UPPSALA
UNIVERSITET

Automatic organization of online Swedish political discourse

Hillevi Hägglöf

Uppsala University
Department of Linguistics and Philology
Master's Programme in Language Technology
Master's Thesis in Language Technology

June 13, 2014

Supervisors:
Jussi Karlgren
Mats Dahllöf

Abstract

Political content is continuously posted online in various textual forms. Some distinct channels in which political texts are frequently submitted have been identified. These include: 1) open parliament, (2) social media and (3) editorial media. In the study at hand, the aim has been to identify *what* is being discussed in the above-mentioned text channels in order to - at some point in the future - be able to compare what is being discussed over the different channels. It is a first step in a larger initiative to open up online political discourse in Sweden.

To do this, topic modeling was performed using two methods: the state-of-the-art method Latent Dirichlet Allocation (LDA) and an experimental method called temporal χ^2 topic modeling. Three corpora were collected and constructed from each data channel respectively.

Topic modeling with LDA was performed on data from the Swedish parliament only as LDA proved to be inappropriate for these particular data channels. Topic modeling with temporal χ^2 was performed successfully for all three data channels. However, formal evidence in the form of a standardized evaluation is lacking. Further, it is clear that further preprocessing of social and editorial media data is needed in order to get truly useful results. It is, for instance, necessary to extract texts that explicitly concerns politics.

The evaluation approach is manual. Requirements of evaluations methods are, however, discussed as the traditional methodology is deemed insufficient.

Contents

Preface **5**

1 Introduction **6**

2 Background **7**
2.1 Confusion of concepts 7
2.2 Topic modeling 7
2.3 Definitions 8
2.3.1 Topic definition 8
2.3.2 Task definitions 9
2.4 Previous topic modeling research 10
2.5 Evaluating topic models 11
2.5.1 The Cranfield paradigm 12
2.5.2 Traditional topic modeling evaluation methodology . . . 12
2.5.3 Pragmatic topic modeling evaluation requirements . . . 13

3 Method **16**
3.1 Data 16
3.1.1 The parliament corpus 16
3.1.2 The social media corpus 17
3.1.3 The editorial media corpus 18
3.1.4 Linguistic features of editorial media text 18
3.2 Preprocessing the corpora 18
3.3 Latent Dirichlet Allocation 21
3.3.1 Model 21
3.3.2 Inference 22
3.4 Temporal chi-squared (χ^2) topic modeling 24
3.5 Evaluation approach 25

4 Experiments **30**
4.1 Experimental design LDA 30
4.1.1 LDA experiments 31
4.2 Temporal χ^2 experiments 32

5 Results of the topic modeling experiments **34**
5.1 LDA results 34
5.1.1 Topic construction 34
5.2 Topic inference 40
5.3 Results of the χ^2 experiments 42

5.3.1 Temporal χ^2 parliament results 43
5.3.2 χ^2 social media results 46
5.3.3 χ^2 editorial media results 48
5.3.4 Results overview 52

6 Discussion **54**
6.1 Discussing the LDA results 54
6.2 Discussing the temporal χ^2 results 55
6.3 Tuning the topic model 56
6.4 Evaluation 56

7 Conclusions **57**

8 Lessons learned **58**

Bibliography **59**

Preface

This thesis puts an end to twenty years of schooling. It has been delayed by depression and the loss of a father, but it has inspired a computational perspective on politics. It has also spawned the birth of a company.

I would like to direct a thank you to my supervisor Jussi Karlgren for keeping my spirits up during this journey.

1 Introduction

The web is overflowing with political content. Original political texts are posted online every minute of every day. It is a diverse and heterogeneous movement that is taking place in every language possible, discussing every topic imaginable. The text types vary from opinion pieces starred on major news sites read by millions of people, to microblog entries posted among friends, and speeches held in various parliaments, read only by journalists and experts. By this, we wish to say that political content is being submitted to the open web in *all possible textual forms* with *drastically varying senders and receivers*. It lies in our interest as researchers to make use of this incredible data source to make sense of the political web.

The aim of this pilot study is to automatically organize the online Swedish political discourse. This, of course, requires appropriate data. As previously mentioned, political discussions are taking place almost everywhere on the web. While this is true, it is easy to identify some data channels in which political discussions are prominent. These include:

1. editorial media,
2. social media and
3. material produced by governmental institutions.

These resources constitute the greater part of the Swedish online political discourse.

Automatical organization of political content is considered with a specific application in mind. This application is a dynamic atlas of online Swedish political discourse. What is a dynamic atlas of political discourse and how can it be achieved? A dynamic atlas - as we understand it here - implies an *automatic organization of political discussion happening online*. To be truly useful, it also needs to be comprehensive and humanly readable. This can be performed by automatically extracting topics, or some other high level semantic entity, given a discourse at hand.

This project, thus, aims to lay the methodological foundation of a topic modeling system that identifies *what* is being discussed in the above mentioned data channels. To do this, a set of topic modeling algorithms are evaluated. Note that this is a *first step in a larger initiative* which aims to open up the online political discourse in Sweden. The atlas is not, however, included in the scope of this project.

2 Background

2.1 Confusion of concepts

Within the field of topic modeling, there is an apparent lack of standardized terminology. In the context of this thesis, the term *topic modeling* refers to the task of automatically detecting topics in natural language. It does *not* refer to the specific approach taken by researchers Croft and Xu (1999) with which the term is generally associated.

Consider the following list:

> *Buzz word monitoring; emerging topic detection; event-based information organization; event detection; event threading; first-story detection; hot topic detection; new event detection; story link detection; topic detection; topic extraction; topic identification; topic labeling; topic modeling; topic tracking.*

This list is comprised of terms for topic modeling or tasks closely related to topic modeling. Many of these are used synonymously. It illustrates the need for a single term and further motivates the use of *topic modeling* as an umbrella term. However, there are sub-tasks that naturally require specific terms. Some of these are listed in section 2.3.

2.2 Topic modeling

Topic modeling is a research field that originates from a body of research called Topic Detection and Tracking (TDT) formed under the Translingual Information Detection, Extraction, and Summarization (TIDES) program, which aimed to find and follow events in streams of broadcast news stories. Today's research in topic modeling is grounded in the TDT initiative but the field has naturally evolved since its founding. The most dramatic shift of focus lies in the types of media that are taken into consideration. In 2013, as opposed to the late 1990's, broadcast news are no longer the dominant source of news. News stories are first and foremost published online via editorial and social media. In some ways, this makes topic modeling in the 21st century easier as researchers are spared the task of automatic speech recognition.Online news streaming from multiple sources quickly measures up to vast amounts of data, which creates the need for highly efficient processing. This introduces new challenges to the field.

