case, three evaluators evaluated all five models. Two of them had participated in the previous evaluation, one was new. The new evaluator was also a Swedish speaker between the age 20-30, with an education level of undergraduate or above.

4.4 Labeling of the topics

As stated above, the chosen models for the newstexts and the Riksdag data was used to classify all the documents in their respective corpora. These classifications were then inspected and used for the analysis of the data. This was mainly done by identifying topics of interest. Here, the interest lies in housing policies, and therefore only topics relating to this area was labeled.

The topics from every parliamentary period was manually inspected by the author and in each period one topic or more related to housing policies was found. All of these topics were simply labeled 'housing policies'. In some periods two topics were found to concern housing policies, these were labeled 'housing policies 1' and 'housing policies 2'. There was naturally a difference in words used in the topics between the periods, but all of them are considered to relate to housing policies.

When manually inspecting the topics from the newstext data, there was a lot more topics found concerning housing policies. This makes sense, because the selection was more narrow and based on a list of terms. In some cases it was difficult to determine if a topic concerned housing policies or not. For example, a topic mentioning architects could be about housing policies, but it could also be unrelated. Because of this, two topics which were clearly concerning housing policies were selected and explored.

Other than this labeling, topics were only labeled when they were cooccurring with the topics of interest or when a sample document was classified. In other cases the topics are instead displayed with their top ten words, see the Results section.

As stated on p.9, the labeling of a topic can be subjective. Some of the labeled topics used here are also displayed together with the top words and the reader can determine for themselves if they agree with the labeling. To the extent it was possible, the labeling was done with the intent to give the topics as general labels as possible, without removing the specificity and usefulness of the topics. Ideally there would be more than one labeler of the topics, but this was not possible due to time limits and resources.

5 Results

All the results below are for models with 75 topics.

5.1 The Riksdag's data

Below the results for the models trained on the differently filtered test sets from the Riksdag are presented.

The top five models according to humans ratings can be seen in the first column of table 11. In the table the mean human rating can also be seen together with the number of 3's (from the mean rating) for each of the topics. The maximum number of 3's is 75, which would mean all human evaluators gave all topics a score of 3.

The highest rated model is the one with only nouns, a stop list and the frequency filter, **filter 1** (words occurring in more than 50% of the documents and words with an occurrence of 5 or less are removed). The words are also lemmatized. In second place comes the same model, but without the stop list. The rest of the top five models are from the POS2 group, but without lemmatization. The third highest rated model is also filtered based on dependency relations, the only one in the top five based on dependency relations.

In the last column of table 11 the perplexity per word is shown, where a low perplexity is desirable. The highest rated model model according to humans would preferably have the lowest perplexity value, if perplexity is to be useful for evaluation purposes. The highest rated model does have a low perplexity, but not the lowest. Table 12 shows the top five models with the lowest perplexity per word. Interestingly, two of the models in the top five (in italics) were disregarded before the human evaluation because the topics were deemed not useful. This, together with the low overlap with the human ratings disqualifies the use of perplexity as an evaluation measure here.

Model	**Mean human judgement**	**Nr of 3's**	**% of total words used**	**Perplexity per word**
NN, Lemma, Stop	2.489	27	18%	259.02
NN, Lemma	2.409	24	23%	189.99
POS2, Stop, Deprel	2.351	16	9%	1754.59
POS2, only freq filter	2.249	14	43%	500.63
POS2, Stop	2.236	13	28%	853.37

Table 11: Top five models according to human rankings.

Model	Perplexity per word
NN, Lemma	189.99
All POS, Lemma	224.28
POS2, Lemma	244.88
NN, Lemma, Stop	259.02
All POS, no freq filter	273.54

Table 12: Top five models according to perplexity per word.

In table 13 the human ratings are compared to the best ranked models according to the two coherence measures, *cv* and *npmi*. The mean topic coherence is also shown, and for both of the measures a high coherence is desired. There is some overlap between the lists. The top model according to human ratings is the second best on the *cv* list. However, the top model according to *cv* is number five according to the human ratings. The top rated model according to humans has fallen yet another place on the *npmi* list, it is ranked as number three. The top model according to *npmi* doesn't occur in the list of the top five highest rated models according to humans.

Top 5 models, human judgement	**Mean human rating**	**Nr of 3's**	**Top 5 models, CV**	**Mean topic coherence**	**Nr of 3's**	**Top 5 models, Npmi**	**Mean topic coherence**	**Nr of 3's**
NN, Lemma, Stop	2.489	27	POS 2, stop	0.57	13	All POS, Lemma, Stop	0.074	13
NN, Lemma	2.409	24	NN, Lemma, Stop	0.566	24	POS 2, only freq filter	0.070	16
POS 2, Stop, Deprel	2.351	16	POS 2, only freq filter	0.562	16	NN, Lemma, Stop	0.068	27
POS 2, only freq filter	2.249	14	All POS, Lemma, Stop	0.558	15	POS 2, Lemma	0.066	6
POS 2, Stop	2.236	13	POS 2, Lemma, Stop	0.553	9	NN, Lemma	0.064	24

Table 13: Top 5 models compared to *cv* and *npmi*.

The correlation between the coherence measures and the human ratings is shown in table 14. Two models were disqualified at the first inspection, these are shown in italics. The *cv* measure performs slightly better in most cases, with a mean correlation of 0.68, compared to the mean of *npmi* which is 0.60. Interestingly, there is a correlation between the correlations of the two measures. Both have the highest correlation for the top ranked model by humans, and both have lower correlation for the models with dependency relations filters, compared to the other models.

All POS	**Npmi**	**CV**	**POS2**	**Npmi**	**CV**
(No frequency filter)	-	-	No frequency filter	0.491	0.664
(Lemma)	-	-	Lemma	0.593	0.769
Lemma, Stop	0.579	0.628	Lemma, Stop	0.667	0.703
Lemma, Stop, Deprel	0.733	0.714	Lemma, Stop, Deprel	0.66	0.749
Lemma, Deprel	0.351	0.73	Lemma, Deprel	0.613	0.68
			Stop, Deprel	0.554	0.572
NN	**Npmi**	**CV**	Deprel	0.579	0.571
No frequency filter	0.555	0.63	Stop	0.748	0.779
Lemma	0.676	0.724	Deprel, no frequency filter	0.613	0.527
Lemma, Stop	0.787	0.811	Only frequency filter	0.34	0.549

Table 14: Correlation with human judgements for CV and NPMI

Based on the results above, the *cv* measure was chosen as the evaluation measure for the newstexts. The model chosen to classify the rest of the Riksdag data is the top rated model by humans.

Returning to the human judgements, table 15 shows all the models with their ratings. The percentage of the original number of words is also shown. However, this number doesn't seem to have an effect on the ratings, rather, it seems it is more important which words are used. For the models using all parts of speech, using a stop list significantly improves the results, as expected. Applying a frequency filter, which has a similar effect to that of a stop list, also improves the result. In fact, in the POS2 group, the frequency filter has a better effect than the stop list, when used alone.

The filter with the selected dependency relations have different effects. This can be seen comparing all parts of speech with and without dependency relations, where the dependency relations filters have a lower ranking. This is also seen in the POS2 group comparing the same groups. However, the POS2 model without lemmatization, stop list and dependency relations has a high score. The POS2 model without any filter except the dependency filter also has a high score.

In the POS2 group models using lemmatized words have lower ratings than their respective models without lemmatization, contrary to what one would expect. However, the NN models using lemmatized words have a higher score than all the POS2 models.

All POS	**Mean human rating**	**Nr of 3's**	**% of all word used**	**POS2**	**Mean human rating**	**Nr of 3's**	**% of all words used**
No frequency filter	-	-		No frequency filter	1.978	0	48
Lemma	-	-		Lemma	2.009	6	42
Lemma, Stop	2.191	15	33	Lemma, Stop	2.200	9	29
Lemma, Stop, Deprel	2.147	13	10	Lemma, Stop, Deprel	1.938	5	9
Lemma,Deprel	1.987	0	19	Lemma, Deprel	1.858	6	12
				Stop, Deprel	2.351	16	9
NN				Deprel	2.058	5	12
No frequency filter	2.102	6	24	Stop	2.236	13	28
Lemma	2.409	24	23	Deprel, no frequency filter	2.231	10	14
Lemma, stop	2.489	27	18	Only frequency filter	2.249	14	43

Table 15: Human ratings for all models.

5.1.1 Inspection of topics

The topics presented in this section are shown with their tags, both the ones for part of speech and dependency relation. For the full list of tags and their meaning, see appendix E. As mentioned earlier, the chosen number of topics was 75. In the initial manual inspection comparing 50, 75 and 100 topics were found to contain the roughly the same topics. However, for the larger the number of topics, the topics were more fine-grained and branched. Similarly, when using 50 topics, the topics were more general and broad. More topics also resulted in an increase of useless topics. Because of this, 75 topics were chosen.

In general, regardless of filter or number of topics, similar topics were found to recur. For example, there were always topics concerning healthcare or the labor market. This is of course hard to conclude without manually labeling all the topics.

Two of the models were disregarded in the initial manual inspection. These were the models using all parts of speech and no stop list. One of them had no frequency filter and no lemmatization, and the other had lemmatized words and a frequency filter. Most of the topics from these models were full of function

words and very frequent words, leading to useless topics. An example of this is shown in table 16, the left topic. This topic is taken from the model with all parts of speech and lemmatization. As can be observed, using no stop list and only the frequency filter was not enough to remove these words. In the same table, an example of a good topic can be seen. This topic comes from the chosen model. All the words are related to education and can easily be connected. This topic also had the highest rating from all evaluators.

Most probable words	Translation	Most probable words	Translation
vi_pn	'we'	*forskning_nn*	'research'
jag_pn	'I'	*högskola_nn*	'university'
man_pn	'one'	*utbildning_nn*	'education'
när_ha	'when'	*universitet_nn*	'university'
komma_vb	'come'	*lärosäte_nn*	'higher education instititue'
denna_pn	'this'	*kvalitet_nn*	'quality'
finnas_vb	'exist'	*student_nn*	'student'
göra_vb	'do'	*utveckling_nn*	'development'
också_ab	'also'	*utbildningsminister_nn*	'minister of education'
nu_ab	'now'	*kunskap_nn*	'knowledge'

Table 16: Examples of a bad and a good topic.

The LDA model returns the topics together with the prominence, or proportion, of them in the corpora. The topics that made least sense often had a lower prominence in the corpora than the more useful topics. The topics with the highest prominence usually contained very general and frequent words, but with a broad underlying concept, such as the labor market or municipalities. Examples of this is shown in table 17, where the left topic can be interpreted as being about the Riksdags work, and the right one about municipalities and finances. However, the proportion of topics about the same thing but in different models differ a lot, probably due to the different linguistic filters. Because of this, when using the proportion of a topic as a measure, this should be kept in mind. Here, the information about prominence or proportion of the topics in the whole corpora is not used.

Most probable words in topic	Translation	Most probable words in topic	Translation
regeringen_nn	'the government'	*kommun_nn*	'municipality'
bör_vb	'should'	*landsting_nn*	'county council'
riksdagen_nn	'the parliament'	*län_nn*	'county'
åtgärder_nn	'measures'	*verksamhet_nn*	'organization'
svenska_jj	'Swedish'	*statsbidrag_nn*	'state funding'
arbetet_nn	'the work'	*bidrag_nn*	'support/funding'
mål_nn	'goal'	*stöd_nn*	'support'
finns_vb	'exists'	*region_nn*	'region'
förslag_nn	'proposal'	*kostnad_nn*	'cost'
anser_vb	'believes'	*service_nn*	'service'

Table 17: Common topics

Turning to the effects of the different filters, one topic from the three different part of speech groups are shown in table 18. These topics are from lemmatized text, a stop list is used and all models had a high rating. The topics are deemed to all be about international aid. As can be expected, the POS2 group contains adjectives and nouns. The all POS topic is similar to the POS2 topic, and also contains mostly

nouns adjectives, but also the proper noun UN. In all the groups, nouns is the most frequently occurring part of speech. In table 19 more topics from the same models are shown. These topics are deemed to be about healthcare and are all very similar, and here all but two words are nouns. In the All POS the term *National Board of Health and Welfare* could be considered a proper noun, but here it was tagged as noun and is therefore treated as one. A note about the all POS, in this model proper nouns were more frequent, and one topic was made up of only of Swedish names.

NN	Translation	POS2	Translation	All POS	Translation
bistånd_nn	'aid'	*rättighet_nn*	'right'	*sverige_pm*	'Sweden'
rättighet_nn	'right'	*bistånd_nn*	'aid'	*land_nn*	'land'
land_nn	'country'	*land_nn*	'country'	*svensk_jj*	'Swedish'
utvecklingssamarbete_nn	'development cooperation'	*konvention_nn*	'convention'	*internationell_jj*	'international'
värld_nn	'world'	*mänsklig_jj*	'human'	*bistånd_nn*	'aid'
organisation_nn	'organization'	*internationell_jj*	'internationell'	*rättighet_nn*	'right'
kärnstöd_nn	'core support'	*utveckling_nn*	'development'	*insats_nn*	'contribution'
stöd_nn	'support'	*värld_nn*	'world'	*värld_nn*	'world'
utveckling_nn	'development'	*global_jj*	'global'	*fn_pm*	'UN'
arbete_nn	'work'	*svensk_jj*	'Swedish'	*utveckling_nn*	'development'

Table 18: Comparison of the three POS groups.

NN	Translation	POS2	Translation	All POS	Translation
vård_nn	'care'	*vård_nn*	'care'	*vård_nn*	'care'
patient_nn	'patient'	*sjukvård_nn*	'healthcare'	*hälsa_nn*	'health'
sjukvård_nn	'healthcare'	*hälsa_nn*	'health'	*sjukvård_nn*	'healthcare'
hälsa_nn	'health'	*behov_nn*	'needs'	*patient_nn*	'patient'
landsting_nn	'county council'	*region_nn*	'region'	*barn_nn*	'children'
kvalitet_nn	'quality'	*patient_nn*	'patient'	*socialstyrelse_nn*	'National Board of Health and Welf
omsorg_nn	'care'	*socialminister_nn*	'Head of the Ministry for Health and Social Affair'	*person_nn*	'person'
behov_nn	'needs'	*öka_vb*	'increase'	*landsting_nn*	'county council'
möjlighet_nn	'opportunity'	*tandvård_nn*	'dental care'	*behov_nn*	'needs'
läkare_nn	'physician'	*kommun_nn*	'municipality'	*sverige_pm*	'Sweden'

Table 19: Comparison of the different POS groups.

An example of the effects of lemmatization is shown in table 20. The topics come from POS2 models with a stop list, one is with lemmatization (the same model as above) and one is without lemmatization. The topic is also the same, it is about health-care. In the topic without lemmatization there is a lot of repetition, but this hasn't affected the rating, both of the topics had the highest ratings.

POS2, Lemma	Translation	POS2, no lemma	Translation
vård_nn	'care'	*hälso_nn*	'health'
hälsa_nn	'health'	*sjukvården_nn*	'the healthcare'
sjukvård_nn	'healthcare'	*vård_nn*	'care'
patient_nn	'patient'	*vården_nn*	'the care'
landsting_nn	'county council'	*sjukvård_nn*	'the healthcare'
behandling_nn	'treatment'	*patienter_nn*	'patients'
socialstyrelse_nn	'National Board of Health and Welfare'	*läkare_nn*	'physician'
person_nn	'person'	*landsting_nn*	'county council'
läkare_nn	'physician'	*patienten_nn*	'the patient'
abort_nn	'abortion'	*personer_nn*	'persons'

Table 20: Comparison of topics with lemmatized and non-lemmatized data.

In table 21 two topics based on a model with dependency grammar is shown. The topics are from the third highest rated model,which is a model with POS2 and stop list, but no lemmatization. As with previous topics, the topics without lemmatization has a a lot of repetition, which is increased by the dependency tags. For example, in the left topic the word *boende* (housing/living) occurs both as a subject (ss) and direct object (oo). This repetition might influence the understandability of a topic.

