# Towards Topic Modeling Swedish Housing Policies: Using Linguistically Informed Topic Modeling to Explore Public Discourse

Anna Lindahl[1] and Love Börjeson[2]

[1] Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, annanlindahl@gmail.com
[2] Visisting Scholar, Graduate School of Education, Stanford University, love.borjeson@hyresgastforeningen.se

**Abstract.** Topic modeling is an unsupervised method for finding topics in large collections of data. However, in most studies which employ topic modeling there is a lack of using linguistic information when preprocessing the data. Therefore, this work investigates what effect linguistically informed preprocessing has on topic modeling. Through human evaluation, filtering the data based on part of speech is found to have the largest effect on topic quality. Non-lemmatized topics are found to be rated higher than lemmatized topics. Topics from filters based on dependency relations are found to have low ratings. To exemplify how topic modeling can be used to explore public discourse the area of Swedish housing policies is chosen, as represented by documents from the Swedish parliament and Swedish newstexts. This subject is relevant to study because of the current housing crisis in Sweden.

**Keywords:** topic modeling,housing policies,LDA,public discourse



## 1 Introduction

In the field of humanities and social sciences the use of computational methods has been argued for by many. Commonly referred to as *Digital Humanities*, the importance of tools for investigation of both digital and printed texts is undeniable. However, as Viklund & Borin [1] 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. [2], where the concept of culturomics is discussed, and the need for good linguistic preprocessing to make this a successful field.

One popular method for investigating text is topic modeling, which is an unsupervised probabilistic method for finding topics in collections of data. It has been 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. [3] use it to study ideas within the computational semantics field over time, DiMaggio et al. [4]

investigate the news coverage of U.S. arts funding and Jacobi et al. [5] use it for following trends in journalistic papers. The most commonly used topic model is the Latent Dirichlet allocation (LDA) model and was developed by Blei et al. [6]. This is also the model used in most of the studies mentioned here.

However, many studies including those above differ in their use and reporting of their preprocessing. Preprocessing is an important step in topic modeling, and it 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 effect of different preprocessing choices has not been studied systematically and there is also a lack of using linguistic information in the preprocessing.

Thus, the aim of the present work is twofold. The first is to investigate how one can adapt and enrich topic modeling with linguistic information and knowledge. The second is to exemplify and explore how one can apply this method to investigate the public discourse of Swedish housing policies. This area is chosen because of its relevance, the housing crisis in Sweden has been ongoing since the 1990's and it has been a source of debate for just as long. Lack of housing is still becoming more widespread, with only a small rise in newly built houses in 2015–2016 [7], further adding to the relevance of this subject.

## 2 Related work

### 2.1 Linguistically informed topic modeling

There are a few studies reporting on the effect of linguistically informed topic modeling. Martin & Johnson [8] conclude that topic modeling is more informative and effective using only nouns. Following Lau et al. [9] they also report that lemmatizing improves the results, but that it slows down the topic modeling. They use semantic coherence for evaluation (see the evaluation section) and find that the coherence of the topics improve using only nouns. Jockers [10] 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.

There are also studies which use linguistic information to develop topic modeling for specific purposes. Fang et al. [11] present a novel cross-perspective topic model which models topics and opinions. The topics are modelled using only nouns from the corpora. The opinions related to the topics are modeled using adjectives, adverbs and verbs. Guo [12] 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.

## 3 Data

The data used here comes from two domains of the public discourse, the Swedish parliament, the *Riksdag*, and Swedish newstexts. Both domains were automati-

cally annotated with help of the corpus infrastructure tool of Språkbanken, Korp [3] [13]. The Riksdag data is already available through Korp, and the newstext data were annotated using the Sparv pipeline[4], which is a part of Korp [14].

It should be noted that the language in the two domains differ, the Riksdag data is formal and contains many domain specific words, while the language in the newstexts is more similar to spoken language and the vocabulary is closer to everyday Swedish.

### 3.1 The Riksdag documents

All documents and records from the Riksdag's proceedings and correspondence are freely available online, known as *Riksdagens öppna data* (The Parliament open data).[5] However, here the documents were downloaded from through Korp.