The topic modeling project at hand is concerned with political texts. However, the news domain, which is the standard topic modeling domain, and the political domain are highly overlapping as political stories are often reported as

news stories and vice verse. This implies some non-standard challenges such as collection of data from atypical sources and the selection for political news from general news sources. However, the difficulties of constructing a topic modeling system of political discourse are largely similar to the difficulties of constructing a topic modeling system for news stories. Research done on news story topic modeling should therefore be a guiding star here.

Topic modeling is a difficult natural language processing task. It not only requires big data and efficient computing *to be truly useful*, but in most cases, it also presupposes an automatic understanding of semantics. A topic modeling system need not only be able to extract important and representative terms but it also needs to be able to efficiently communicate this to the system's end user. These are no simple requirements.

This background aims to give an orientation to the field of topic modeling research by defining some crucial tasks and concepts as well as by giving an overview of state-of-the-art topic modeling research. Some important tasks and concepts in topic modeling are defined and explained in section 2.3. In section 2.4, the standard approaches to topic modeling as well as state-of-the-art research are accounted for. Section 2.5 describes the topic modeling evaluation methodology.

2.3 Definitions

The following sections serve to introduce and define some concepts crucial to the topic modeling community. In section 2.3.1, the notion of a topic is defined. Section 2.3.2 goes on to introduce the field of topic modeling by briefly defining the characterizing tasks of standard systems.

2.3.1 Topic definition

The work presented here spans multiple genres of texts. The features of a topic will inevitably vary with the genre. Therefore, a broad and genre-independent definition of a topic is needed.

Very generally, a topic is a non-trivial event, activity or abstract entity taking place somewhere at some point in time. This is also the official definition of a topic according to the first TDT joint task (Fiscus and Doddington, 2002). It is an accurate definition of a topic, albeit somewhat imprecise. It tells us very little about the manifestation of a topic in the context of language.

In topic modeling of news stories, with which the topic modeling community is largely preoccupied with, a topic is - quite accurately - defined as a collection of interrelated stories about some seminal event (Brants et al., 2003; Cieri et al., 2000; Lam et al., 2001; Nallapati et al., 2004). This definition suffers the same flaws as the above mentioned definition - it tells us nothing about how such an event is manifested in natural language. It is also insufficient as it does not take into consideration the distinct nature of various genres. For example, it seems likely that microblogs, parliament statements and editorial media have varying degrees of sensitivity to seminal events.

In fact, previous research on topic modeling is rarely concerned with definitions of topics. Most definitions are brief and non-formal descriptions (Brants

et al., 2003; Jurgens and Stevens, 2009; Nallapati et al., 2004; Stoyanov and Cardie, 2010) or altogether non-existent (Canhui Wang et al., 2008). There are, however, exceptions to this rule. One line of research, whose advocates are primarily active in social media research, defines topics as coherent sets of semantically related terms (Blei, 2011; Fung et al., 2007; Sahlgren and Karlgren, 2008). We will adopt this definition here as it allows us to define and redefine a given topic in accordance with its desired scope. Further, it is a useful definition as semantically related terms are dynamic entities of terms, as are topics per se.

What is semantic coherency? Semantic coherency is a term used in both computational linguistics and general linguistics and describes what makes a text or an utterance semantically meaningful. Semantic coherency is defined by Beaugrande (1996) as a *the mutual access and relevance within a configuration of concepts and relations*. That is, semantically coherent words needs to be somehow connected according to the logic of the text.

An example of trivial coherent and incoherent sets of lexical items can be viewed in 2.1 where the coherent set of lexical items are related words to the Wikipedia entry *Linguistics*[1] and the incoherent set of lexical items are randomly selected words from a dictionary.

Illustration of semantic coherency
Coherent set of lexical items: phonetics, phonology, morphology, syntax, semantics, pragmatics, stylistics, semiotics.
Incoherent set of lexical items[2]: scholasticism, contraception, indicated, june, adipoceriform, shittleness, scyphiform, vinasse.

Table 2.1: Constructed examples of coherent and incoherent sets of lexical items.

To automatically determine if a set of lexical items are semantically coherent is a difficult task. It requires an understanding of the underlying semantics of the lexical items as well as the relatedness between the words. Semantic coherence can be computed in a number of ways with varying quality. However, it is an extensive task due to the fact that a topic needs to be evaluated not only in regards to coherency but also relevance.

2.3.2 Task definitions

Topic modeling is an umbrella term which makes a reference to a vast field of research that attempts to automatically attribute documents or similar subsets of data with relevant topics in order to organize, summarize and orient users in large collections of documents. It is a complex natural language processing task that is comprised of several sub-tasks. Some of these sub-tasks are essential to the system, while others are not. These may be excluded, depending on the application and the requirements of the system. The most crucial tasks of topic modeling will be briefly defined below.

[1] `http://en.wikipedia.org/wiki/Linguistics`

1. ***Topic detection*** (also called *topic extraction*) refers to the process of automatically attributing topics to a given document. Topic detection is a query-free task. That is, it requires no user input.

2. ***Hot topic detection*** (also called *buzz word monitoring* or *emerging topic detection*) refers to the detection of topics that occur frequently in a data stream within a given time interval, but that rarely occurs outside of that specified time frame. Hot topic detection is a query-free task.

3. ***Topic tracking*** refers to the tracking of documents in which a given topic occurs. Similar to document search. The given topic is specified by the user.

4. ***New topic detection*** (also called *new event detection* or *first-story detection*) refers to the task of identifying a topic that occurs in the document collection for the very first time. New topic detection is a query-free task.

The scope of the system at hand is limited to topic detection. Topic tracking, hot topic and new topic detection are, however, obvious extensions suitable or future research.

2.4 Previous topic modeling research

Topic modeling systems typically aim to organize text quantities that are too large for humans to review. Chong Wang et al. (2011) writes about topic modeling algorithms:

> Topic modeling algorithms are statistical methods that analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time. [...] Topic modeling algorithms do not require any prior annotations or labeling of the documents - the topics emerge from the analysis of the original texts. (Chong Wang et al., 2011, p. 754)

This is a suggestive summary of the state of the art algorithms used within the topic modeling community. In the following sections, some methods of performing topic modeling are presented.