Topic	Translation	Topic	Translation
bostäder_oo	'residences'	*elever_ss*	'students'
bostad_oo	'residence'	*skolan_ss*	'the school'
kommunerna_ss	'the municipalities'	*skolan_ra*	'the school'
boende_oo	'living'	*eleverna_ss*	'the students'
kommuner_ss	'municipalities'	*möjlighet_oo*	'opportunity'
bostäder_ss	'residences'	*ungdomar_ss*	'youth'
möjlighet_oo	'opportunity'	*lärare_ss*	'teachers'
boende_oa	'living'	*skolor_ss*	'schools'
förutsättningar_oo	'conditions'	*skolan_oa*	'the school'
möjligheter_oo	'oppotunities'	*elever_oa*	'students'

Table 21: Two topics from dependency parsing based data.

While inspecting these topics it is obvious that the LDA model captures topics based on semantic coherence, as is the purpose of the model. However, there were a few curious cases of topics. English words were always assigned to a topic of their own, as was names for persons. This raises the question of what kind of semantic information is captured by the model.

5.1.2 Performance of the selected model

Before moving forward with analysis with the help of the highest rated model, the model with only lemmatized nouns and a stop list, there is a need to check if it correctly classifies documents. To do this fully would require labeled documents and would defeat the purpose of using topic modeling. But, it is important to check the classifications of a few documents to have an idea of the reliability of the model's classifications. Here, a few of the documents identified as containing the topic labeled 'housing policies' were inspected. The document classified as having the highest probability of this topic can be seen below.

The document is a *motion*, translated from Swedish. For the original document, see Appendix F. This is deemed to be an accurate classification of a document containing the 'housing policies' topic. A few other documents were also inspected and deemed to be accurately classified with the 'housing policies' topic.

> Motion to the Riksdag 2010 / 11:C264 The building of cheap tenancies s34010 Proposal for decision in the Riksdag. The Riksdag announces to the Government their opinion of what is presented in the motion about stimulus for the production of tenancies. Motivation. The Reinfeldt government have removed the housing policy goals from before. This was done at the same time as when they removed the support for interest and investment in relation to the construction of new tenancies. The result has been disastrous consequences for the building of tenancies. The pursued politics have failed. What is needed now is politics which defends the tenancy and stimulates the building of tenancies for a reasonable cost. Such politics consists of introducing an investment support for tenancies and a review of the neutrality at the housing market. Stockholm the 26th of October 2010 Ann-Christin Ahlberg (S) Hans Olsson (S) Phia Andersson (S)

A few random documents from each document type were also classified and inspected. For the shorter document types, *interpellations*, *motions* and *skriftliga frågor*, most topics were deemed suitable. An example of this is seen below. The document is of the type *skriftliga frågor*, and the top four topics for this document are labeled 'Fishing & Culture, 'Environment & Water', 'Food' and 'Pollution'. All of these topics are deemed to fit the document. The original document in Swedish can be read in Appendix F.

> The 7th of September / Answer to question / 2009/10:99 The Swedish fermented herring. / Agriculture minister Eskil Erlandsson / Peter Hultqvist have asked me what measures I plan to take by reason of the worry for the future of the fermented herring. The question has its origins in the temporary exception Sweden has from the EU's regulations regarding limits of some toxins in certain fish from the Baltic Sea area, which ends the 31th of Decemeber 2011 / Sweden has together with Finland a time limited exception from the regulation in the EU for highest allowed amount of dioxin and PCB in fish, the exception holds for among others herring from the Baltic Sea. The exception is to a largely based on the fact that Sweden has a well-functioning system for protection of the public health, through information directed to the consumers about the consumption of certain fish species from the Baltic Sea area might need to be limited. / The fermented herring is a part of our food heritage and an important tradition in large parts of Sweden, and I want to protect that. I have therefore personally been in touch with the commission for health and consumer issues, John Dalli, and told him how important the Swedish tradition surrounding fermented herring is. The consumption is also usually very limited in both geographic region and time. / Would the exception end and one does not find another solution it would with all probability mean that all the small-scale fishing at the Baltic Sea coast would cease to exist. I don't want to participate in that. Instead I want to protect the herring, and if our experts at the agencies makes the assessment that through the present dietary recommendations and other measures one can ensure the public health I will act to find a solution that enables further fishing. / Before the coming negotiation about a continuing exception the government has delegated the National food agency and the National board of fisheries, after consulting with the Environmental protection agency, to investigate what options there is and what the consequences of thes will have. The agencies will leave their final report the 1st of March 2011. / Last but not least I want to emphasize that to solve the problem with pollution in fish from the Baltic Sea the supply of the toxins to

the Baltic Sea must be limited. Sweden actively works with this nationally as well as within the EU and globally. /

When inspecting the classifications of the document it was found that topics which had a high probability were often correct. But topics with a low probability were often not correct. (These topics were still classified as being in the document, just not with a high probability.) The *protocols* were classified with lots of topics, which is natural because they are from debates and proceedings which contain many different subjects. These topics were often of a more general type, for example about employment and finances. This is similar to the results for the *betänkande*, *SOU* and *propositions*. These document types were summarized. Despite this, the retrieved topics for these documents were deemed suitable but quite broad, usually concerning finances and organizations. However, this might be more due to the specificity of the topics and the nature of the documents than bad classification. The chosen topic model doesn't have a topic for the finances of the Royal household, for example, which is a very specific topic and therefore a topic concerning finances in general should be good enough. It is also in the nature of these documents to contain lots of words related to finances, they are political documents.

5.2 The newstext data

Following the results for the Riksdag, five filter combinations corresponding to the filters from five highest rated models for the Riksdag were chosen for the newstext data. The difference between the models is that the Riksdag data has a stop list and the newstexts instead use a frequency filter, called **Filter 2**. This filter removes the 300 most frequent words, and words that occur in more than 75% of the documents.

The results from the human judgements can be seen in table 22. This table corresponds to the top five models for the Riksdag in table 13, but without the results for the *npmi* measure. The highest rated models differ from the Riksdag data. Here, the highest rated model is with the POS2 group, frequency filter 2, and no lemmatization, as opposed to lemmatized nouns with a stop list, which had the highest scores in the Riksdag. The second place is the same as the Riksdag, but the rest of the models have different rankings. The frequency filter named **Filter 1** is the previously used filter which removes words occurring in 50% or more of the documents.

It should be noted that both the mean ratings and number of 3's are lower overall for the newstext data than for the Riksdag. here the maximum number of 3's is also 75. The lower ratings might be due to the replacement of the stop list with a frequency filter, or the nature of the data itself. It could also be that the results from the Riksdag didn't generalize well to the newstexts. Also, since the number of evaluators is quite small, differences between the evaluators will have a large effect on the ratings.

Model	Mean human rating	CV correlation	Nr of 3's	Top models, CV
POS2, Filter 2	2.08	0.616	10	POS2, Filter 1
NN, Lemma	2.036	0.759	5	NN, Lemma
POS2, Filter 1	1.933	0.37	3	POS2, Filter 2
NN, Lemma, Filter 2	1.871	0.789	4	NN, Lemma, Filter 2
POS2, Filter 2, Deprel	1.636	0.322	0	POS2, Filter 2, Deprel

Table 22: Results for the chosen models for the newstext data

As with the Riksdag data, the *cv* measure does not correspond to the ratings by humans. Here, the difference between the human ratings and the *cv* measure is larger than for the Riksdag data. The correlation is also the highest for models with only nouns.

5.2.1 Inspection of topics

The same patterns found in the topics from the Riksdag can be seen in the topics for the newstexts. Even in the POS2 topics, nouns occur the most frequent. When the topics are not lemmatized, there is a lot of repetition in the most probable words. Two examples of two top rated topics from the highest rated model (POS2, note lemmatized, **filter 2**) can be seen in table 23, showing the

Topic	Translation	Topic	Translation
kök_nn	'kitchen'	*hyra_vb*	'rent'
taket_nn	'the roof'	*hyr_vb*	'rents'
badrum_nn	'bathroom'	*bostadsrätt_nn*	'condominium'
tak_nn	'roof'	*hyra_nn*	'rent'
köket_nn	'the kitchen'	*kö_nn*	'queue'
kyrkan_nn	'the church'	*söker_vb*	'seeks'
väggarna_nn	'the walls'	*betalar_vb*	'pays'
hem_nn	'home'	*hyresrätt_nn*	'tenancy'
trädgård_nn	'garden'	*uthyrning_nn*	'rental'
byggdes_vb	'was built'	*flyttar_vb*	'moves'

Table 23: Top rated topics from the newstext data.

When inspecting the topics, some major areas or subjects can be found throughout the topics. One is the area or topic originally searched for, the housing policies area. As opposed to the Riksdag, there are more than one topic related to this. There are topics related to incidents mentioning housing or living, for example in reporting of crimes or fires. There are also topics related to the housing market, where words describing interiors or renovation are common. There are also several topics relating to finances, both private and political. Two examples of topics can be seen in table 24, which comes from the highest rated model model. The left topic concerns housing policies and finances, while the right is about crime.

Topic	Translation	Topic	Translation
lån_nn	'loan'	*mannen_nn*	'the man'
bolån_nn	'mortgage'	*brott_nn*	'crime'
bankerna_nn	'the banks'	*misstänkt_pc*	'suspected'
räntan_nn	'the interest'	*åklagaren_nn*	'the prosecutor'
ränta_nn	'interest'	*greps_vb*	'was arrested'
hushållens_nn	'the households' '	*mord_nn*	'murder'
banken_nn	'the bank'	*hittades_vb*	'was found'
hushållen_nn	'the households'	*män_nn*	'men'
räntor_nn	'interests'	*misstänkta_pc*	'suspected'
låna_vb	'loan'	*fängelse_nn*	'prison'

Table 24: Examples of topics from related to housing.

Interestingly, there are also a few topics containing words related to time, including months and weekdays. In table 25 examples of this is shown. The examples are taken from the top two models, but similar topics were found in all the models, except for the ones with dependency relations filter. This is similar to the topics found in the Riksdag, which had topics for names of persons.

Topic	Translation	Topic	Translation
november_nn	'November'	*onsdag_nn*	'Wednesday'
april_nn	'April'	*fredag_nn*	'Friday'
maj_nn	'May'	*torsdag_nn*	'Thursday'
mars_nn	'March'	*euro_nn*	'euro'
oktober_nn	'October'	*tisdag_nn*	'Tuesday'
juni_nn	'June'	*måndag_nn*	'Monday'
februari_nn	'February'	*vecka_nn*	'week'
måndag_nn	'Monday'	*börs_nn*	'exchange'
tisdag_nn	'Tuesday'	*juni_nn*	'June'
danske_jj	'Danish'	*nyhetsbyrå_nn*	'news agency'

Table 25: Examples of topics with months and weekdays.

5.2.2 Performance of the selected model

In the newstext data, there were more topics found concerning housing policies than in the Riksdag. This is further discussed in section 5.3.2. As with the Riksdag data, a few of the documents/newstexts which had the highest probability of one of the topics labeled 'housing policies' were inspected. The model were found to identify documents good. An example can be seen below. This newstext was classified with the 'lack of housing' topic. This is also what this newstext is about. The Swedish original can be found in Appendix F.

> Dangerous silence in housing policies. Signed Karlsson. In almost half of the country's 290 municipalities there is a lack of housing. This is showed in a report made by LO. The lack of housing is the largest in and around the country's largest cities. The lack of housing in Sweden is not news. But the opinions differ when it comes to what should be done to resolve the low building rate. LO advocates for, unsurprisingly, stimulus from the state.With 15 millions in direct support to new tenancies, LO wants to solve the housing crisis where the needs are the highest. The building of housing should not be fueled with state funding. Instead, to accelerate the building has to be more profitable, partly through simplification and standardization of building regulations, partly through the removal of the regulation of rents. Stefan Attefall (KD) has during his time as housing minister started a number of investigations. These have resulted in concrete proposals, especially in relation to simplification of the building process and to create standardized building regulations. It is plausible that a new legislation will be here soon. On the other hand, there is no solution to the possibly biggest problem, the lack of tenancies. The regulation of the rents of today is counterproductive to its purpose to break the segregation of housing, as it instead creates another division in the population, between the ones who have a housing of their own, and the ones who doesn't. This especially has an effect on young people, who are about to enter the housing market. To reduce the large alienation in the Swedish housing market the thresholds need to be lowered. This is done most easily with simpler and cheaper tenancies and student housings and with the phasing out of the regulation of rents. Unfortunately, it is all too quiet within the political arena, housing issues are not hot topics. But if there is no acceleration of the building of houses the risks for the future are a much more dysfunctional work market and a suppressed growth. Mattias Karlsson, Kristianstadsbladet, 20131005

A few random newstexts were also chosen and classified by the model, below is one of them. The top four topics for this newstext are labeled 'Common verbs', 'Fire & Emergency', 'Family' and 'Building

plans'. The first topic is very general, and should maybe have been removed. However, this and the other three topics do occur in the newstext.

> "Their home was eaten by the big fire". Göteborg. Renovation of the whole house at Decembergatan 35-53 is ongoing. Move in is planned to October 2016 for the first, and the last is planned for January 2017. The ones who move back arrives to a fully renovated house with everything new. Windows, kitchen, bathroom, everything. Many faults are the reason for the fire, but not everyone who were evacuated wants to return. It took a couple of phone calls before we reached someone was sure of what she wanted. – I want to return for the sake of my child. He just started school, says Malin Forsell. Her son Liam has been able to stay in his school because they weren't moved very far. The housing company Poseidon says that care has been taken to current placements at kindergartens and schools when replacement apartments were distributed. Families have in general been able to stay in Kortedala, most of the other ended up in Biskopsgården. Second hand housing in other parts of the city also occur. – I can certainly understand the ones who doesn't want to return, says Malin Forsell. The fire has left more or less deep wounds at all the tenants who were evacuated from their apartments the night between the 10 and 11 th of August. Malin had to wait about a week before she was allowed to return to her apartment. – It was a chock, Our bedroom was gone. All that was left was bars. A couple of months later she was offered a second chance to see the apartment. – Everything was still there. I was there to look after some old things, not the least some pictures. They had made it, but they smelled of smoke. They still smell, Malin says. Malins apartment on Decembergatan 51 didn't burn and she has had some help from Poseidon and her insurance company. – But you don't get a compensation for everything. Our whole life just went up in flames, says Malin who later now have started to miss things that can't be replaced. A bit worse is it for her friend Linn Olin who lived at Decembergatan 35. There was never a fire there. But the apartment was damaged by water from the extinguishing. – We could rescue some of the furniture, but I had had big problems with the insurance company, who among other things questions that I had so much food in my fridge and they don't answer and are hard to get in touch with, says Linn. Both Malin and Linn also thinks that it's hard to get in touch with someone at Poseidon who can answer questions. – Everyone refers to someone else. Why isn't there someone designated who we can turn to with everything related to this, says the both of them." Bertil Guslén, Göteborgs-Posten, 20121225

5.3 Analysis of the corpora with the help of the models

5.3.1 Riksdagen

After selecting the highest rated model for the Riksdag, this model was trained on all the parliamentary periods, respectively. The model for each period was then used to classify all the documents in the same period. The topics for each period was manually inspected, and in every period a topic corresponding to housing policies was found. In some of the periods, two topics were found. Examples of topics labeled 'housing policies' can be seen in table 26. In all the periods the topic or topics labeled 'housing policies' were fairly similar.

Topic	Translation	Topic	Translation
åtgärd_nn	'measure'	*bostad_nn*	'residence'
bostad_nn	'residence'	*hyresrätt_nn*	'tenancy'
lägenhet_nn	'apartment'	*boende_nn*	'living'
byggnadsprojekt_nn	'building project'	*lägenhet_nn*	'apartment'
lån_nn	'loan'	*fastighet_nn*	'property'
krav_nn	'demands'	*hyresgäst_nn*	'tenant'
lokal_nn	'room'	*kommun_nn*	'municipality'
kostnad_nn	'cost'	*bostadsmarknad_nn*	'housing market'
ändring_nn	'change'	*möjlighet_nn*	'opportunity'
småhus_nn	'houses'	*hyra_nn*	'rent'

Table 26: Example of topics labeled 'housing policies' from the period 1979–82 and the period 2006–2010

With the help of the chosen model classification, there is now a classification for each document. This is a list of topics the models finds in the document and the proportion of the topics. With this, and a topic labeled 'housing policies', the data can be inspected in different ways. Here follows a few suggestions and examples.