The documents span between 1971 to present day, with the exception of a few document categories missing from the earlier years. There are 20 different document categories and from these seven were chosen. Documents deemed to cover debates, discussions and proposals are chosen. An overview of the selected documents can be seen in table 1. Only the first 3000-4000 words were used from the longer document types except for the *protocols*. This was done with the hope that this part covers the document's topics well enough. The *protocols* will have topics distributed throughout the documents and therefore these were kept long.

| **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 1.** Overview of the chosen document types.

[3] `https://spraakbanken.gu.se/swe/node/1535`
[4] `https://spraakbanken.gu.se/swe/node/19799`
[5] `https://data.riksdagen.se/data/dokument/`

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 long time span. Topics will have varied over time and this might affect the topic modeling. The parliamentary periods with respective document and word count can be seen in table appendix A.

### 3.2 Newstexts

To analyze the media, newspaper and magazine articles have been downloaded from the Media Archive provided by Retriever. [6] The access was provided by the Swedish Union of Tenants(SUT).

In order to find all the newstexts concerning housing policies a search term list was made together with people from SUT who are knowledgeable of housing policies. See Appendix B for the search terms. All 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 used. 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 can be seen in table 2 together with the number of tokens and documents. 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.

| **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 2.** The different periods for the newstexts data.

## 4 Method

In order to compare the effects of different linguistic preprocessing, a number of filters based on linguistic information were designed and applied to a test set of the data. An example of a filter can be selecting all words in the documents which

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

are tagged as nouns or words participating in a specified dependency relation. The filters are described in more detail below.

A topic model was trained on each of the filtered versions of the test set, and the models were evaluated using semantic coherence and human judgement, see below.

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 used for the rest of the parliamentary periods of the Riksdag data, which are then used for exploration of the data.

As previously stated, the language in the two data sets differ, and because of this the highest rated combination of filters for the Riksdag data is not used for the newstexts. Instead, the top five highest rated combinations of filters from the Riksdag 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. The five resulting models from the newstexts are then evaluated in the same way as the Riksdag data.

### 4.1 Preprocessing and linguistic filters

Punctuation and numbers are removed from all documents, and all words are changed to lower-case. Frequent words are removed, words which occur in 50% or more of the documents and words which occur in less than 5 documents are removed. Here, this frequency filtering is referred to as **filter 1**. Unless stated otherwise, this is applied to all documents.

A stop list was used, also defined as a filter. This list was made from a general stop list for Swedish, but it was necessary to manually add domain-specific words to this list.

Through the Korp annotation there is information about lemma, part of speech and dependency relation for every token. From this, a filter of lemmas of words was used, this filter simply replaces words with their lemmas.

Three filters based on part of speech were tested. The first filter uses all parts of speech, called *all POS*. The second filter removes all words which are not nouns, verbs, adjectives and participles, from here on called *POS2*. The third, following [8] uses only nouns.

A filter based on dependency relations was also made. This filter only uses words participating in seven specified dependency relations, chosen with the aim to find the meaningful parts of the sentence. These relations are: agent, object adverbial, direct object, predicative attribute, place adverbial, subject predicative complement and other subjects.

In table 3 an overview of the combinations of filters tested is shown. 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 was chosen for further investigation has thus 5 more filters applied to it.

| **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 1 | - |

**Table 3.** Filters for the Riksdag test set.

The linguistic filters applied to the newstext data can be seen in table 4. 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 was also made up of domain specific and 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.

| **POS2** | **NN** |
|---|---|
| POS 2, Filter 1 | NN, Lemma |
| POS 2, Filter 2 | NN, Lemma, Filter 2 |
| POS 2, Filter 2, Deprel | - |

**Table 4.** Filters for the newstexts test set, filter 2 replaces the stop list.

### 4.2 Topic modeling

The topic modeling was implemented using the python library Gensim.[7] The LDA implementation in Gensim uses a modified version of variational Bayes, made to handle documents in a stream, which makes handling large corpora more effective [15][16]. Part of the evaluation was also carried out with methods in the library, see next section.

When training an LDA model the number of topics needs to be provided. Guided by previous papers, experiments were run between 50 - 200 topics. After

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

manual inspection 75 number of topics where selected for the filter tests. Other than this the default configurations of Gensim were used.