Previous research on Latent Dirichlet Allocation

The most commonly used topic model is based on the Latent Dirichlet Allocation (LDA) method (Blei et al., 2003). LDA is a probabilistic model that assumes that all documents in a given set of documents share the same set of topics, but that each document exhibits the topics at hand to varying degree (Blei, 2011). The underlying intuition of this assumption is that a given document consists of a number of topics.

LDA is a generative model, generating observable data given some hidden structure. In the case of topic modeling, the documents and the words of the documents are the observed variables while the underlying topics are the hidden

structures that the model infers from the observed variables. It is useful to think of this as a reversed generation process: The observed states, i.e. the documents, are generated by the hidden states, i.e. the topics. A topic model like LDA can thus answer the question: what do the hidden structures, that likely generated the observed states, look like?

Technically, this is done by specifying a joint probability distribution over the words of the documents and over the topic structure. Using this distribution, a conditional distribution of the topics given the words of the documents is computed. Inference of the topic structure from the words of the documents is then performed using the conditional distribution, also called the posterior distribution. A more in depth explanation of the LDA procedure can be found in section 3.3.

LDA has proven to be a successful topic model (Blei et al., 2003; Hoffman et al., 2010; Porteous et al., 2008; Chong Wang et al., 2011). However, the method has some serious flaws. The biggest known issue is that LDA makes a bag of words assumption. That is, the model does not take the word order of the documents into consideration. From a linguistic point of view, this is an unrealistic assumption as there is absolutely no reason to believe that syntax carries no information about topic structure. Chong Wang et al. (2011), however, argues in favor of the bag of words assumption, stating that it is in fact a reasonable assumption "if our only goal is to uncover the coarse semantic structure of the texts" (Blei et al., 2003; Chong Wang et al., 2011). In previous research, however, attempts have been made to relax - or altogether disregard - the bag of word assumption. Wallach (2005) has performed topic modeling with an LDA model that generates words dependent on previously occurring words of the document outperforms the standard LDA model (Blei et al., 2003) as well as hierarchical LDA models (Blei et al., 2003). Griffiths et al. (2005) have successfully used a hybrid method for topic modeling by combining the standard LDA model with a Hidden Markov Model.

Another serious drawback of the LDA method is the need to initialize the LDA model with a pre-defined number of topics. This is a problem as there is no way of estimating a reasonable number of topics beforehand due to the field's uncompromisable need for data-driven methods.

Apart from standard, hybrid and hierarchical LDA methods[3], several other versions of the algorithm have been proposed over the years. These include, for example, Pachinko allocation machines, spherical topic models, correlated topic models and sparse topic models (Chong Wang et al., 2011). These are, however, of little or no use to the project at hand as it is highly recommended to initialize experiments using the standard algorithm. Additional methods will therefore not be presented here.

2.5 Evaluating topic models

Evaluation of information retrieval systems is generally a difficult task. These difficulties are often attributed lack of gold standards within the field. However, this is far from the whole story. In section 2.5.1 and 2.5.2, the traditional

[3] Only the standard method is presented in the study at hand.

methodology of evaluating topic modeling systems and its paradigm are accounted for. In section 2.5.3, the traditional methodology will be compared to the pragmatic requirements of an evaluation framework for information retrieval systems.

2.5.1 The Cranfield paradigm

The traditional methodology of information retrieval evaluation is based on the Cranfield paradigm, also known as batch-mode evaluation. It it the most widely used paradigm for evaluation of information retrieval systems and is characterized by the metrics precision and recall. The Cranfield paradigm-based evaluation consists of:

- an information need,
- a set of documents and
- a given set of relevant documents for each information need.

This evaluation is composed of two main components: (1) a static test collection (consisting of a document collection, a query set and judgments of relevance) and (2) evaluation metrics. The core of this evaluation is that once a test collection is established, it can be re-used to evaluate any information retrieval system. This has made the field of information retrieval very successful. Nonetheless, it might be time to rethink this tradition. Why? Because the Cranfield paradigm makes neutrality assumptions that it cannot live up to. Consider the following: a set of pre-defined queries and a pre-defined understanding of the term *relevance* – is that neutral? More on the difficulties of the traditional methodology can be found in section 2.5.3.

With the rise of new challenges, new evaluation solutions are inevitably necessary. If there are no well-defined search queries, such as "display all documents about the Atlantic Coast Line Depot in Florence, South Carolina", but rather fuzzy queries such as "show me what is going on in the Swedish parliament today", there can be no well-defined answers. Therefore, the evaluation paradigm needs to evolve.

Read more about the guiding principles of an extended Cranfield paradigm in section 2.5.3.

2.5.2 Traditional topic modeling evaluation methodology

With all due respect to the Cranfield paradigm, but how is it implemented in topic modeling?

In order to evaluate the performance of topic modeling systems, previous research has typically made use of manually annotated resources (Brants et al., 2003; Chen et al., 2007; Kumaran and Allen, 2004; Lam et al., 2001; Nallapati et al., 2004; Stoyanov and Cardie, 2010). The typical procedure of this approach is simple:

1. Put aside a subset of a corpus labeled with topics as test set.

2. Tune topic detection and tracking system on the remainder of the unlabeled corpus, i.e. the training set.

3. Evaluate the performance of the model on the test set using preferred precision and recall metrics.

Using this methodology, topic detection and tracking models are most often evaluated in regards to detection error trade-off and detection cost functions using an external resource that serves as a gold standard. The corpora most often used are the Linguistic Data Consortium's TDT Corpora[4] that consists of sub-corpora in Arabic, English and Mandarin. In other words, there are no topic annotated corpora of Swedish texts available, which makes the use of the TDT corpora impossible in this project.

There are also some qualitatively influenced approaches of evaluating topic detection and tracking systems. These methods rely primarily on human judgments. Jurgens (2010) manually identifies a set of topics that are likely to occur in a corpus at hand and subsequently evaluates the system's success in verifying the presence of these topics. Cataldi et al (2010) similarly evaluate emerging topics on Twitter using case studies. However, these evaluation methods require a large set of manually analyzed topics in order to give a fair evaluation of a system. This is a complex, laborious, time consuming and error-sensitive task. Hulpus et al. (2013) address these difficulties and suggest a comparative evaluation based on crowd-sourced human judgments.