First, using the classifications one can see how many documents contain the 'housing policies' topic. To filter out the document with a low proportion of the 'housing policies' topics, documents with less than 35% of the topic was removed. Figure 2 shows the proportion of documents which contains over 0.35 of this topic in all the *motions*. Figure 3 shows the proportion of 'housing policies' in all the shorter documents. The time span is from 1998, because *interpellationer* and *skriftliga frågor* aren't available before. Figure 4 shows the same, but for the longer document. Note that the *SOU* aren't available until 1994.

Inspecting the figures on can see a peak for almost all of the documents in the period 1998 to 2002. There are also individual peaks for the different document types.

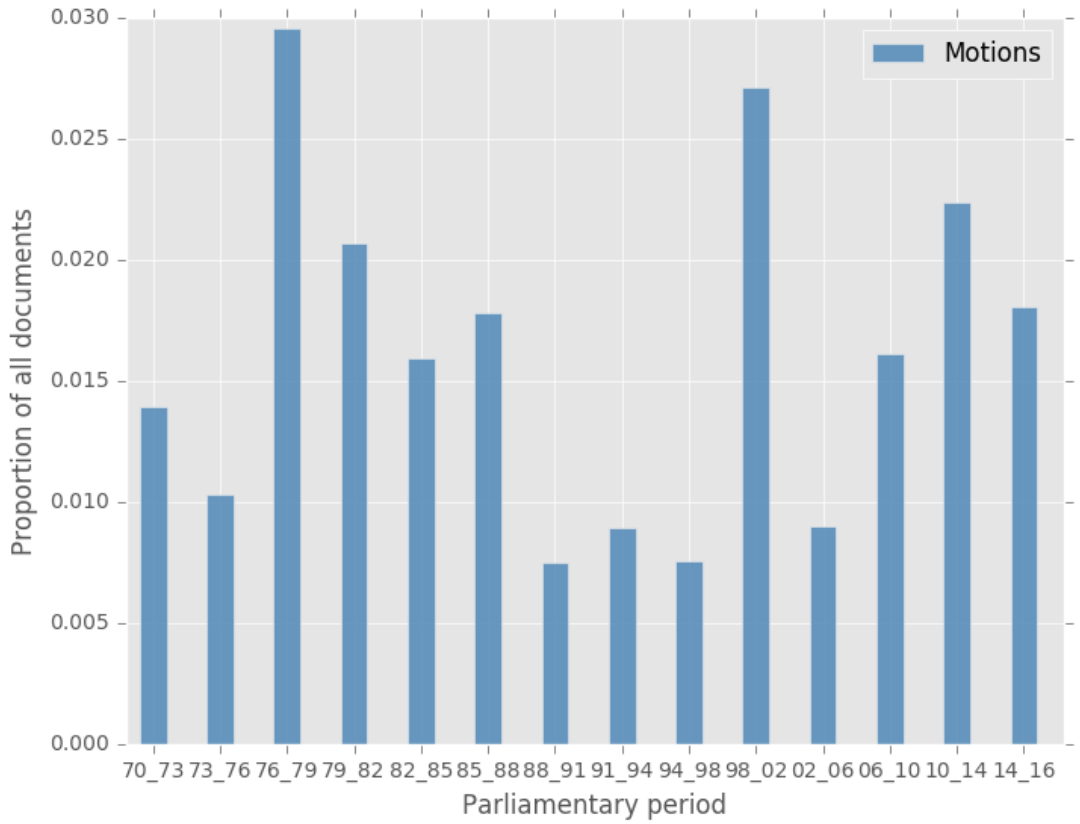


Figure 2: Proportion of documents with a proportion over 0.35 of the topics labeled 'housing policies' in the *motions*.

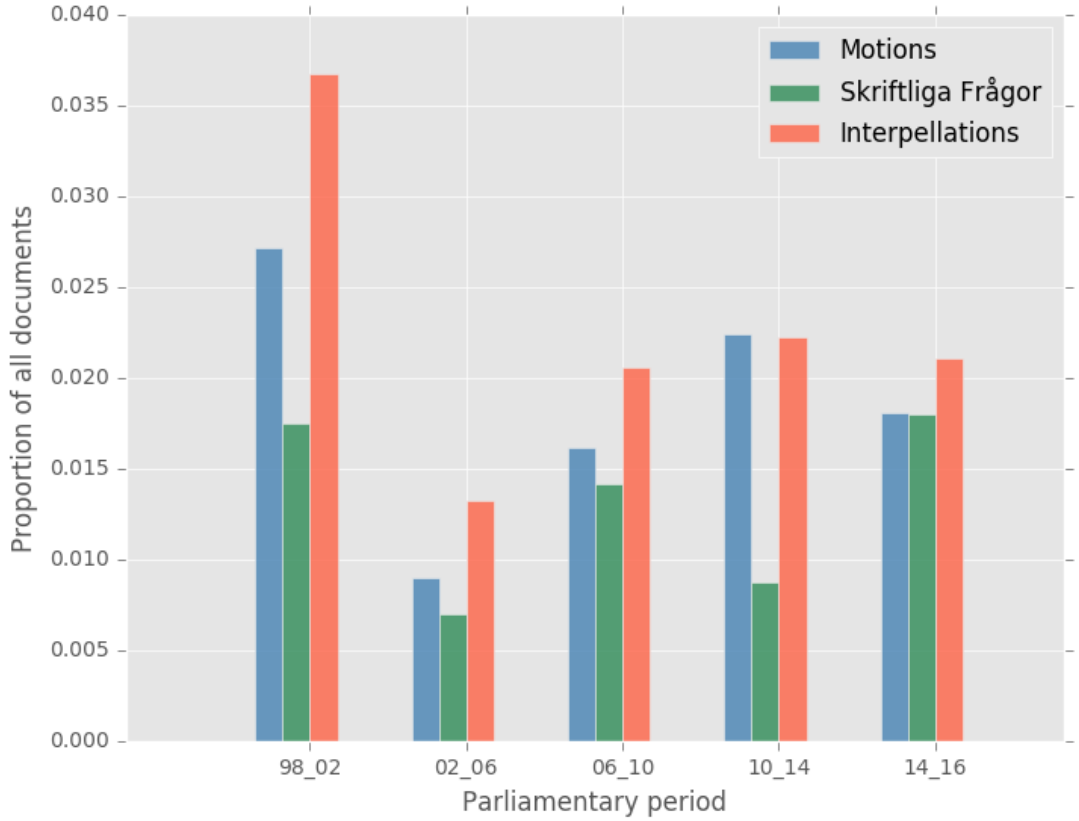


Figure 3: Proportion of documents with a proportion over 0.35 of the topics labeled 'housing policies' in the short documents.

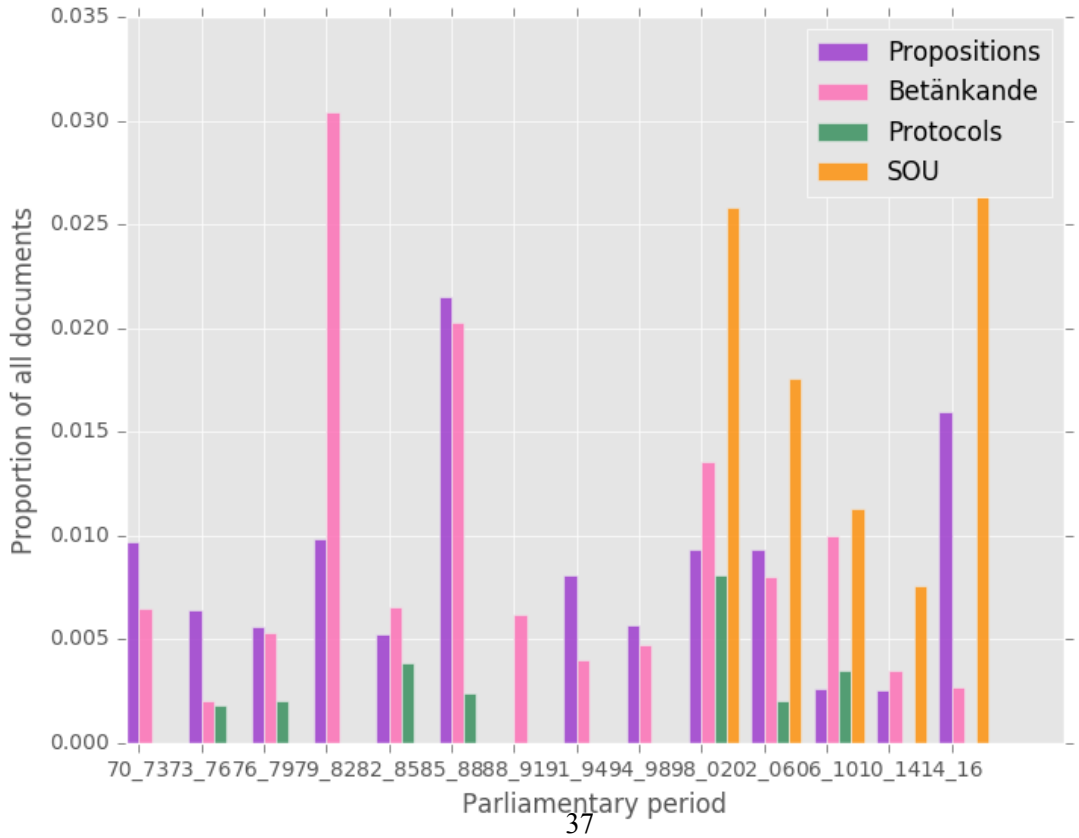


Figure 4: Proportion of documents with a proportion over 0.35 of the topics labeled 'housing policies' in the long documents.

To examine what has been talked about together with the ’housing policies’ topics one can compare what other topics are found together with this topic. In figure 5 all the *motions* from 1974 containing the ’housing policies’ topics are shown. They are shown together with the most frequently cooccurring topics with the ’housing policies’ topic. Each column represents a document, with the date they were handed in. The y-axis is in chronological order. The proportion of each topic in respective document is shown in the intersection between the rows and the columns. The most frequently cooccurring topic is about the municipalities and taxes.

The cooccurrence is calculated from the whole period 1973 to 1976, the topics shown are the most frequently cooccurring topics with the ’housing policies’ topic from that period. Ideally, the motions from the whole period would be shown, but the figure would be too condensed and hard to interpret if all *motions* were shown. Note that the cooccurring topics are also calculated from all the document types, not only the *motions*.

The most frequently cooccurring topics for the ’housing policies’ topics from the period 1991 to 1994 are shown in figure 6, together with the *motions* from 1992 and their proportion of these topics. Only the *motions* from January for that year are shown, because as with the previous figure this figure would also be to condensed to interpret if more document were added. Here there is a small shift in topics, although most of them are still concerning economy in different ways, and municipalities.

In figure 7 the most frequently cooccurring topics for the ’housing policies’ topics is again shown, but for the period 2010 to 2014. The *motions* are also shown here, but only for the month September. This is for the same reason as the previous figures. There were no *motions* from January containing the ’housing policies’ topic, otherwise these would have been shown for comparison. Here ’Healthcare’ has come up as the most frequently cooccurring topic. In 2012 the ’housing policies’ topics were found in 228 *motions*, but only in 4 of the *propositions*. It was also only found in 22 of the *interpellations*, these can be seen in figure 8.

The number of documents in the following figures is different, and this results to slight differences in size of the figures. For readability reasons these differences are kept.

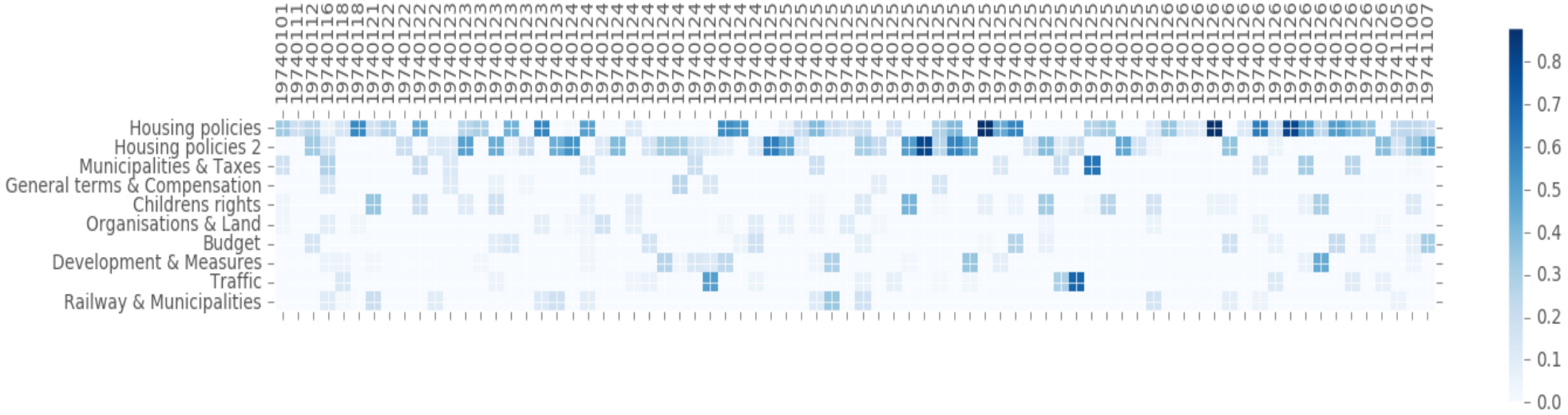


Figure 5: The occurrence of the 'housing policies' topics together with the most frequently occurring topics in *motions*, 1974. Each column represents a document. The color bar shows the proportion of topic.

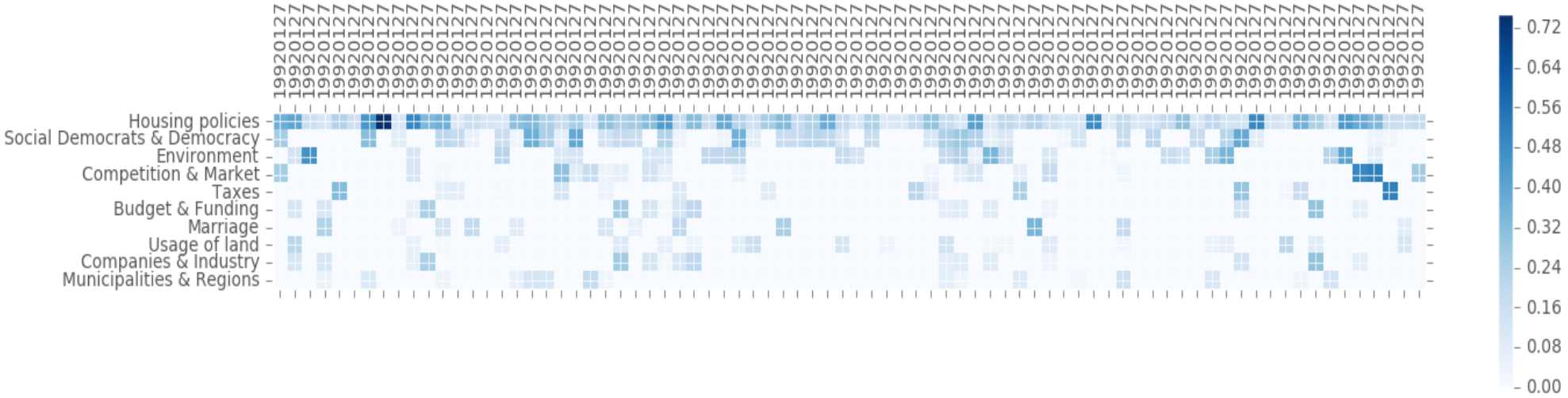


Figure 6: The occurrence of the 'housing policies' topic together with the most frequently occurring topics in *motions*, 1992, January. Each column represents a document.The color bar shows the proportion of topic.