### 4.3 Evaluation

There are several ways to evaluate a topic model. It has previously been shown that held out likelihood of a model doesn't always correspond to human judgement [17]. Here the focus lies instead on the interpretability of the generated topics. This is evaluated both computationally and with humans. Using the coherence model available in Gensim, the two semantic coherence measures *cv* and *npmi* were calculated. These measures calculate the semantic coherence between the words in a topic by using probabilities derived from word co-occurrence statistics. If a topic has high coherence between its words it is presumably also a good topic. The two measures differ in how the probabilities are calculated, see [18] for more details. [18] also finds *cv* to be the best measure, but is contradicted by [19] who finds *npmi* to be the best measure, and therefore these are compared.

To assess the performance of the coherence measures and evaluate topic quality, human judgements were collected. Before this, a short manual inspection of the models were done by the authors. This resulted in two models being disregarded due to them containing mostly useless topics. The rest of the models were kept, in total 16. These models can be seen in table 6 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 [20] and [9], 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 5. 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 5.** 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. As stated in the previous section, five models corresponding to the five top rated combinations of filters from the Riksdag test set was chosen for this.

## 5 Results

Below the results for the models trained on the filtered Riksdag data are presented. In table 6 all the models with their ratings are shown. 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 percentage of the original number of words is also shown. However, this number doesn't seem to have an effect on the ratings.

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 a stop list. The following top ranked models are from the POS2 group, but without lemmatization. The third highest rated model is also filtered based on dependency relations.

For the models using all parts of speech, using a stop list significantly improves the results, as expected. Applying frequency filter 1 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 dependency relations filter 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. 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 6.** Human ratings for all models.

The results from the human judgements for the newstexts can be seen in table 7. 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. Note that the frequency filter 2 replaces a stop list here. The mean ratings and number of 3's are lower overall for the newstext data than for the Riksdag.

| **Model** | **Mean human rating** | **Nr of 3's** |
|---|---|---|
| POS2, Filter 2 | 2.08 | 10 |
| NN, Lemma | 2.036 | 5 |
| POS2, Filter 1 | 1.933 | 3 |
| NN, Lemma, Filter 2 | 1.871 | 4 |
| POS2, Filter 2, Deprel | 1.636 | 0 |

**Table 7.** Results for the chosen models for the newstext data.

When inspecting the topics from the different filters a few patterns were found. In all topics, nouns were the most frequent part of speech, regardless of POS-filter. Non-lemmatized topics had more repetition of the same words but different word forms. The dependency relations captured mostly nouns due to the nature of the chosen relations, but still these topics where not rated as high as the others.

The rankings from the two coherence measures, *cv* and *npmi*, did not correspond to the human rankings for the Riksdag test set. *cv* however has the top ranked model as the second best model. The calculated correlation for the *cv* measure is almost always higher than for the *npmi*, with a mean correlation of 0.68 and 0.60, respectively. 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. See appendix C for more details.

### 5.1 Exploring the public discourse

The highest rated combination of filters from the Riksdag, which was lemmatized nouns, with a stoplist, was used on the rest of the data. The resulting models and classifications of documents is here used to exemplify how one can use topic modeling for examining public discourse. The same was done for the newstexts, but with the highest rated model for this data, the POS2 group with filter 2.

For the Riksdag, 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. In the newstexts, more topics was found relating to housing policies as compared to the Riksdag, due to the selection process.

With this information, one can track changes in the topic over time. For example, figure 1 shows the proportion of documents which contains over 0.35

of this topic in all the *motions*. To filter out the document with a low proportion of the *housing policies* topics, documents with less than 35% of the topic was removed. Inspecting the figure one can see that the topic has a peak in 1998–2002 and 1976–1979.

To further inspect the data, interactive plots were made with the help of the Python library Bokeh.[8] A static version of this is seen in figure 2. It 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.

With these kinds of plots, co-occurring topics can also be examined. Figure 3 is based on newstexts, and shows the mean of each topic for every month during 2014. Only newstexts containing a topic labeled *the lack of housing* are used. The lack of housing topic is removed (nr 25), to be able to see the other topics more clearly.

In the figure, topic nr 33, which is about student housing is slightly more co-occurring during July, August and September, possibly due to the start of the academic year in September. 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*.

## 6 Conclusions

In this work we have shown how one can examine the discourse of Swedish housing policies with the help of topic modeling. The method is deemed suitable for the intended analysis, although there is more work to do for a full analysis of the public discourse.