2.5.3 Pragmatic topic modeling evaluation requirements

As previously stated, there are multiple problems associated with the traditional evaluation methodology of topic modeling systems. Disregarding those problems, is the traditional methodology really what we are looking for? The following section will make a deep dive into the field of information retrieval evaluation to try to answer this.

Traditional methodology for evaluation of natural language processing applications and information retrieval tasks advocates the measurement of success using metrics such as precision and recall paired together with the computational complexity of the task at hand. Much of the success of these research fields is owed to their convenient and formal evaluation schemes. To a large extent, however, the methodology at hand concerns research tasks where the end user is a system. For example, a word sense disambiguation module in an information retrieval task.

If the scope of the system (the system in which the topic modeling at hand is intended to included) is extended to human end users, the evaluation framework needs to be sensitive to a broad range of usage scenarios. The outcome variables should be based on user and usage factors rather than on the characteristics of the database of documents.

In other words, *a human has more complex information needs than a computer*. This is true in the case of topic modeling. Therefore, traditional evaluation methodology is insufficient. It lacks the real requirements of a useful topic modeling evaluation.

[4] `http://projects.ldc.upenn.edu/TDT/`

So, what should a proper topic modeling evaluation look like? At the convergence of understanding users, information needs, items of information and interaction lies the concept of *relevance*. Relevance is a function of a task and its characteristics, user preferences, background, situation, tools, temporal constraints and numerous hidden factors. The target concept of relevance is based on the everyday definition of that term, but should be operationalized to be a topical and task-oriented relation between query and information item.

As contextual factors are left out in the traditional topic modeling evaluation, as they are in information retrieval tasks at large, the sections at hand will discuss and perhaps establish how relevance can be extended to formally cover contextual information and formulate experiment goals to verify the new formalizations.

For the proposed system, the notion of context-sensitive relevance is crucial to evaluation and experimentation. For the community at large, a well-crafted context-sensitive relevance measure will be a durable contribution to the field, which thereafter will only be able to ignore context and usage factors at its own peril. However, how this target for quantitative relevance can be extended to non-topical access scenarios with uncertain link to specific tasks is not obvious – should it be supplanted by more general notions such as user satisfaction or pertinence or should the notion of relevance be enhanced and extended? This is not an easy question.

First of all, the rationale of users may be different. For usage which is more directed towards less urgent entertainment, which may in fact be the case of the system at hand, rather than fulfilling timely information needs, user satisfaction is less obviously modellable. This is an argument for more wide-ranging target metrics, beyond topical relevance.

Second, the media itself and its content are factors against topical evaluation. While some material may be topically analyzable, other may be intended to serve other needs – perhaps intended to provoke a sensation or provide momentary enjoyment. Finding common content features over a set of such materials is not obviously possible. It can be argued that most important factual content is available in linguistic form (and thus amenable to topical indexing) and that other materials should be accessed in completely different ways. This calls into question the primacy of topical indexing.

Third, interaction is diverse. If the interaction is based on system-initiated recommendation rather than on user-initiated search; on likeness to examples, rather than goal-directed matching to queries; on satisfaction rather than optimization, the target metrics must be different. The information system's ranking of information items, and the evaluation of that ranking must be done using schemes different from those in current use.

Typically, systems for entertainment and diversion are evaluated based on less formal and formidable target notions of "user experience" or "user satisfaction". One long term vision of the evaluation of the topic modeling system at hand is that system designers will be able to take into account affective factors in their design - moving beyond the current focus of information systems as primarily topical and task-based to include *information systems that are genre- and appeal-based*, while retaining the valuable qualities of quantitative information system evaluation, albeit with a wider scope of target notions.

In recent years a surge of interest in sentiment analysis, attitude identifi-

cation, and opinion mining has shown how much of that signal is explicitly identifiable and potentially useful. Observation of the potential usefulness is especially pertinent to the study of multimedia information access, where topic- and task-based information access recedes from the forefront compared to appeal-based information access; where users are less inclined to expend effort to specify and optimize their needs as compared to accepting or discarding offerings and satisfying their needs. The models used in sentiment and attitude analysis are often based on rather narrow scope annotation schemes with respect to coarse categories. While these models form a starting point, the link between affect, content and user satisfaction needs further attention and a solid grounding in contemporary behavioral psychology, linguistics and information access research.

What does all this mean? How do we construct a better evaluation methodology from these conclusions? This is still unclear. All of these above mentioned factors clearly need to be included in to the evaluation. But into what framework will we insert them? The evaluation framework needs to rely on:

1. **Reproducibility.** If the experiments and its results are not reproducible, it is not scientific. This is a fact. Nonetheless, it poses a strict and flawed requirement on an evaluation framework, namely *the use of a static set of documents*. Needless to say, an evaluation methodology that requires the use of a static set of documents is far from an ideal evaluation. In the age of big data, it is not realistic to demand that a researcher produce the exact same results as a previous researcher, but rather similar. If similar or equivalent material has been used to perform the experiments, the results should be similar – not exactly the same.

2. **Transparency.** Relevance is not binary. Due to this, an evaluation that is transparent and that allows researchers and users to understand the results is necessary.

3. **Portability.** Information retrieval evaluation needs to become portable. To date, topic modeling evaluation is impossible for more than a handful of languages. It is also impossible in a domain distinct from that of editorial media. It is crucial that new frameworks needs to allow - and foster - portability.

This project intends to take the first steps to close some of the gaps in the description above. Please note that there are no intentions to build such an evaluation system within this project, as it is beyond the scope of the study. The matter requires urgent discussion.

3 Method

3.1 Data

The current experiments have been performed with open access web data exclusively. Data channels include:

1. parliament,
2. social media and
3. editorial media.

In the following chapter, the corpora of the project are presented. That is, resources, preprocessing and characteristics of the corpora are accounted for. Linguistic features of the different texts are given in section 3.1.1. Preprocessing steps are explained in section 3.2.

3.1.1 The parliament corpus

The Public Sector Information Directive[1], also known as the PSI Directive, states that public sector information (that is, information produced, collected or financed by public bodies) should be freely available to the public. In accordance with the PSI Directive, the Swedish parliament has since 2010 maintained an application programming interface (API) from which anyone is able to anonymously and without registrations or fees collect various documents produced by the parliament. The present API is a simple representational state transfer API. The API is available via `data.riksdagen.se`.