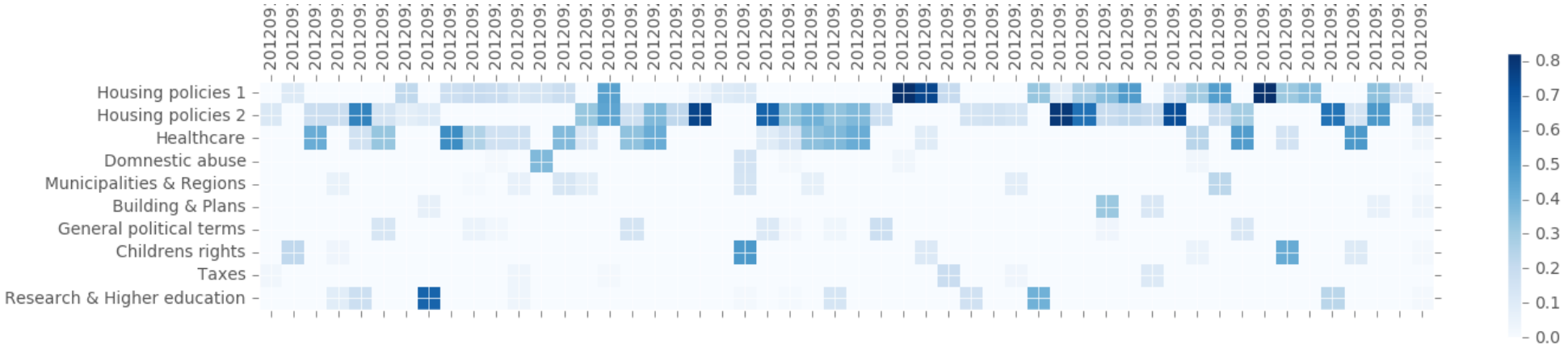


Figure 7: The occurrence of the 'housing policies' topics together with the most frequently occurring topics in *motions*, 2012, September. Each column represents a document. The color bar shows the proportion of topic.

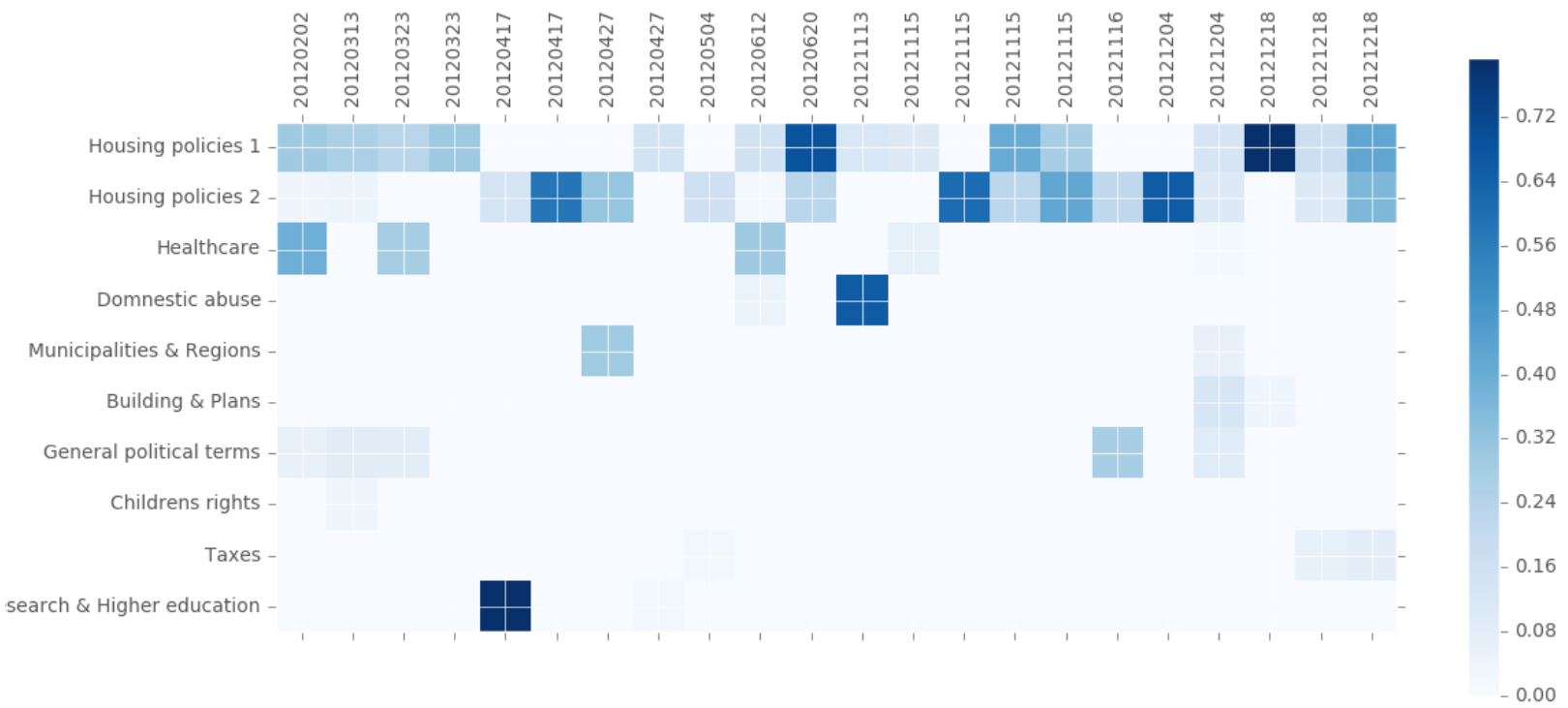


Figure 8: The occurrence of the 'housing policies' topics together with the most frequently occurring topics in *interpellations*, 2012. Each column represents a document.The color bar shows the proportion of topic.

To explore the results and the topic model more thoroughly, an interactive plot was made with the help of the Python library Bokeh.[8] A screenshot of this plot is seen in figure 9. The static plot is seen in figure 10. The static figure is similar to the previous figure, but includes all topics. It also shows all documents, not just the ones containing the 'housing policies' topics. The documents on the y-axis are in chronological order. As can be seen in the screenshot, when hovering the mouse over a square, the name of the document it represents is shown, in this case *Livet efter skyddat boende* (Life after protected housing). The topic is unnamed, but the top ten words of the topic are displayed. They include *våld*, (violence), *kvinna* (woman), and *barn* (children). The proportion of the topic is also shown. Together with the title, one can assume that the document is classified in a correct way.

This interactive plot or visualization is thus both a way to explore the data, but also a way to examine how the model classifies documents.

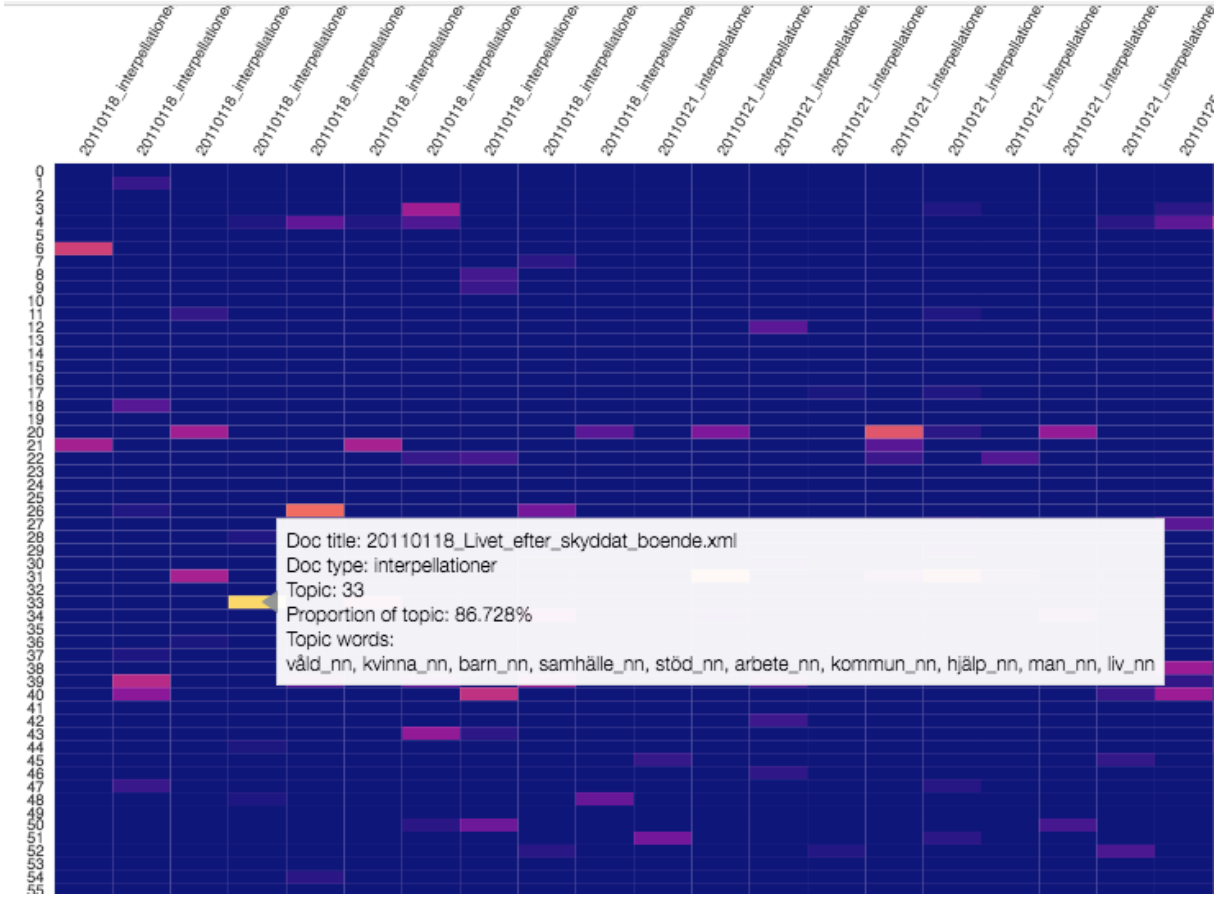


Figure 9: A screenshot of the interactive plot showing the occurrence of the housing topics together with the most frequently occurring topics in *interpellations*, 2012. Each column represents a document. The color indicates the proportion of topic.

[8] `https://bokeh.pydata.org/en/latest/`

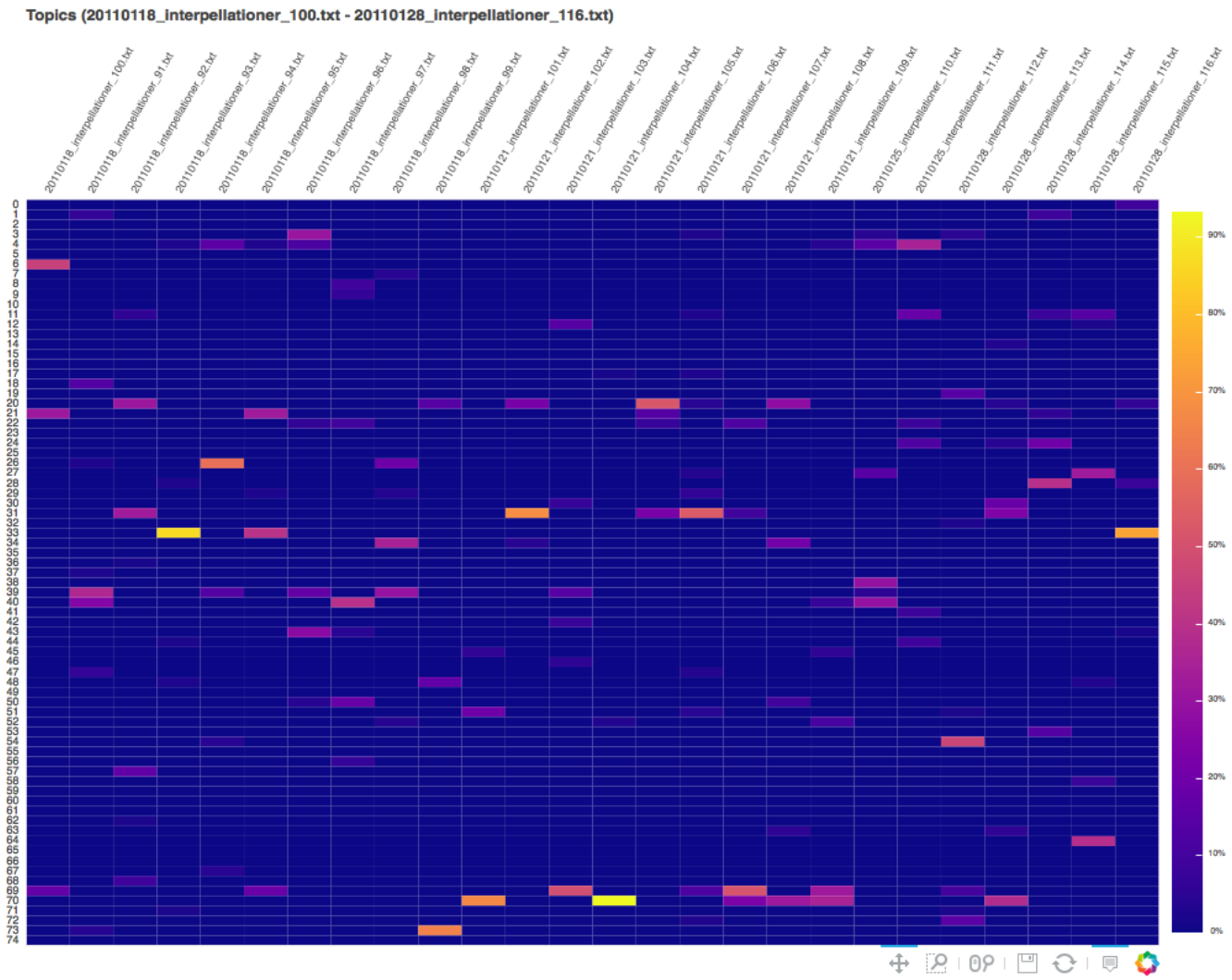


Figure 10: The occurence of the housing topics together with the most frequently occurring topics in *interpellations*, 2012. Each column represents a document.The color bar shows the proportion of topic.

5.3.2 The newstext data

As previously stated, there were more topics concerning ’housing policies’ in the newstext data then in the Riksdag data. Two topics were thus chosen for further investigation. The first topic chosen was a topic concerning rents. See table 27 for an example of this topic. We can see how the cooccurring topics for the ’rents’ topic has varied in this span in the figures 11 , 12 and 13.

The second topic is *bostadsbrist*. The topic can also be seen in table 27. It was only found in the last period, 2010–2015, something which may say something in itself. In figure 14 this is seen, together with the most occurring topics.

Topic	Translation	Topic	Translation
hyrorna_nn	’the rents’	*bostadsbristen_nn*	’the lack of housing’
hyror_nn	’rents’	*bostadsbrist_nn*	’lack of housing’
hyran_nn	’the rent’	*brist_nn*	’lack’
höja_vb	’raise’	*byggandet_nn*	’the building’
hyra_nn	’rent’	*byggande_nn*	’the building’
bostadsbolagen_nn	’the housing companies’	*bristen_nn*	’the lack’
underhåll_nn	’maintenance’	*lösa_vb*	’solve’
bostadsbolaget_nn	’the housing company’	*billiga_jj*	’cheap’
hyresgästföreningens_nn	’the union of tenants’ ’	*råder_vb*	’is’
förhandlingarna_nn	’the negotiations’	*bostadsbyggande_nn*	’building of housing’

Table 27: The rents topic from 2005–09 and the ’lack of housing’ topic from 2010–15.

The number of documents in the following figures is different, and this results to slight differences in size of the figures. For readability reasons these differences are kept.

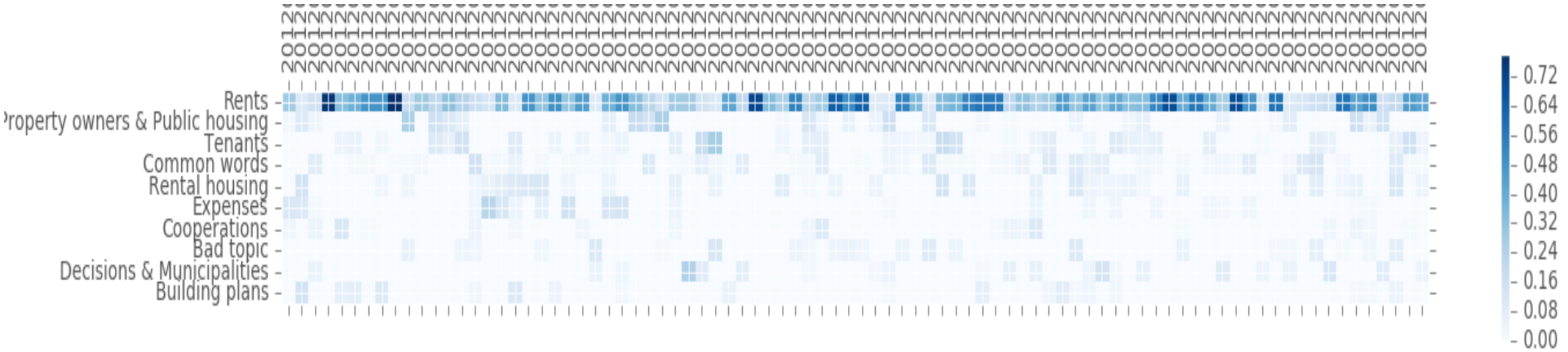


Figure 11: The occurrence of the rents topic together with the most frequently occurring topics in newstexts, first half of March, 2012. Only newstexts containing the lack of rents are included. The color bar shows the proportion of topic.