By using human evaluators, the effects of different kinds of linguistic preprocessing were investigated. 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. Lemmatized data is not rated as high 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

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

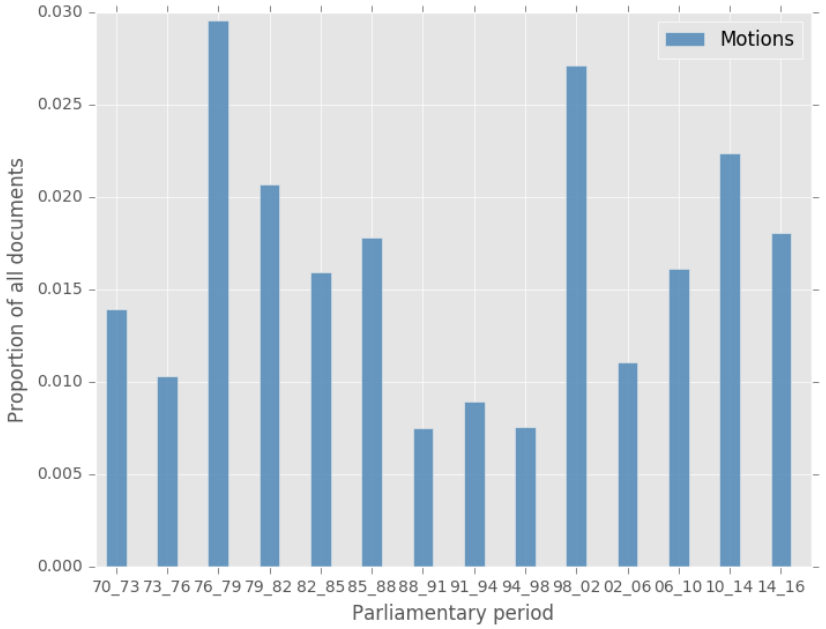


**Fig. 1.** Proportion of documents with a proportion over 0.35 of the topics labeled 'housing policies' in the *motions*.