The resources include data sources such as calendars, documents, information about members of parliament (current and historically), reports, voting results, protocols, speeches, interpellations, replies and motions. The current corpus includes speeches, interpellations, replies and motions. This is first and foremost due to the fact that these are the most natural language heavy documents.

To collect parliament data, a combined data collector and parser was constructed in Python. To download XML documents from the Swedish parliament's API, the Python module `urllib2`[2] was utilized. The resulting XML and HTML documents[3] were parsed using the Python HTML/XML scraping library `Beautiful Soup`[4].

[1] 2003/98/EC

[2] `docs.python.org/2/library/urllib2.html`

[3] The output from `data.riksdagen.se` depends on the type of document requested.

[4] `crummy.com/software/BeautifulSoup`

Linguistic features of parliament text

Speeches, interpellations, replies and motions are delivered by ministers and members of the parliament. Transcriptions from the Swedish parliament are intended to be verbatim to the utterances as they are worded in the chamber. They are, however, professionally edited and standardized in reference to morphology and syntax (Dahllöf, 2012). In other words, the documents provided by the parliament are in fact more linguistically homogeneous than the actual speeches. Given that it is a strictly formal domain, the difference is estimated to be marginal. This is further emphasized by the fact that the speakers most often read directly from manuscripts. This makes the communicative style monologic rather than interactive (Dahllöf, 2012).

3.1.2 The social media corpus

The social media corpus is primarily collected from the microblog platform Twitter and the blog manager Twingly. The data was made available to the project by Jussi Karlgren at Gavagai AB. The social media corpus is comprised of two weeks worth of data from Twitter and Twingly in October and November of 2012. The social media data partially overlaps with the parliament test set. Note that even though the time period of the social media corpus is much smaller than the parliament period, the final data quantity is much larger than for the parliament corpus as the social media stream[5] is much wider than the parliament stream.

Preprocessing of the social media corpus does not include filtering to obtain material in which politics is explicitly discussed. This is due to an underlying assumption that all synchronized topics will discuss politics or - somewhat more broadly - news. Therefore, no filtering should be necessary. There are, however, ways of retrieving entries with a higher precision that could be used if needed. For instance, it is possible to use hashtags of different types. On Twitter, for instance, there is a hash tag `#svpol`, which is used as a means of declaring the content of a tweet as belonging to the topic SWEDISH POLITICS. Further, it is possible to make use of similar annotations on, for example, blog portals.

Linguistic features of social media text

The term *social media* refers to a broad category of text, spanning from high quality factual prose posted on well-known blogs to five character spam messages tweeted by bots on Twitter. Therefore, it is generally difficult to list the linguistic features of social media text as it is in no way a unified domain. Even though this makes it difficult to make generalization about the linguistic features of social media text, some things are clear. In contrast to, for instance, parliament data, social media text is obviously less formal. Entries from Twitter never exceed 140 characters. Frequent domain-specific features of a tweet might include user links, hash tags, shortened URLs and abbreviations. Abbreviations may refer to lexicalized items. An example of this is *RT* which stands for *retweet*.

[5]The term *stream* refers to a complete set of data from a given domain. Please note, however, that this is a term that describes an abstract entity and that there is no way of collecting an entire stream of any domain

It indicated that the tweet at hand is a copy of another user's tweet. Other abbreviations are less general and may be specific to a certain group of users or even individual users. Another prominent linguistics features of social media text is the interlinking. Both tweets and blogs are often interlinked. Blog entries are often oriented towards seminal events. Blogs, however - like tweets - are highly individual which makes it difficult to make generalizations about them without first performing a proper register analysis. Research about the linguistic features of the language of the web in this decade is desperately needed.

3.1.3 The editorial media corpus

The editoral news media corpus was collected by Moreover Technologies and was made available to the project by Jussi Karlgren at Gavagai AB. The texts of the corpus originates from various Swedish editorial news sites.

3.1.4 Linguistic features of editorial media text

There is a distinction between hard and soft news in press text. Hard news are reports of accidents, crimes and other events, including announcements of various kinds. An important feature of the hard news story is that the coverage of the story is published while it is still occurring or has very recently occurred. Soft news, on the other hand, are feature stories. Soft news stories are explanatory and opinionated. In these stories, backgrounds are often given; conclusions are drawn and judgments are passed.

Press text, and hard news text in particular, are characterized by a highly standardized type of language. The message of the text is clear and the news of the story is maximized. One salient feature of press text is the lack of redundant information. The standardization of the language of press text is due to the medium's heterogeneous reader group. This is a natural consequence of the public accessibility of news text. It should also be noted that all news text is editorial. That is, all published stories have been reviewed by an editor who cross-checks facts, contents and style. This factor contributes adds to the standardized language of news texts.

3.2 Preprocessing the corpora

Preprocessing of the corpora consists of some basic NLP operations. Some amount of preprocessing is desirable in order to reduce the complexity of the system while other processing steps are required to make the most of the data. The preprocessing steps undertaken include:

- Tokenization
- Punctuation removal
- Digit removal
- Conversion of all letters from uppercase to lowercase
- Stop word filtering

- Verb removal
- Hapax removal

Punctuation and digit removal as well as conversion of all letters from uppercase to lowercase are self explanatory. Punctuation is removed to standardize the corpora so that *wolf* and *wolf.* will not appear as two different lexical items. This is also the motivation behind performing case conversion. Digits are removed as they are assumed to have little or no linguistic information.

Stop word filtering is performed because high-frequency words are assumed to hold little or no information about semantics (Jurafsky and Martin, 2008). To reduce the noise generated by these words, a stop list can be constructed and implemented. A stop list is a list of words that are filtered out and subsequently removed from the corpus. A stop list implementation will reduce noise in any application in which the corpus is utilized, but it is also important in order to reduce the computational expense. A smaller corpus means a reduced computational cost.

As three different corpora from three distinct domains are used, three different stop lists were constructed and applied to the corpora. According to Zipf's law, the frequency of any word is inversely proportional to its rank in a frequency table. The most frequent word will, in accordance with this law, occur approximately twice as often as the second most frequent word. The most frequent words, however, do not carry much meaning (Li, 1992). They are most often grammatical words such as *the*, *for*, *of* etc. As it is not possible to completely rule out that the most frequent words for the three different corpora are completely overlapping, term frequency distributions were, as previously mentioned, created for each corpus. These can be viewed below in figure 3.1, 3.2 and 3.3. An ocular inspection of these figures, however, reveals that the frequency patterns of the different corpora are similar. The frequency cutoff has been set to 200 as the frequency distribution for all corpora seem to stabilize at this threshold. That is, all occurrences of the 200 most frequent words have been removed from the corpora at hand.