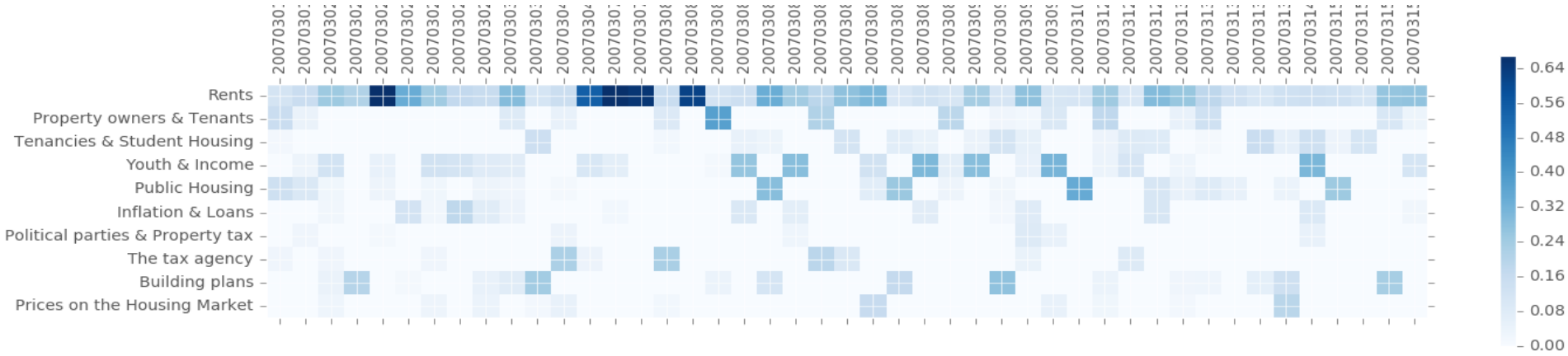


Figure 12: The occurrence of the rents topic together with the most frequently occurring topics in newstexts, first half of March, 2007. Only newstexts containing the rents topic are included. The color bar shows the proportion of topic.

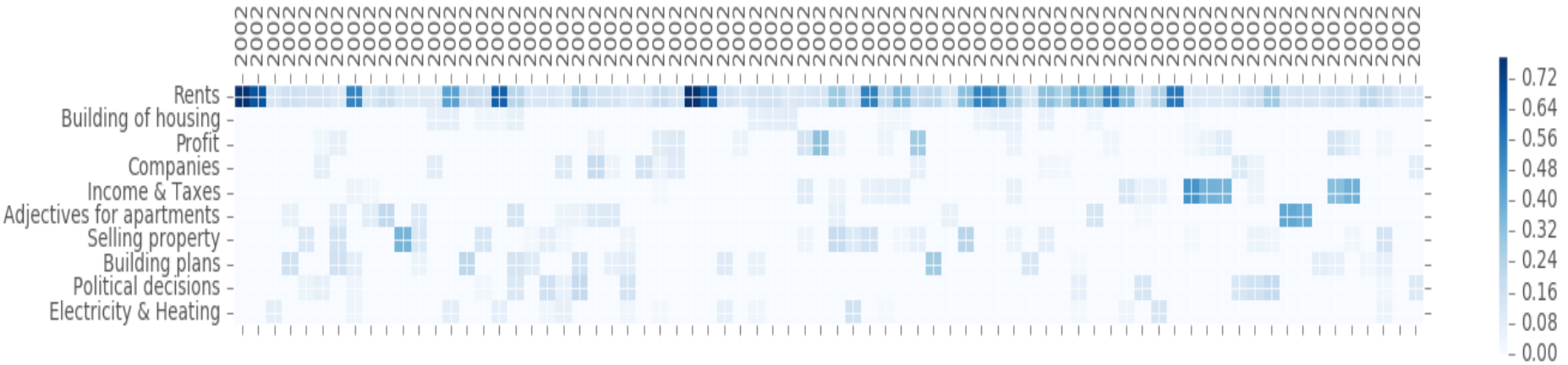


Figure 13: The occurrence of the rents topic together with the most frequently occurring topics in newstexts, first half of March, 2002. Only newstexts containing the rents topic are included. The color bar shows the proportion of topic.

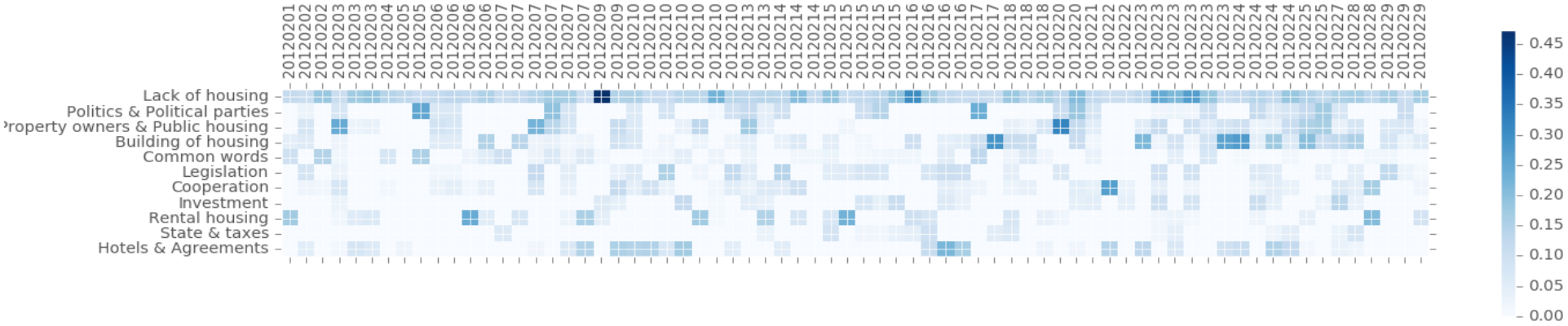


Figure 14: The occurrence of the lack of housing topic together with the most frequently occurring topics in newstexts, February, 2012. Only newstexts containing the lack of housing topic are included. The color bar shows the proportion of topic.

As with the Riksdag, an interactive plot was made. An example of how this looks is seen in figure 15. The full plot, but static, is seen in figure 16. This figure shows the 50 first newstexts in July, 2014. However, only newstexts containing the lack of housing topic (25) were selected, and therefore this topic is visible as a thick line. Naturally, the proportion of the lack of housing topic varies between newstexts, and this can be seen in the color nuances in the line. The most frequently cooccurring topics are seen further below, number 66 and 67. These are about politics and political parties (67) and property owners and public housing (66). Topic number 33 is also cooccurring for a few newstexts. This topic is about student housing, and one would assume this topic would be more cooccurring during the months before a new semester starts. This can be further investigated if we inspect figure 17. This figure shows the mean of each topic for every month during 2014. As with the earlier figures, only newstexts containing the lack of housing are used. The lack of housing topic is removed, to be able to see the other topics more clearly. Figure 18 shows the plot with the lack of housing topic included.

In figure 17, topic nr 33 (student housing) is slightly more cooccurring during July, August and September, as expected. This is more visible in this plot, as compared to the plot including the lack of housing topic (figure 18). The most frequent topics in figure 16 are frequent over the whole year. Topic number 67, which concerns political parties and politics have a strong peak in August. In September 2014, general elections were held in Sweden, and this could explain this peak. Other frequent topics are number 39 and 57. 39 is about investments and growth, and 57 are a topic of general words such as *said*. Another trend more visible in the plot without the lack of housing topic, is that topic nr 63 is more frequently cooccurring during the first two months of the year than the other months. This topic is about legislation and proposals.

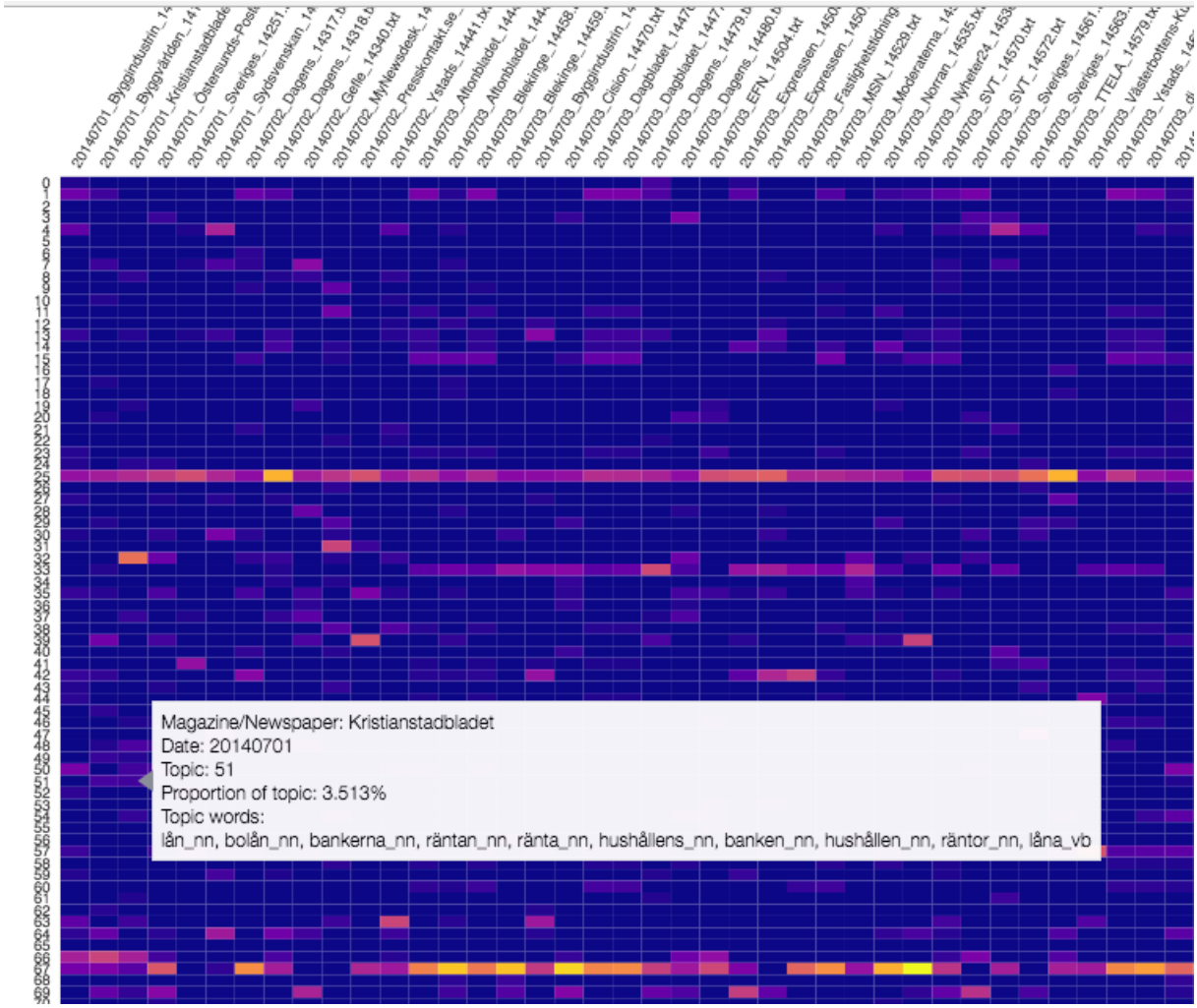


Figure 15: A screenshot of the interactive plot of newstexts.

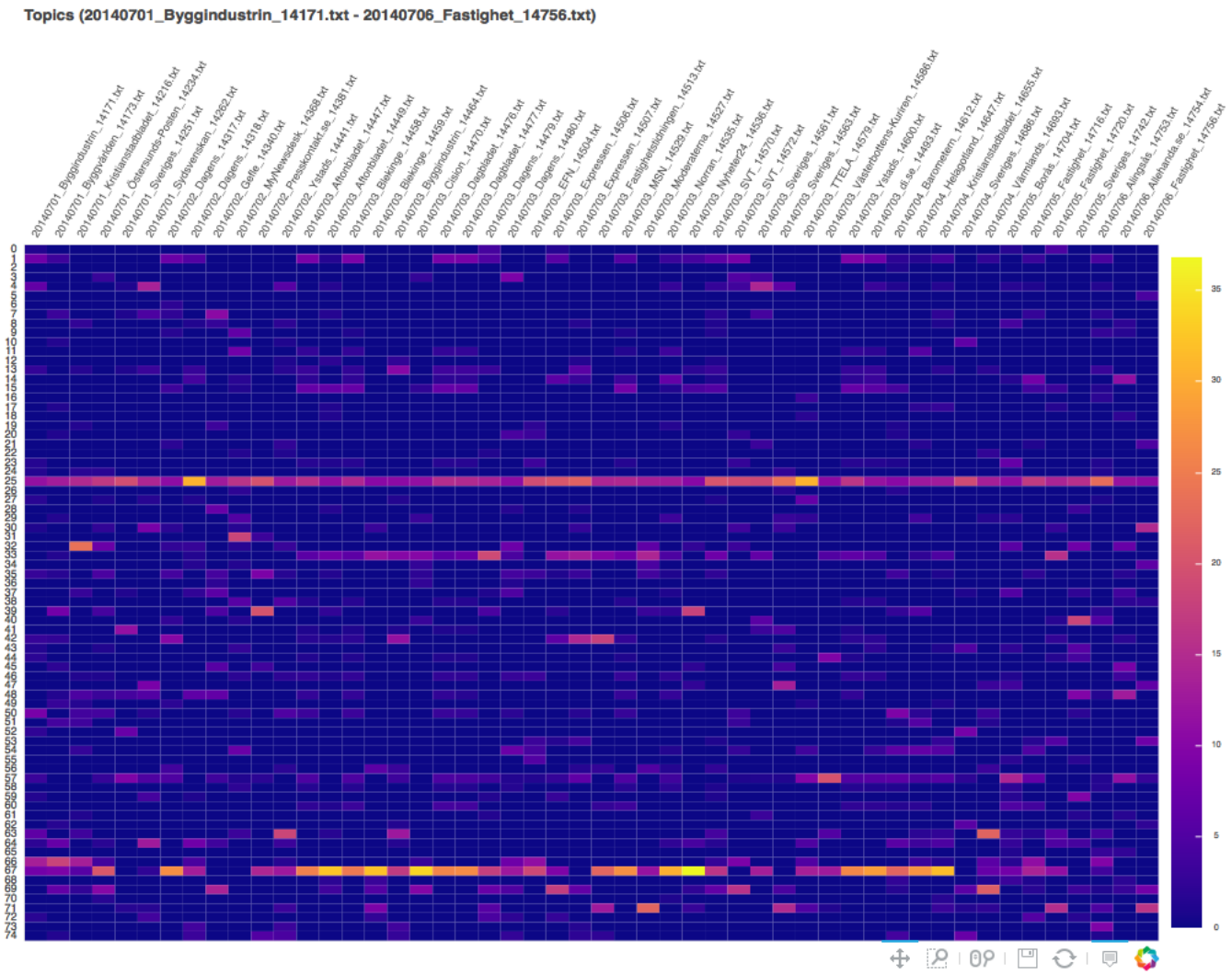


Figure 16: A plot over the 50 first newstexts in July, 2014, containing the lack of housing topic.