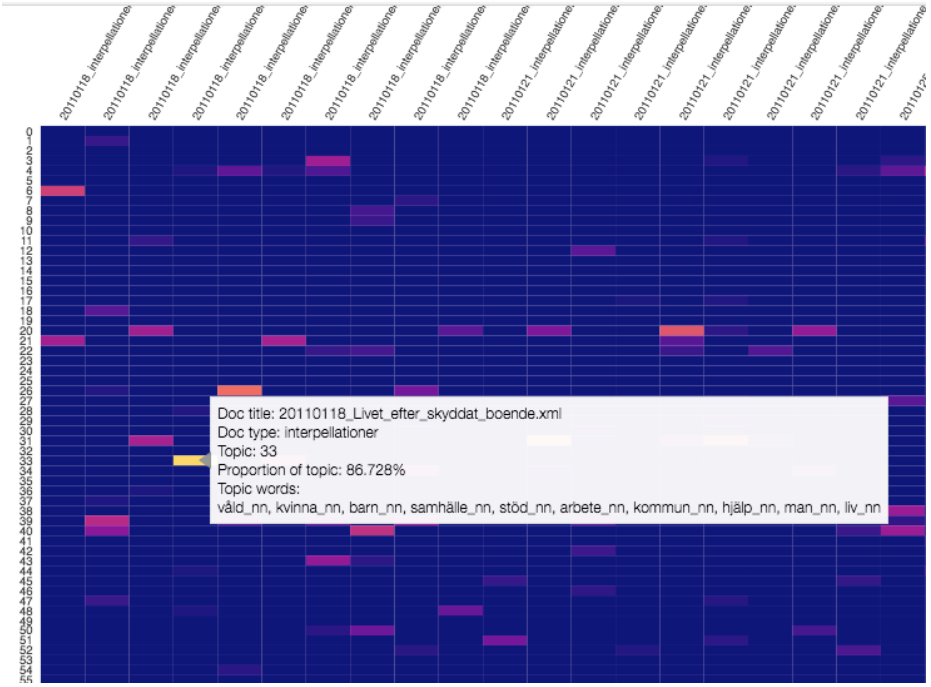


**Fig. 2.** A screen shot of an interactive plot. Columns represent documents and rows represent topics.

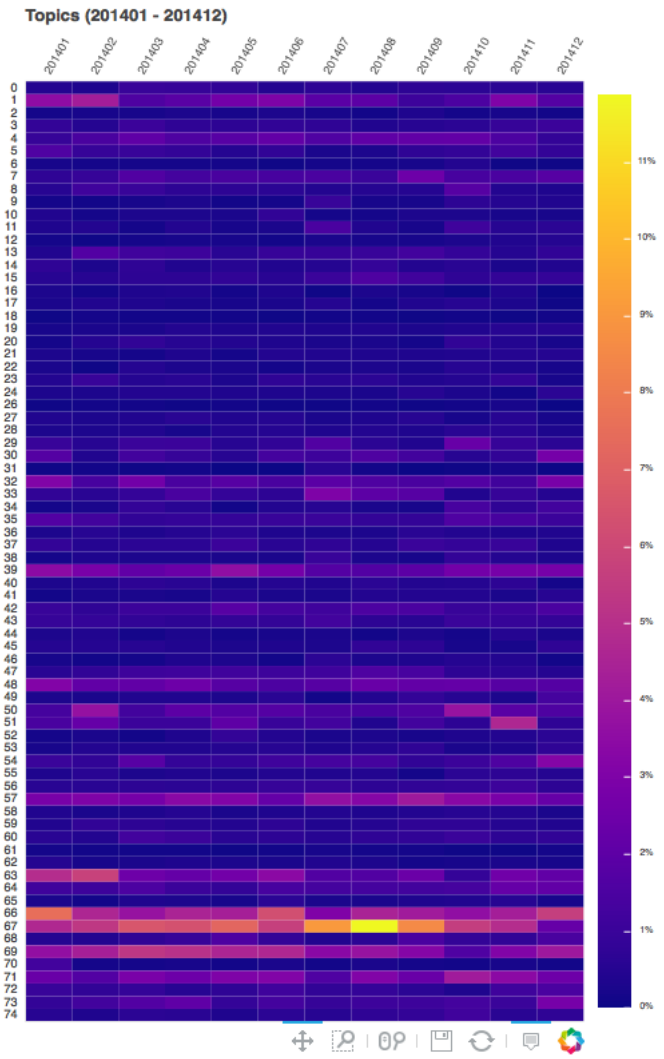


**Fig. 3.** A plot over the mean of each topic for all the newstexts containing the lack of housing topic for each month during 2014. The housing topic is removed.

data is preferred. Using data selected based on dependency relations does not result in topics with high ratings, 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.

**Acknowledgments.** This work has been supported in part by a framework grant for the project *Towards a knowledge-based culturomics*[9], awarded by the Swedish Research Council (contract 2012-5738).

This work has also been carried out with the support from the Swedish Union of Tenants[10], which has provided part of the data used.

## References

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993–1022.
7. Höjer, H. (2017). Därför kan byggboomen inte lösa bostadskrisen. *Forskning och Framsteg*, (2), 61–74.
8. Martin, F. & Johnson, M. (2015). More efficient topic modelling through a noun only approach. In *Australasian Language Technology Association Workshop 2015* (pp. 111).
9. 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).
10. Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history.* University of Illinois Press. pp: 128-133.
11. 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.
12. Guo, S. (2012). *Using Dependency Parses to Augment Feature Construction for Text Mining.* Virginia Polytechnic Institute and State University.
13. Borin, L., Forsberg, M., & Roxendal, J. (2012). Korp-the corpus infrastructure of Språkbanken. In *LREC* (pp. 474–478).
14. Borin, L., Forsberg, M., Hammarstedt, M., Rosén, D., Schumacher, A., & Schäfer, R. (2016). Sparv: Språkbanken's corpus annotation pipeline infrastructure.

[9] `https://spraakbanken.gu.se/eng/culturomics`
[10] `https://www.hyresgastforeningen.se/`

15. Rehurek, R. & Sojka, P. (2010). Software framework for topic modelling with large corpora.
16. 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).
17. 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).
18. 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.
19. van der Zwaan, J. M., Marx, M., & Kamps, J. (2016). Topic Coherence for Dutch.
20. 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.

## Appendix A - Parliamentary periods for the Riksdag data

| **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 8.** Periods for the topic modeling.

## Appendix B - 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*

## Appendix C - Top 5 models from the Riksdag compared to *cv* and *npmi* measures

| **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 |