Verbs are also known to hold less semantic information than, say, nouns or adjectives. Verbs also hold little information about the content of a topic. Therefore, verbs were filtered out of the corpora at hand. This was done by compiling a verb list using a part-of-speech tagged version of the Stockholm-Umeå Corpus (SUC) Ejerhed and Källgren, 1997, which is a balanced corpus of Swedish texts from the 1990s consisting of one million words in total. This particular corpus often represents "standard Swedish" in Nordic linguistic research. Verbs outside of the scope of the standard Swedish domain are thus kept in the corpus. The motivation behind keeping these verbs is that they are, in some sense, unique to the domain at hand and might therefore add to the semantic content of the corpus. 12062 verbs were filtered out of the corpora.

Hapax legomena are words that occur only once in a given corpus. Just as words that occur very frequently, hapax legomena also hold little or no semantic information (Jurafsky and Martin, 2008). Removing them is a good way of further reducing the size of the corpora which in turn minimizes the computational cost of the algorithms. Hapax legomena were therefore removed from the corpora at hand.

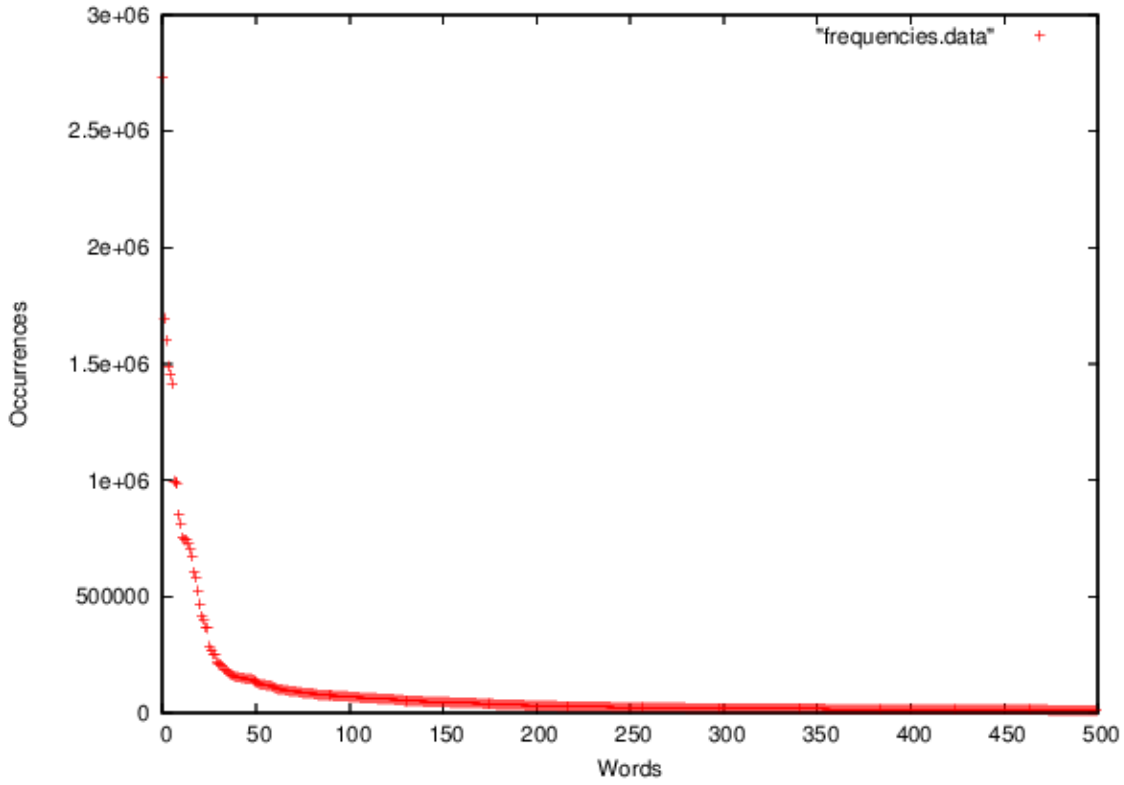

Figure 3.1: Term frequency distribution of the parliament corpus.

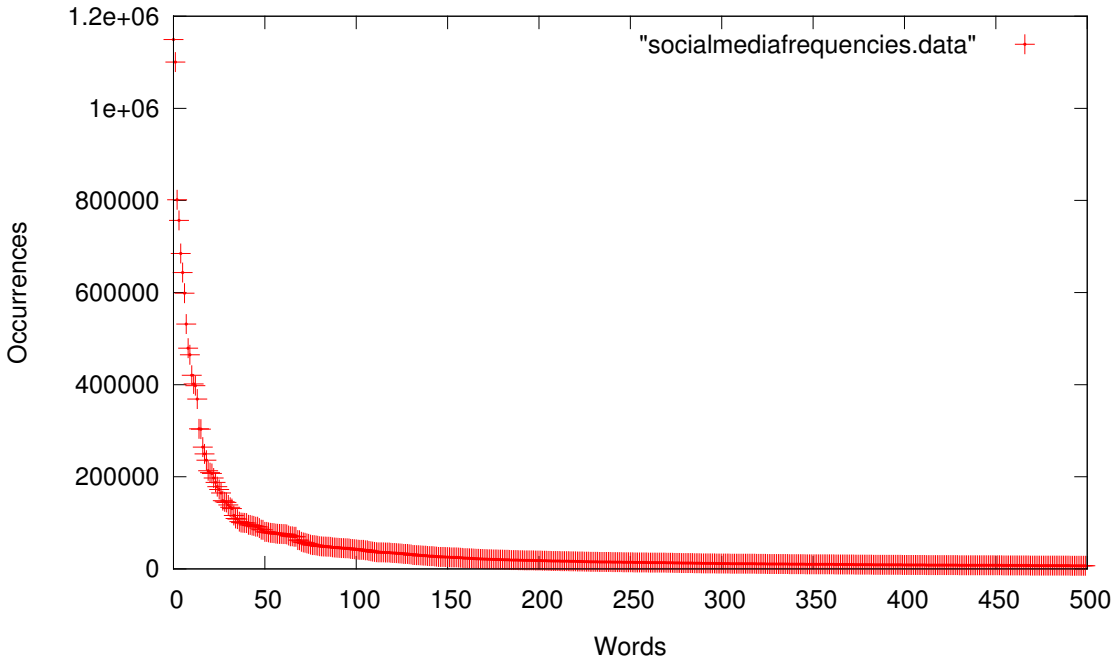

Figure 3.2: Term frequency distribution of the social media corpus.