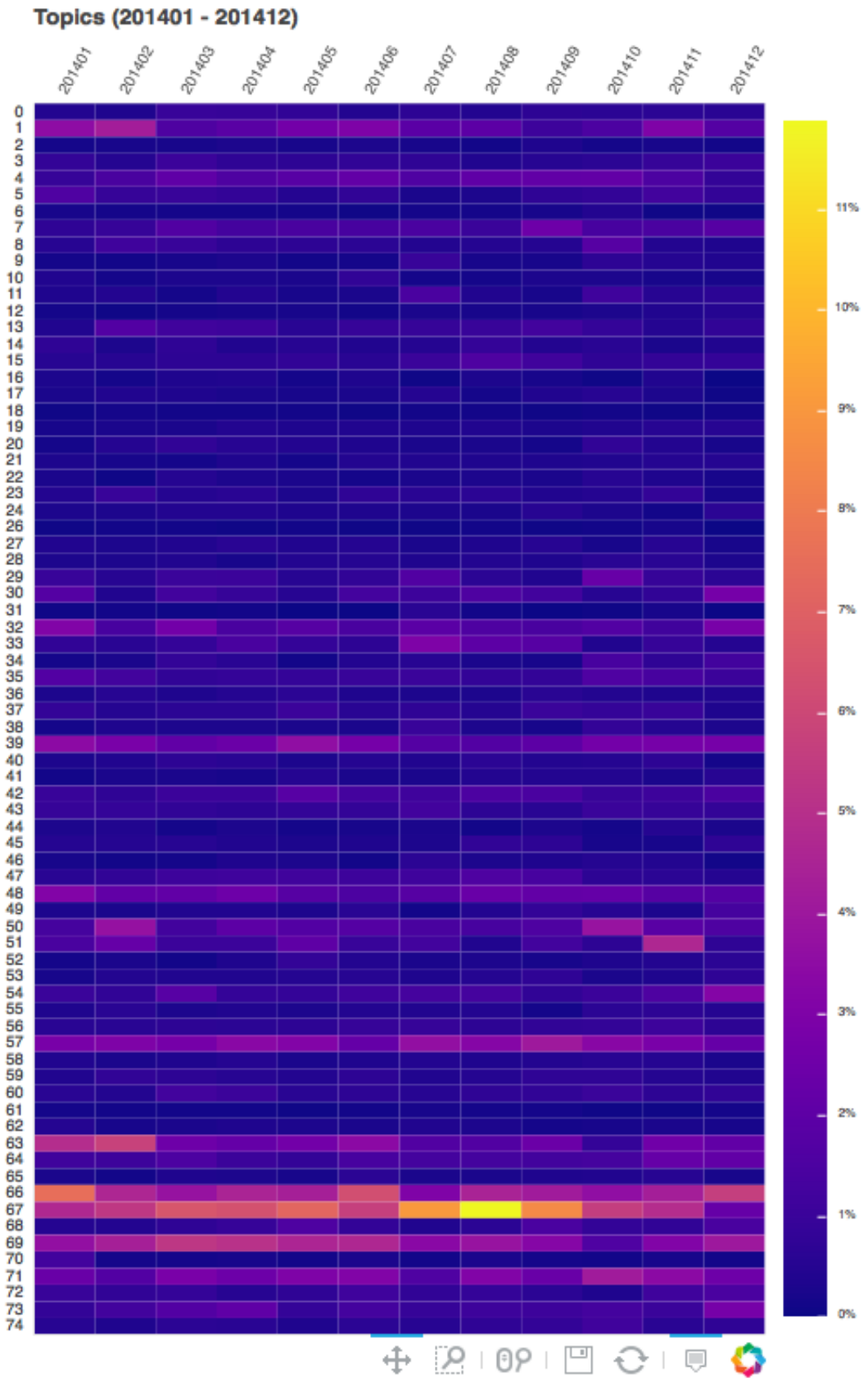


Figure 17: A plot over the mean of each topic for all the newstexts containing the lack of housing topic for each month during 2014. The housing topic is removed.

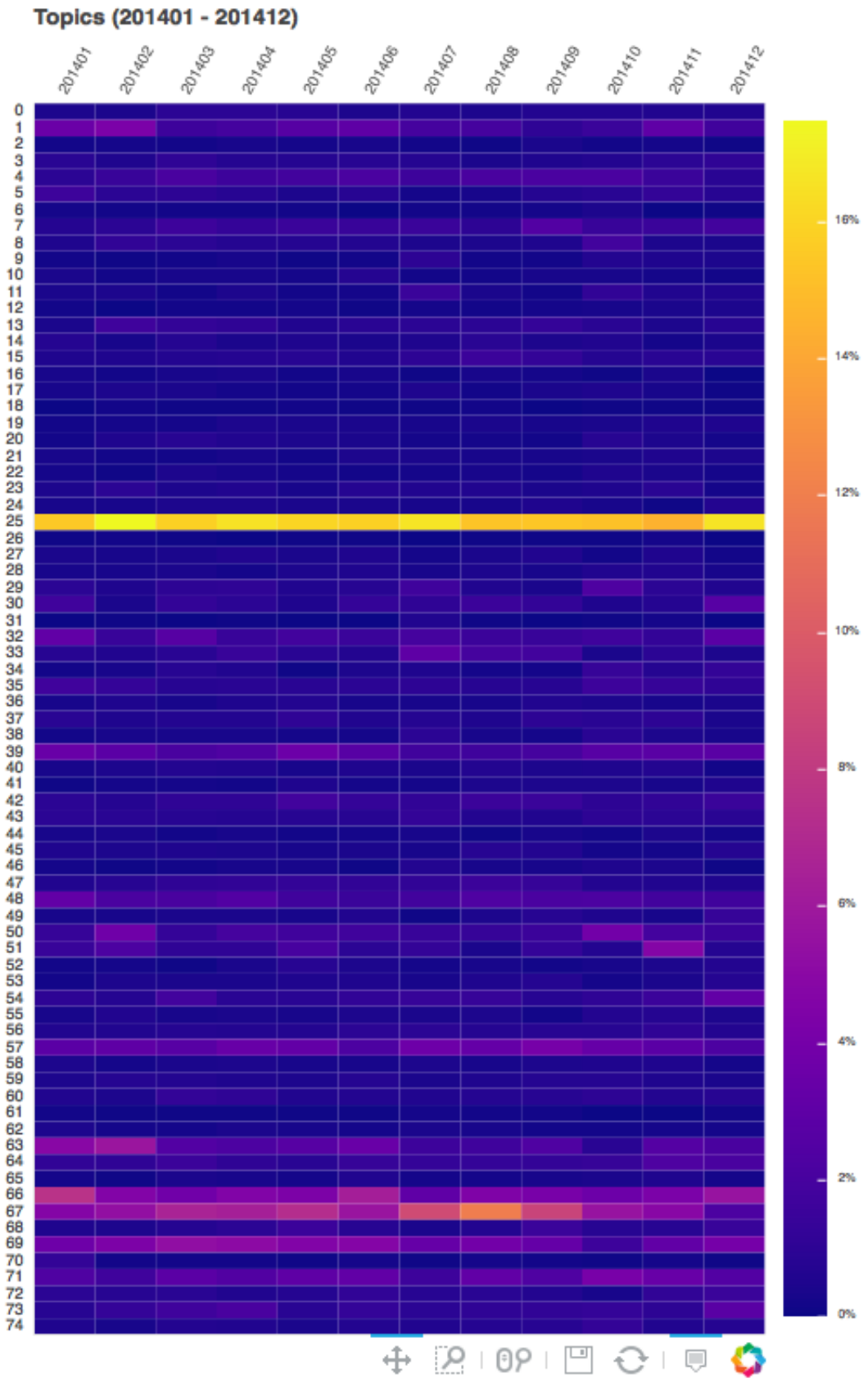


Figure 18: A plot over the mean of each topic for all the newstexts containing the lack of housing topic for each month during 2014. The housing topic is seen as the thick yellow line.

6 Discussion

This section discusses the results from the previous section.

6.1 Evaluation methods

Evaluating evaluation methods wasn't one of the original aims for this work, but by using both human judgement and computational methods this came naturally. Early in the process the perplexity measure was found to be not good, and this result replicates previous research (Chang et al., 2009). However, this measure is easily accessible in many toolkits, including the one used here. Possibly this measure could be used as a first step towards developing a good model.

The results of the two semantic coherence measures follow the results of Röder, the *cv* measure performs better than the *npmi*. The correlation values here are not as high as in their work, which reports a mean of 0.731, as compared to a mean of 0.68. But, this mean is much higher than the results of van der Zwaan et al. (2016), were *cv* is reported to have a correlation of 0.129. A difference in correlation should also be expected, the topic modeling has been carried out on different data sets, with different languages, preprocessing and filters. The expected effect of language specificity was not present.

For the newstexts, *cv* performed similarly to the Riksdag data. Models with nouns have the highest correlation, however two of the POS2 models have very low correlation. *cv* also ranks the models differently than the humans. These similarities between the models could indicate that the measure generalizes well between the two data sets.

The *cv* measure's ranking of the models differ from the human ratings, although with a small margin (0.57 vs 0.566) for the Riksdag. This suggests that while *cv* can be used as a guide for selecting topic models, one should also use complementary evaluation methods. It should also be noted that the correlation was the highest between the human rankings and both measures for models using only nouns.

The human judgement has been used as a gold standard here, but it is important to note that because there was a low number of evaluators the results are sensitive to small differences in the ratings. This might be the explanation for the lower results for the newstext data, where a new evaluator was employed. This could also be the reason for why the highest reported number of 3's was 27, when the highest possible number was 75. Note that the number of 3's is from the mean rating for every topic.

6.2 Effects of the linguistic filters and preprocessing

One of the two aims of this work was to examine the effect of different kinds of preprocessing. The results have been presented in previous sections, and here it will be discussed.

The highest rated models in the Riksdag data are trained only on nouns. This result is also partly found in the top models according to the semantic coherence. This follows the results of Jockers (2013), who reports better topics with only nouns, and the results of Martin & Johnson (2015), who report higher coherence for nouns.

Martin & Johnson (2015) also report higher coherence with lemmatized data. The highest rated model from the Riksdag does have lemmatized nouns. However the model with non-lemmatized nouns has no frequency filter. This will have had a negative effect on the ratings for that model, and may influence the comparison between the models. The highest rated model for the newstexts is also not from lemmatized data.

Looking at the effect of lemmatization overall, for the all POS group and the NN group, lemmatized models have higher ratings and coherence. In fact, the non-lemmatized all POS model was disregarded before the evaluation. However, as with the noun model above, this model also lacks frequency filter and stop list. This is probably the reason for the low ratings.

Comparing models based on the POS2 group with same filters, with the exception of lemmatization and no lemmatization, it can be observed that the later has higher ratings. This could indicate that non-lemmatized data is to be preferred, contrary to what was initially assumed here. This initial expectation was due to the fact that Swedish has a richer morphology than English and therefore the results was expected to improve with lemmatization. However, this effect might not be present.

But, lemmatization can't be disregarded completely. By observing the topics from non lemmatized and lemmatized models, it was found that models without lemmatization often had repetition of the same word in the top words. The evaluators were asked how understandable a topic was, and the same word but in different forms would presumably be interpreted as understandable. This might be the reason why non-lemmatized data is rated higher. Furthermore, understandable is not the same as informative, which perhaps the evaluators should have been asked for instead, although this criteria is harder to assess without knowing the task at hand. This is also more subjective. To conclude, the effect of lemmatization is unclear.

Repeating what is stated above, the highest rated model in the Riksdag is trained on only nouns, despite that they are lemmatized. The highest rated model from the newstexts is based on the POS2 group, non lemmatized, but with the lemmatized nouns as number two. This indicates that part of speech has more influence on the ratings of the topics than lemmatization. In all the inspected topics from all the models, the most frequent part of speech was nouns, so it is not surprising that models trained on nouns have high ratings.

The dependency relations filter had a very obvious effect, it worsened both the ratings and the coherence measures. This is a bit surprising, because most of the selected words were nouns, which improved the results for the other models. The choice of dependency relations could be the cause of this, or using dependencies for the selection might not suit the LDA model. There is also one exception, the POS2 model with a stop list had improved ratings after the dependency relations filter was added.

In general, in order to draw more conclusions, more tests need to be done, both on other dependency relations then the ones tested here and on the NN and the All POS group. Both of these groups needs to be tested with frequency filters but without lemmatization, in order to further examine the effect of lemmatization.

Using a frequency filter and stop list almost always improved the result for both data sets, which is to be expected. There are two exceptions, for the newstext data the nouns with a lower frequency filter was ranked higher than the one with a higher frequency filter. This could be the frequency filter removing too many words. The other example is POS2 with dependency relations, from the Riksdag, which has a higher result without a frequency list than with it. However, with a stop list and frequency list the models have the higher rankings. This could also be due to the frequency list removing important word, but leaving words the stop list would then remove.

The stop list for the Riksdag data was based on the period 2010–2014, and then used for all the periods. This is not ideal, because the time span is quite large and the vocabulary surely will have changed over time. However, the stop list is deemed to generalize ok, because there were good topics from all of the periods. This might be because of the formality of the Riksdags language, possibly the vocabulary changes slower and is more fixed as compared to other domains due to this.

For the newstext data using frequency only was deemed good enough. The difference in results, and the difference in the ranking of the models for the two data sets, highlight the need for different preprocessing for different domains. The newstext domain presumably has a less formal language, with fewer domain specific words than the Riksdag and this might be the reason a frequency filter is suitable instead of a stop list. That the style of the newstext language is closer to spoken language might also be the reason. However, when inspecting the topics from the newstext, at least one topic was found containing general words such as *says*. There might be a need to use a stop list for this domain too.

As stated above, the order of the top five models differ a lot between the two data sets, and the highest rated model is different between the data sets. The newstexts models had lower ratings overall. This indicates that the five highest rated combination of filters from the Riksdag data might not generalize well to the newstexts, however the highest rated model for the newstexts was still found to be useful.

As stated in the background section, many papers about topic modeling don't explain or elaborate on which preprocessing methods have been used. The aim here has been to systematically explore the effects of preprocessing, and this becomes even more important if different domains require different preprocessing and even different linguistically informed preprocessing, as suggested by the results here.

6.3 The chosen model and topics

The fact that there is a good LDA model for the Riksdag contradicts the results of Hägglöf (2014) who found LDA not be a suitable model for the same data. But, one can always discuss how good the chosen model is. In the Background section, the robustness of topic models was briefly mentioned. Here there has been no thorough investigation of robustness, but through the manual inspection by the author the same topics were found to recur. But, this is of course something that needs to be further investigated.

As discussed in the previous section, the evaluators were asked to rate according to understandability. But how understandable, or interpretable, are the final models? For the Riksdag, the model with only nouns were chosen. Jockers (2013) argues that if one wants to explore sentiment, adjectives should be included in the analysis. However, in most of the models here nouns were the dominant part of speech in the topics, regardless of which POS filter was used. For the newstext data a model from the POS2 group was chosen as the best model, but with nouns as the second highest rated model. This means that even if one wants to explore sentiment through adjectives, simply including them in the data might not be enough.

Both of the two issues above could be further investigated, as could the many other parameters involved in topic modeling. However one must also focus on the usefulness of the model, and what is needed from it. Inspecting the data sets with the help of the models is indeed feasible, as demonstrated in section 5.3. As mentioned in the evaluation of topic models, Jockers (2013) discusses if one can trust a model with not only good topics. He comes to the conclusion that one model cannot capture everything, and that there is no reason not use it if it describes the topics one needs. Of course, a good topic model is preferred over a mediocre one. In the same spirit, continuing with the lens analogy fromDiMaggio et al. (2013), here a model is found that can describe what was searched for. To conclude, the chosen models here serve their purpose, but there is always room for improvement.

The labeling of the topics was as previously stated done by the author. To label topics can be subjective, but in order to interpret the topics this was needed. However, in the interactive plots the topics are also presented with the words from the topics instead of labels. This gives the reader the opportunity to by himself/herself explore the results and topics.

6.4 Data analysis

One of the aims of this thesis was to show how one could use language technology for analysis. In section 5.3 this has been exemplified. There is a lot left to do for a full analysis, and most of the data is ready to be examined further.

It has been shown that the chosen method, LDA topic modeling, is suitable for identifying and following trends in the chosen corpora. This was slightly more successful in the newstext data, possibly because of the less formal language, although the topics were rated higher in the Riksdag data. That the trends were harder to follow in the Riksdag might be because the data wasn't filtered like the newstexts. In the newstext data, a filter had been applied before the preprocessing, for selecting documents, and thus more relevant document were found.

The topics found in the documents of Riksdag were more diverse than the newstexts. In all the inspected documents from the riksdag there seems to be fewer cooccurring topics and this will also have had en effect on the following of trends. Although this must be further investigated.

In retrospect, one could have wished for some of the topics could have been more narrow in their scope for a more precise analysis. Also, despite thorough preprocessing, there were still a few topics full of general words. There were also some topics about very general subjects. For the Riksdag data, this could be because the topic modeling was done on all the document types together. However, the topics serve their purpose in identifying documents. The next step in the analysis could be to do topic modeling on only these, in order to find more nuances in the data.

7 Conclusions and future work

7.1 Conclusions

The first aim of this work was to exemplify how topic modeling could be used for analysis of public discourse, by investigating how the topics labeled 'housing policies' have varied over time and what is has cooccurred with. In section 5.3 this was shown, and the conclusion is that the topic modeling method LDA can be used for investigation of public discourse regarding housing policies.

The second aim was to investigate how topic modeling is affected by different preprocessing, based on linguistic information. Of the three categories investigated here, part of speech had the largest impact on the results. Using nouns improved the topics. Models based on verb, adjectives, participles and nouns also improved the topics, however the most frequent part of speech in these models is nouns.

Lemmatization data is not as good as non-lemmatized data, however without lemmatization the same words are repeated in the topics. This might have an effect on the topics usefulness and interpretability and it is thus unclear if non-lemmatized data is preferred. Using data selected based on dependency relations does not result in good topics, however this might change if one uses different dependency relations.

The evaluation of the topic models showed that the *cv* measure has a better correlation with human judgements than the *npmi* measure. Both of the measures has the highest correlation for models using only nouns.

7.2 Future work

There are many directions in which the results of this work could be taken further.

First and foremost, the results of the topic modeling and the classifications need to be further investigated for a full analysis. For example, the full time span could be considered and different periods could be compared against each other. The results could also be incorporated in some corpus exploration tool such as Korp. The interactive visualization could in these cases have a link to the specific documents for further exploration. The interactive plot could also be used as a means for evaluation, at least for an initial manual evaluation.

Regarding evaluation, the evaluation of the coherence measures could also be further explored. More measures need to be compared, as only two such measures were compared here. The effect of different reference corpora for the word statistics for the coherence measures is something that needs to be evaluated as well. Here, the same corpora was used both for topic modeling and for the coherence measures.

The topic modeling in itself could also be further explored. For example, the robustness of the model could be checked by labeling topics and documents. There is more work to be done when using a stop list, or trying to avoid using one. Different number of frequent words could be tested, to find the optimal number of words to avoid.

There is more linguistic information to use, for example word sense or named entity recognition, which are both already available through the annotation. Other possible extensions could be to split compounds into parts, or using bigrams or multiword expressions. The information used here also needs to be further investigated, especially the effect of both lemmatization and dependency parsing.

References

Blei, D., Carin, L., & Dunson, D. (2010). Probabilistic topic models. *IEEE signal processing magazine*, 27(6), 55–65.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84.

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

Borin, L., Forsberg, M., Hammarstedt, M., Rosén, D., Schäfer, R., & Schumacher, A. (2016). Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *SLTC 2016. The Sixth Swedish Language Technology Conference, Umeå University, 17-18 November, 2016*.

Borin, L., Forsberg, M., & Lönngren, L. (2013). SALDO: a touch of yin to WordNet's yang. *Language resources and evaluation*, 47(4), 1191–1211.

Borin, L., Forsberg, M., & Roxendal, J. (2012). Korp-the corpus infrastructure of Språkbanken. In *LREC* (pp. 474–478).

Boverket (2017). Bostadsmarknadsenkäten 2017. [Online; accessed 10-May-2017, `http://www.boverket.se/sv/samhallsplanering/bostadsplanering/bostadsmarknaden/bostadsmarknadsenkaten-i-korthet/`].

Boyd-Graber, J. L. & Blei, D. M. (2009). Syntactic topic models. In *Advances in neural information processing systems* (pp. 185–192).

Bäck, H., Erlingsson, G., & Larsson, T. (2015). *Den svenska politiken*. 4 edition.

Carron-Arthur, B., Reynolds, J., Bennett, K., Bennett, A., & Griffiths, K. M. (2016). What's all the talk about? Topic modelling in a mental health Internet support group. *BMC psychiatry*, 16(1), 367.

Carter, D. J., Brown, J., & Rahmani, A. (2016). Reading the high court at a distance: Topic modelling the legal subject matter and judicial activity of the high court of Australia, 1903-2015. *UNSWLJ*, 39, 1300.

Chang, J., Boyd-Graber, J. L., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Nips*, volume 31 (pp. 1–9).

Dahllöf, M. (2012). Automatic prediction of gender, political affiliation, and age in Swedish politicians from the wording of their speeches—A comparative study of classifiability. *Literary and Linguistic Computing*, 27(2), 139–153.

Darling, W. M., Paul, M. J., & Song, F. (2012). Unsupervised part-of-speech tagging in noisy and esoteric domains with a syntactic-semantic bayesian hmm. In *Proceedings of the Workshop on Semantic Analysis in Social Media* (pp. 1–9).: Association for Computational Linguistics.

Delpisheh, E. & An, A. (2014). Topic modeling using collapsed typed dependency relations. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 146–161).: Springer, Cham.

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics*, 41(6), 570–606.

Fang, Y., Si, L., Somasundaram, N., & Yu, Z. (2012). Mining contrastive opinions on political texts using cross-perspective topic model. In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 63–72).: ACM.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. In *Advances in neural information processing systems* (pp. 537–544).

Guo, S. (2012). *Using Dependency Parses to Augment Feature Construction for Text Mining*. Virginia Polytechnic Institute and State University.

Hägglöf, H. (2014). Automatic organization of online Swedish political discourse. Master thesis, Uppsala university.

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 363–371).: Association for Computational Linguistics.

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *advances in neural information processing systems* (pp. 856–864).

Höjer, H. (2017). Därför kan byggboomen inte lösta bostadskrisen. *Forskning och Framsteg*, (2), 61–74.

Jacobi, C., van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 4(1), 89–106.

Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press. pp: 128-133.

Jurafsky, D. & Martin, J. H. (2009). *Speech and language processing*, volume 2. Pearson London. pp: 448-450.

Kuang, D., Brantingham, P. J., & Bertozzi, A. L. (2017). Crime Topic Modeling. *arXiv preprint arXiv:1701.01505*.

Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *EACL* (pp. 530–539).

Martin, F. & Johnson, M. (2015). More efficient topic modelling through a noun only approach. In *Australasian Language Technology Association Workshop 2015* (pp. 111).

McFarland, D. A., Ramage, D., Chuang, J., Heer, J., Manning, C. D., & Jurafsky, D. (2013). Differentiating language usage through topic models. *Poetics*, 41(6), 607–625.

Mohr, J. W. & Bogdanov, P. (2013). Introduction—Topic models: What they are and why they matter. *Poetics*, 41(6), 545–569.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 100–108).: Association for Computational Linguistics.

Regeringen (2017). How sweden is governed. [Online; accessed 06-June-2017, `http://www.government.se/how-sweden-is-governed`].

Rehurek, R. & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*: Citeseer.

Riksdagen (2017a). Debates and decisions in the chamber. [Online; accessed 06-June-2017, `http://www.riksdagen.se/en/how-the-riksdag-works/the-work-of-the-riksdag/debates-and-decisions-in-the-chamber`].

Riksdagen (2017b). Elections to the Riksdag. [Online; accessed 06-June-2017, http://www.riksdagen.se/en/how-the-riksdag-works/democracy/elections-to-the-riksdag].

Riksdagen (2017c). Forming a Government. [Online; accessed 06-June-2017, http://www.riksdagen.se/en/how-the-riksdag-works/democracy/forming-a-government].

Riksdagen (2017d). Riksdagens historia. [Online; accessed 06-June-2017, http://www.riksdagen.se/sv/sa-funkar-riksdagen/demokrati/riksdagens-historia].

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399–408).: ACM.

SCB (2017a). Framtidens befolkning. [Online; accessed 10-May-2017, http://www.scb.se/hitta-statistik/sverige-i-siffror/manniskorna-i-sverige/framtidens-befolkning/].

SCB (2017b). Sveriges folkmängd från 1749 och fram till idag. [Online; accessed 10-May-2017,http://www.scb.se/hitta-statistik/sverige-i-siffror/manniskorna-i-sverige/befolkningsutveckling].

Tahmasebi, N., Borin, L., Capannini, G., Dubhashi, D., Exner, P., Forsberg, M., Gossen, G., Johansson, F. D., Johansson, R., Kågebäck, M., et al. (2015). Visions and open challenges for a knowledge-based culturomics. *International Journal on Digital Libraries*, 15(2-4), 169–187.

Tangherlini, T. R. & Leonard, P. (2013). Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research. *Poetics*, 41(6), 725–749.

van der Zwaan, J. M., Marx, M., & Kamps, J. (2016). Topic coherence for dutch. [Online; accessed 06-June-2017, https://zenodo.org/record/46377/files/case_study.pdf].

Viklund, J. & Borin, L. (2016). How Can Big Data Help Us Study Rhetorical History? In *Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland*, number 123 (pp. 79–93).: Linköping University Electronic Press.

Wilkerson, J. D. & Casas, A. (2017). Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges. *Annual Review of Political Science*, 20(1).

A Appendix A - Categories of Riksdagens öppna data

Document	Description	Span
Betänkande	Committee reports with proposals for decisions in the Riksdag.	1971-present
Departementserien	Investigations made by the Ministries of the Government.	1990-present
EUN	Documents from the proceedings of the committee on European Union affairs	1990–present
Föredragningslista	Agendas of the Chambers meetings.	2000–present
Faktapromemoria	Explanatory memoranda on EU-proposals	2000–present
Framställning/redogörelse	Submissions and reports by bodies appointed by the Riksdag. Submissions are proposals for legislation, reports are annual reports from the bodies.	1971–present
Interpellation	Written questions from a member of the Riksdag to a Government minister.	1998–present
Kammaraktiviteter	Lists of the proceedings/activities in the Chamber.	2010–present
KOM	Submissions and proposals of the European Commission.	1990–present
Motion	Proposals for legislation from a member of the Riksdag.	1971–present
Proposition	Proposals for legislation from the Government.	1971–present
Protokoll	Records of the proceedings/meetings in the Chamber.	1971–present
Riksdagsskrivelse	Documents informing to the Government of the decisions the Riksdag has taken in different parliamentary matters.	2000–present
Sammanträden	Lists of the meetings in the Chamber	2010–present
Skriftliga frågor	Written questions from a member of the Riksdag to a Government minister. Answers to the questions are also included in this category.	1998–present
Statens offentliga utredningar	Reports from committees of inquiry appointed by the Government, in preparation for submitting a proposal.	1990–present

Talarlista	List of speakers from the Chambers meetings.	2000–present
Utredningar	Committee terms of reference and reports from the Government-appointed committees.	1961–present
Utskottsdokument	Documents from the committees including agendas, speakerlists and annual reports.	1971–present
Yttrande	Opinions on matters to the responsible committee from another committee	1971–present
Övrigt	Miscellaneous documents including reports from the Riksdag, reports from the National Audit Office and old investigations.	1971–present

Source: http://www.riksdagen.se/en/documents-and-laws/ & http://www.riksdagen.se/sv/sa-funkar-riksdagen/arbetet-i-riksdagen/dokumenttyper-i-riksdagen/

B Appendix B - List of Newspapers and Magazines

Magazine/Newspaper	Number of newstexts
Helsingborgs Dagblad	10242
Dagens Nyheter	9475
Göteborgs-Posten	9467
Svenska Dagbladet	7456
Sydsvenskan	6294
Upsala Nya Tidning	6252
Aftonbladet	6020
MyNewsdesk	5548
Östgöta Correspondenten	5235
Webfinanser	4918
Arbetarbladet	4621
Expressen	4337
Hem -och- Hyra	4187
Dalademokraten	4072
Helagotland	3957
Sydsvenska Dagbladet	3878
Norrköpings Tidningar	3865
Skånska Dagbladet	3803
Hallands Nyheter	3632
Norrbottens-Kuriren	3614
Värmlands Folkblad	3598
Sundsvalls Tidning	3234
Ystads Allehanda	3187
Nerikes Allehanda	3049
Vestmanlands Läns Tidning	3035
Hallandsposten	3023
Gefle Dagblad	3014
Barometern	3002
Borås Tidning	2897
Dalarnas Tidningar	2895
Östersunds-Posten	2846
Nya Wermlands-Tidningen	2822
Kristianstadbladet	2732
di.se	2724
Dagens Industri	2716
Länstidningen Östersund	2671
Blekinge Läns Tidning	2655
Smålandsposten	2556
Fastighetstidningen	2507
Norrländska Socialdemokraten	2494
Affärsvärlden	2334
Eskilstuna-Kuriren	2316
Fastighet -och- Bostadsrätt	2315
TTELA	2276
Folkbladet	2228
Newsdesk	2207
Cision Wire	2134
Västerviks Tidningen	1944

Magazine/Newspaper	Number of newstexts
Södermanlands Nyheter	1892
Sydöstran	1883
Riksdagen	1860
FastighetsSverige	1854
Piteå-Tidningen	1849
Vår Bostad	1751
Dagbladet Sundsvall	1747
Motala -och- Vadstena Tidning	1747
Trelleborgs Allehanda	1721
GT	1711
Newsdesk pressmeddelanden	1640
Folket	1619
Norra Skåne	1560
SVT Nyheter	1533
Västerbottens-Kuriren	1522
Länstidningen Södertälje	1415
Östran-Nyheterna	1394
Privata Affärer	1382
Bohusläningen	1339
Aktietips	1330
Byggindustrin	1266
Hela Hälsingland	1256
TV4	1227
Fastighetsnytt	1222
Byggnyheter	1201
Västerbottens Folkblad	1162
Sveriges Radio Stockholm	1147
Metro	1139
E24	1111
Lokaltidningen Mitt i	1107
Tidningen Fastighetsaktien	1077
Kvällsposten	1060
Nya Ludvika Tidning	1050
Allehanda.se	1045
SVT ABC	1043
Katrineholms-Kuriren	1036
Stockholm City	1034
Eskilstuna Kuriren	1001
Byggkontakt	985
Norran	982
Norra Västerbotten	969
Avanza	955
Norrländska socialdemokraten	949
Sveriges Radio Ekot	938
ETC	929
Östran	919
Norrtelje Tidning	898
Affärsvärlden Delv lösenord	884
Fastighetsvärlden	856
Sveriges Radio Malmö	837

Magazine/Newspaper	Number of newstexts
Nacka Värmdöposten	815
Sveriges Radio Sörmland	815

C Appendix C - List of parliamentary periods in the Riksdag

- 1970–1973
- 1973–1976
- 1976–1979
- 1979–1982
- 1982–1985
- 1985–1988
- 1988–1991
- 1991–1994
- 1994–1998
- 1998–2002
- 2002–2006
- 2006–2010
- 2010–2014
- 2014–2018

D Appendix D - Search terms for newspapers and magazines

affordable housing
andrahandshyra
andra hand
andrahandskontrakt
bolån
bolåneränta
boverket
brf
bruksvärde
byggnorm
bygga
detaljplan
fastighetskatt
fastighetsskatt
fastighetsägarföreningen
fastighetsägarna
flyttskatt
förort*
första hand
förstahandskontrakt
gentrifiering
hyresgäst*
hyreskontrakt
hyresreglering
hyresrätt
innerstad*
kontantinsats
lägenhet
marknadshyr*
plan och byggnadslagen
presumptionshyra
rot
rut
ränteavdrag
rörlighet
segregation
segregerade områden
social housing
studentbostäder
sverigeförhandlingen
trångboddhet
villamatta
ytterområde*

E Appendix E - Part of speech and dependency relations tags

Part of speech tag	Meaning
Tag	Meaning
AB	Adverb
DT	Determiner
HA	WH-adverb
HD	WH-determiner
HP	WH-pronoun
HS	WH-possessive
IE	Infinitival marker
IN	Interjection
JJ	Adjective
KN	Coordinating conjunction
NN	Noun
PC	Participle
PL	Particle
PM	Proper Noun
PN	Pronoun
PP	Preposition
PS	Possessive pronoun
RG	Cardinal number
RO	Ordinal number
SN	Subordinating conjunction
VB	Verb
	Supplementary Tag
UO	Foreign word
	Delimiters
MAD	Major delimiter
MID	Minor delimiter
PAD	Pairwise delimiter

Dependency tag	Meaning
++	Coordinating conjunction

+A	Conjunctional adverbial
+F	Coordination at main clause level
AA	Other adverbial
AG	Agent
AN	Apposition
AT	Nominal (adjectival) pre-modifier
CA	Contrastive adverbial
DB	Doubled function
DT	Determiner
EF	Relative clause in cleft
EO	Logical object
ES	Logical subject
ET	Other nominal post-modifier
FO	Dummy object
FP	Free subjective predicative complement
FS	Dummy subject
FV	Finite predicate verb
I?	Question mark
IC	Quotation mark
IG	Other punctuation mark
IK	Comma
IM	Infinitive marker
IO	Indirect object
IP	Period
IQ	Colon
IR	Parenthesis
IS	Semicolon
IT	Dash
IU	Exclamation mark
IV	Nonfinite verb
JC	Second quotation mark
JG	Second (other) punctuation mark
JR	Second parenthesis
JT	Second dash
KA	Comparative adverbial
MA	Attitude adverbial
MS	Macrosyntagm
NA	Negation adverbial
OA	Object adverbial
OO	Direct object
OP	Object predicative
PL	Verb particle
PR	Preposition
PT	Predicative attribute
RA	Place adverbial

Dependency tag	Meaning
SP	Subjective predicative complement
SS	Other subject
TA	Time adverbial
TT	Address phrase
UK	Subordinating conjunction
VA	Notifying adverbial
VO	Infinitive object complement
VS	Infinitive subject complement
XA	Expressions like ”så att säga” (so to speak)
XF	Fundament phrase
XT	Expressions like ”så kallad” (so called)
XX	Unclassifiable grammatical function
YY	Interjection phrase
	New Categories
CJ	Conjunct (in coordinate structure)
HD	Head
IF	Infinitive verb phrase minus infinitive marker
PA	Complement of preposition
UA	Subordinate clause minus subordinating conjunction
VG	Verb group

F Appendix F - Examples of classified documents

Swedish original of the *motion* classified with the 'housing policies' topic.