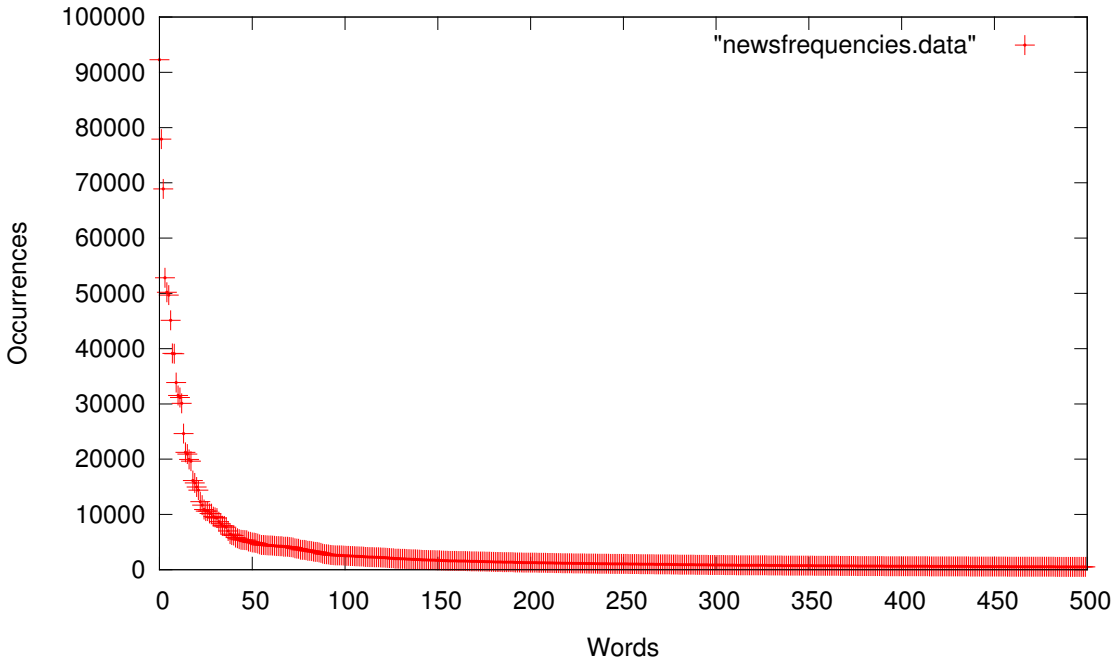

Figure 3.3: Term frequency distribution of the news media corpus.

3.3 Latent Dirichlet Allocation

LDA represents a document d in a document collection D as a mixture of topics. The topics are represented by probability distributions consisting of words with certain probabilities. It is a generative model in which hidden variables - topics - give rise to the observed variables – the words of a document. A given topic, therefore, has different probabilities of generating different words. For instance, a topic such as FISCAL POLICY would, according to LDA, have high probabilities of generating words like *taxes*, *funding* and *public* and low probabilities of generating words like *pets*, *rainbows* and *strawberries*. Note that this is a constructed example. Actual output from LDA does not have a label such as FISCAL POLICY.

3.3.1 Model

LDA is a three-tier hierarchical model in which each document of a corpus is modeled as a finite set of topics, where each topic is represented as a probability distribution over a vocabulary, i.e. the words of the corpus. This means that a topic, according to this model, is the sum of the items in the vocabulary that are assigned a high probability by the model. Using these topics, we can thus describe the content of a document. The generative process of the LDA can be defined as follows:

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

where d is a document in a document collection D which is represented by a topic distribution $\theta_{1:K}$. A topic θ_k is, in turn, represented by a probability distribution over a finite vocabulary in which all of the words of a given corpus are represented. The proportions of the topic distribution for the dth document is denoted θ_d. The topic assignment for a document d is denoted z_d, where $z_{d,n}$ is the topic assignment for the nth word in the dth document. A topic assignment is an instance of a topic and should not be confused with the topic itself. Lastly, the observed words from a document d are denoted w_d, where $w_{d,n}$ is the nth word in the dth document.

As previously stated, documents are generated by hidden variables. This is done by making an independent sampling of a topic assignment z for each word from θ_d, making $z_{d,n}$ dependent on the topic proportion for the dth document, θ_d. Further, the observed variable $w_{d,n}$ is dependent on the topic assignment $z_{d,n}$ by independently sampling the word $w_{d,n}$ from the current topic $z_{d,n}$[6].

The equation above can be explained by the following procedure for each d in D:

1. Sample a random topic distribution according to a Dirichlet distribution over a finite set of topics for document d.
2. For each topic, pick a distribution of words from the Dirichlet for that topic.
3. For word w_i in d:
    a) From the distribution of topics selected for d, sample a topic.
    b) From the probability distribution selected for that topic, pick the current word.

This model gives us the joint probability distribution of the topic model. Please note that we assume symmetric Dirichlet priors.

3.3.2 Inference

Using the joint probability distribution, we can, given a document collection D and an n number of topics to discover, learn the latent topics of a corpus by examining the posterior distribution of the topics θ, the topic proportions β and the topic assignments z. That is, we compute the conditional distribution of the topic's structure given some observed set of documents. The posterior is

[6]The fact that the model makes independant samplings specifies the bag of word assumption of the LDA model.

defined as follows:

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

In this equation, the numerator corresponds to the joint probability distribution. The denominator corresponds to the marginal probability of the observed variables, i.e. the probability of seeing the observed corpus under any topic model. However, the posterior distribution cannot be directly computed, as the number of possible topic structure is exponentially large and intractable to compute. We therefore have to make an approximation of the equation above by forming an alternative distribution over the latent topic structure that is adapted to be close to the true posterior.

The most commonly used method of approximation is Gibbs sampling. It is a Markov Chain Monte Carle method of obtaining a sequence of observations, whose limiting distribution is equal to the posterior distribution. The Markov chain, a dependent sequence of random variables, is defined on the hidden topic structure for a given document collection. Samples are then collected from the limiting distribution of this chain. The posterior distribution is subsequently approximated from these samples.