> Motion till riksdagen2010 / 11:C264av Ann-Christin Ahlberg m.fl. (S) Byggandet av billiga hyresrätter s34010 Förslag till riksdagsbeslut Riksdagen tillkännager för regeringen som sin mening vad som anförs i motionen om stimulanser för produktionen av hyresrätter . Motivering Regeringen Reinfeldt har slopat de bostadspolitiska målen som tidigare fanns . Detta gjordes samtidigt som de slopade ränte- och investeringsstöden vid nybyggnation av hyresrätter . Resultatet har givit förödande konsekvenser för byggandet av hyresrätter . Den förda politiken har misslyckats . Det som nu behövs är en politik som värnar om hyresrätten och stimulerar byggandet av hyresrätter till en rimlig kostnad . En sådan politik består i att införa ett investeringsstöd för hyresrätter samt en översyn av neutraliteten på bostadsmarknaden . Stockholm den 26 oktober 2010 Ann-Christin Ahlberg (S) Hans Olsson (S) Phia Andersson (S)

Swedish original of the written question used as an example on page 30.

> den 7 september / Svar på fråga / 2009 / 10:999 Den svenska surströmmingen / Jordbruksminister Eskil Erlandsson / Peter Hultqvist har frågat mig vilka åtgärder jag tänker vidta med anledning av oron för surströmmingens framtid . Frågeställningen har sin upprinnelse i det tillfälliga undantag Sverige har från EU:s regler om gränsvärden för vissa miljöföroreningar i viss fisk från Östersjöområdet som upphör den 31 december 2011 / Sverige har tillsammans med Finland ett tidsbegränsat undantag från de regler som gäller inom EU för tillåtna högsta halter av dioxiner och PCB i fisk , undantaget gäller bland annat strömming från Östersjön . Undantaget baseras till stor del på att Sverige har ett väl fungerande system för att skydda folkhälsan , genom information riktad till konsumenterna om att konsumtionen av vissa fiskarter från Östersjöområdet kan behöva begränsas . / Surströmmingen är en del av vårt matarv och en viktig tradition i stora delar av Sverige , och det vill jag värna . Därför har jag personligen haft kontakt med kommissionären för hälso- och konsumentfrågor John Dalli och berättat om hur viktig den svenska traditionen kring surströmming är . Konsumtionen är dessutom vanligtvis begränsad såväl geografiskt som tidsmässigt / Skulle undantaget upphöra och man inte finner någon annan lösning så innebär det också med all sannolikhet att den övervägande delen av det småskaliga fisket längs ostkusten skulle slås ut . Det vill jag inte medverka till . I stället vill jag värna surströmmingen , och om våra experter på myndigheterna gör bedömningen att man genom nuvarande kostråd och andra åtgärder säkerställer folkhälsan kommer jag att verka för en lösning som möjliggör fortsatt fiske . / Inför den kommande förhandlingen om ett fortsatt undantag har regeringen givit Livsmedelsverket och Fiskeriverket i uppdrag att , efter samråd med Naturvårdsverket , utreda vilka handlingsalternativ som finns och vilka konsekvenserna blir av de olika handlingsalternativen . Myndigheterna ska lämna sin slutrapport senast den 1 mars 2011 . / Sist men inte minst vill jag betona att för att få bukt med problemet med miljöföroreningar i fisk från Östersjön måste tillförseln av dessa föroreningar till Östersjön begränsas . Sverige arbetar aktivt med detta såväl nationellt som inom EU och globalt . /

Swedish original of an newstext classified with the lack of housing-topic.

> Farlig tystnad inom bostadspolitiken Signerat Karlsson . I nästan av hälften av landets 290 kommuner råder det bostadsbrist . Det visar en rapport som LO tagit fram . Störst är bristen

på bostäder i och omkring landets storstäder . Bostadsbristen i Sverige är ingen nyhet.Men vad som ska göras för att råda bot på det låga byggandet råder det delade meningar om . LO förespråkar föga överraskande statliga stimulanser . Med 15 miljarder kronor i direkt stöd till nya hyresrätter och studentboende vill LO lösa bostadskrisen där behoven är som störst . Bostadsbyggande ska inte dopas med statliga medel . För att sätta fart på byggandet måste det istället bli mer lönsamt , dels genom att byggkraven förenklas och standardiseras , dels genom att hyresregleringen tas bort . Stefan Attefall (KD) har under sin tid som ansvarig bostadsminister dragit igång en rad utredningar . Dessa har landat i konkreta förslag , framför allt vad gäller förenklingar av byggprocessen och att skapa enhetliga byggkrav . Det är troligt att en ny lagstiftning är på plats inom kort . Däremot finns det i dag ingen lösning på det kanske största problemet , bristen på hyreslägenheter . Dagens hyresreglering motarbetar sitt syfte att bryta bostadssegregering då den istället skapar en annan uppdelning av befolkningen , de som har en egen bostad och de som inte har . Framför allt drabbar detta unga människor som ska ta de första kliven in på bostadsmarknaden . För att minska det stora utanförskapet på den svenska bostadsmarknaden måste trösklarna sänkas . Det görs med enklare och billigare hyreslägenheter och studentboenden och genom att hyresregleringen fasas ut . Tyvärr är det allt för tyst inom politikens område , bostadsfrågor är inte hett . Men får man inte fart på det svenska bostadsbyggandet är riskerna för framtiden betydande med en dysfunktionell arbetsmarknad och en tillväxt som hämmas . Mattias Karlsson,Kristianstadsbladet,20131005

Swedish original of the newstext used as an example on page 34.

Deras hem slukades av storbranden - Göteborg Renovering av hela huset Decembergatan 35-53 pågår . Inflyttning är planerad till oktober 2016 för de första , januari 2017 för de sista . De som flyttar tillbaka kommer till ett totalrenoverat hus med nytt allt . Fönster , kök , badrum , allt.Flera brister orsak till branden Men alla som evakuerades efter branden är inte säkra på att de vill återvända . Det tog ett antal telefonsamtal innan vi stötte på en som var helt säker på vad hon vill . - Jag vill återvända för mitt barns skull . Han har just börjat skolan , säger Malin Forsell . Sonen Liam har kunnat gå kvar i sin skola då de inte måste flytta särskilt långt bort . Bostadsbolaget Poseidon säger att hänsyn tagits till befintliga platser på dagis och i skola när ersättningslägenheter fördelats . Barnfamiljer har i allmänhet fått vara kvar i Kortedala , merparten av övriga har hamnat i Biskopsgården . Bostäder i andra delar av stan förekommer också . - Jag kan mycket väl förstå de som inte vill återvända , säger Malin Forsell . Branden har lämnat mer eller mindre djupa sår hos alla hyresgäster som blev evakuerade ur sina lägenheter natten mellan den 10 och 11 augusti . Malin fick vänta en dryg vecka innan hon tilläts återvända till sin lägenhet . - Det var en chock . Vårt sovrum var borta . Allt som återstod var ett galler , säger hon . Ett par månader senare erbjöds en andra chans att återse bostaden . - Allt stod kvar . Jag var där för att se efter en del gamla saker , inte minst bilder . De hade klarat sig , men luktade ju rök . De luktar än , berättar Malin . Malins lägenhet på Decembergatan 51 brann och hon har fått en del hjälp från Poseidon och sitt försäkringsbolag . - Men inte ersättning för allt . Hela vårt liv brann ju upp , säger Malin som på senare tid börjat sakna en del saker som aldrig kan ersättas . Lite värre är det för kompisen Linn Olin som bodde på Decembergatan 35 . Där brann det aldrig . Lägenheten blev dock allvarligt vattenskadad av släckningen . - Vi kunde bärga en del möbler , men jag har haft stora problem med försäkringsbolaget , som till exempel ifrågasätter att jag hade så mycket mat i min kyl och som inte svarar eller är svåra att få tag i , säger Linn . Både Malin och Linn tycker också att det är svårt att få tag i någon på Poseidon som kan svara på frågor . - Alla hänvisar till någon annan . Varför finns det inte någon utsedd vi kan vända oss till med allt i just denna saken , säger båda . Bertil Guslén 031-62 41 74 bertil.guslen@gp.se Mycket försvann för Malin Forsell och sonen Liam i branden på Decembergatan . För Malin Forsell och sonen Liam försvann allt i branden på Decembergatan i augusti . Därefter bodde

de på hotell innan de fick en ersättningslägenhet på Julaftonsgatan . - Vi ska flytta tillbaka , säger Malin . Bertil Guslén,Göteborgs-Posten,20121225

G Appendix G - Stop list and lemma list for the Riksdag data

List of words/punctuation marks to be excluded

-	antal	bra	du
–	applåder	både	då
,	april	bättre	där
;	att	bör	därför
:	augusti	böra	efter
!	av	c	eftersom
?	avse	dag	egentligen
”	avser	de	eller
”	avslag	debatt	emot
(	avslutningsvis	debatten	en
)	avtal	debatterar	ena
±	bara	december	endif
<	bedömning	del	enligt
=	behöver	delar	ens
a	beslut	dem	ersättare
all	bestämmelser	den	ett
alla	besvara	denna	ex
allt	besvaras	deras	exempel
alltid	betänkande	dessa	exempelvis
alltså	betänkandet	det	faktiskt
andra	bidra	detta	fall
anf	bifall	dig	februari
anförs	biföll	din	fick
anledning	bl.a.	direkt	finnas
anmälan	bland	direktiv	finns
annan	bli	direktivet	fler
annat	blir	diskussioner	flera
anse	borde	diskuterat	fp
anser	bort	dock	fram

fru
fråga
frågan
frågar
frågat
frågor
från
frånvarande
få
får
fått
följa
följande
följande
för
föreslår
föreslås
förslag
förslaget
ganska
ge
genom
gjort
god
gå
går
gälla
gäller
gör
göra
ha
hade
han
handla
handlar
har
hela
helt
herr
hon
hoppas
hos
hur
håller
här
i
idag
igen
ihåg
in
införa
ingen
ingå
innebär
inom
inte
interpellanten
interpellation
istället
ja
jag
januari
ju
juli
juni
just
kammaren
kan
kanske
kap
kl
kom
komma
kommer
kommissionen
konvertering
kunna
känna
lag
lagen
ledamot
ledamöter
lite
liten
lägga
länge
m
maj
man
mandat
mars
med
medan
mellan
men
menar
mening
mer
mfl
mig
min
mina
minister
ministern
minst
mitt
mm
mot
motion
motionen
motionstiden
mp
mycket
myndigheter
myn-
digheterna
många
mången
måste
möte
möten
naturligtvis
ned
nej
ni
november
nr
nu
ny

nya
någon
någonting
något
några
nämligen
när
o:p>
och
också
ofta
ok
okej
oktober
olika
om
op>
organ
oss
partier
pga
pratat
precis
procent
productid=
prop
proposition
propositionen
protokoll
på
redan
redogjorde
regering
regeringen
regeringens
res
reservation
riksdag
riksdagen
riksdagsbeslut
rösterna
s
s.k.
sade
sagt
saker
samt
sd
se
sedan
sen
september
ser
sidan
sig
sin
sina
sitt
själv
ska
skall
skriftliga
skulle
som
s
statsminister
statsministern
statsråd
statsrådet
stmetric-converter>
stor
stora
stå
står
svar
svaret
så
sådan
sådana
såg
såklart
sån
säga
säger
sätt
t
t.ex.
ta
taget
talman
talmannen
tanke
tar
tas
tex
tid
tidigare
till
tillkänna
tillkännager
tillkännagi-vande
tillstånd
titta
tog
tro
tror
trots
tycka
tycker
tyckte
tydlig
typen
tänker
under
upp
ur
ut
utan
utreda
utskott
utskottet
v
vad
var
vara
varannan
varenda
varför

varit
vecka
veckan
veckor
verkligen
vet
veta
vi
vice
vid
vidare
vidta
viktig
viktigt
vilja
vilka
vilket
vill
visa
vissa
votering
vår
våra
vårt
väl
väldig
väldigt
w:st=
wst=
yrka
yrkande
yrkar
å
år
åt
återkomma
än
ändå
ännu
är
även
över
överens
särskilt
ofta
oftast
aldrig
fortfarande
ibland
före
jo
människor
nästa
tack
ungefär
ihop
ca
kr
per
bla
dvs
ovan
ner
längre
kring
senaste
antalet
antal
samtidigt
tittar
härmed
överläggnin-
gen
överläggning
ute
trots
grund
längs
norr
södra
mest
självklart
förra
höra
alldeles
överens
fungerar
särskild
dnr
interpella-
tionsdebatt
förstod
stället
rett
dnr
kap
a
gått
oerhört
möjligt
apropå
utifrån
större
möjligt
heller
bakom
er
samma
tillbaka
viktiga
klart
ställa
samma
varje
framför
tillsammans
behövs
behöva
ytterligare
diskutera

List of lemmas to be excluded

all	finnas	kunna	redogöra
anföra	fortsätta	lag	redovisa
anförande	framställa	ligga	regering
ange	fredag	lyfta	riksdag
anmäla	fråga	låta	riktig
anmälan	frånvarande	lördag	riktigt
ansvar	få	meddela	runt
avse	föreslå	mena	röster
avtal	förhandling	mest	sak
behandla	förslag	motion	samtidigt
behöva	förstå	motionstid	se
bestämmelse	försöka	måndag	sig
besvara	förut	mången	sitta
betänka	ge	människa	skäl
betänkande	gång	möte	statsminister
bidra	gälla	nog	säga
bifalla	göra	ny	sätt
bli	ha	ni	söndag
bra	han	någon	ta
böra	handla	något	tack
debatt	hon	nämna	tala
debattera	inboka	onsdag	tid
direktiv	införa	ord	tillkännage
diskussion	ingå	papper	tisdag
dröjsmål	inlägg	prata	torsdag
du	innebära	problem	tro
en	inkomma	process	tycka
engagemang	interpellation	proposition	tydlig
exempel	jag	protokoll	tänka
fall	komma	punkt	uppdrag

utredning	yrka	heta	framför
utskott	yrkande	ses	avstå
vara	återkomma	känna	inne
varje	överens	viktig	ytterligare
veta	överläggning	ställa	klart
vi	skäl	ingen	rätt
vilja	leverera	gå	fel
vår	antal	lägga	varje
välja	senaste	tillbaka	