Another popular method of approximating the posterior distribution is by the use of various variational Bayes inference method. These methods use optimization to find a distribution over the latent variables that are close to the true posterior distribution. The inference problem is, using this method, transformed into an optimization problem.

If we allow a document to use an arbitrary number of topics, topic modeling with LDA is an NP-hard problem (Sontag, Roy, 2011, p. 8). Therefore, efficient algorithms for the approximation of the posterior distribution are essential.

Gibbs sampling can be performed in polynomial time, but it requires a full pass through the corpus for each iteration of the algorithm and therefore does not scale to big data sets. The traditional variational Bayes runs in constant time, but also suffer from the serious flaw of having to pass through the corpus for each iteration. Therefore, an online variational inference algorithm developed to address this precise issue is utilized in the current experiments. Online variational Bayes inference performs as well or better than the standard variational Bayes in a fraction of the time.

Online variational inference for LDA

Traditional variational Bayes runs in batches. It iteratively analyzes all documents of a corpus while continuously updating the parameters that best describe the topic model. As previously stated, online variational Bayes is significantly faster than the traditional method and is therefore the algorithm that will be used in the current experiments. Online variational Bayes uses an online stochastic optimization on the variational objective function (the function that decides whether q is a good approximation to p) with a natural gradient step.

Online variational inference updates observation-specific parameters using coordinate ascent to make the variational distribution close to that of the true posterior distribution. This procedure is equal to that of the traditional

variational inference. However, when updating the global parameters of the model, stochastic gradient ascent is utilized, where a noisy version of the gradient is followed to update the global parameter value - as long as the noisy gradient is equal to the true gradient in expectation. From here, smaller and smaller steps in the direction of the noisy gradient will be taken and thus resulting in their convergence. For a detailed accounting of the algorithm, see Hoffman et al. (2010).

3.4 Temporal chi-squared (χ^2) topic modeling

Chi-squared (χ^2) statistics is a method commonly used in descriptive statistics to characterize two groups of observations. It is a measurement of how a set of results compare to the expectations of those results. χ^2, very naively stated, finds and displays unexpected distributions. Using χ^2, we can find out if a distribution of terms is skewed in some way.

The criteria of the data used in the calculation of a χ^2 statistic are the following: random, raw, mutually exclusive and drawn from independent variables. χ^2 is computed as follows:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \tag{3}$$

where O_i is the *observed* frequency of a term and E_i is the *expected* frequency of a term. Thus, χ^2 sums the differences between the expected frequencies of a bigram and the observed frequencies of a bigram.

In corpus linguistics, χ^2 can be described as a statistical approach to characterizing the distribution of terms in a given corpus. We can make use of this in a topic modeling context. The traditional usage in computational linguistics is term similarity statistics, which allows a frequency distributions of terms to be tested and compared to the expected distribution of terms in a corpus.

The intuition behind the use of χ^2 statistics in a topic modeling setting is that unexpected results can help determine important topics during a given period of time. That is, if information about the period of time is available.

The given corpora come with a generous amount of metadata. Parliament texts are accompanied by information about speaker, party affiliations, time and date among other things. Social media and news media texts also are also accompanied by information about the context in which they are produced: time, date, domain and user identities. Depending on user settings, information about the geographical location of the poster is also available.

Using the temporal metadata at hand, it is possible to simulate a stream of data. By tagging the corpora using discrete temporal entities symbolizing from millennium to decade to a fraction of a second (given that we have that kind of data at hand), we can determine what terms have been especially important during the given period of time using χ^2.

The intuition behind the use of χ^2 in a topic modeling setting is thus that the temporal information about a given text can constitute the glue that collocates related words, i.e. topics.

The procedure is performed by, initially, calculating an observed value matrix, where w_i corresponds to a term in the corpus and t_i to a time period of interest.

	w_i	$\neg w_i$	$\sum$
t_i	w_i,t_i	t_i, $\neg w_i$	t_i
$\neg t_i$	$\neg t_i$,w_i	$\neg w_i$, $\neg t_i$	$\neg t_i$
$\sum$	w_i	$\neg w x_i$	

The expected distributions are subsequently calculated based on the sums of the observed distributions of the time period of interest. Using these values, an expected value matrix is computed. If this matrix is distinctly separate from the observed value matrix in regards to the time period of interest, the term has a skewed distribution and is thus regarded as topical for the period at hand.

3.5 Evaluation approach

Evaluation of the topic modelings systems at hand is performed manually. This is due to the lack of a gold standard but also due to the difficulties of implementing a proper and pragmatic evaluation system as discussed in 2.5.3.

Evaluation is performed by comparing experimental results to a gold standard *when a gold standard is available*. In cases where no gold standard is available, an ocular inspection of the results is performed. The only existing gold standard is that of the parliament corpus. This is due to the fact the parliament corpus is the only corpus for which manual annotation of documents has been possible as it is significantly smaller than the news and social media corpora. The parliament gold standard is comprised of manually annotated topics from the parliament test set. As stated in section 3.1.1, the parliament test set is comprised of documents collected in the parliament during week 43 of 2012. To evaluate the topic models at hand, the gold standard was manually created by labeling the test set with manually annotated documents. Random samples of the gold standard were verified by human evaluators. Samples of the gold standard are displayed below.

Patrik Björk (S) on unemployment benefits (2012-10-25)[7]	
Herr talman! Statsministern har i dag i sina svar gett en lite glättig syn på svensk arbetsmarknad. Om man däremot tar del av SCB:s senaste arbetskraftsundersökning ser man att den moderatledda regeringens misslyckande beskrivs både hårt och tydligt. Arbetslösheten ökar från en hög nivå. Långtidsarbetslösheten permanentas. Fler och fler kommer längre bort från arbetsmarknaden. Det de arbetslösa efterfrågar är två saker. Det är dels en aktiv arbetsmarknadspolitik som gör deras tid i arbetslöshet så kort som möjligt, dels en möjlighet att försörja sig och sina familjer under tiden. Regeringen, som statsministern leder, är ansvarig för den hopplösa situation som allt fler nu lider under. Tidigare har statsministern hänvisat dem som drabbats av arbetslöshet till föräldrar, eventuella partner eller kommunens försörjningsstöd. Är detta fortfarande statsministerns svar till den ökande grupp som nu ställs utanför arbetsmarknaden? Eller skulle statsministern här i dag kunna ge besked om att det är dags att höja ersättningen i arbetslöshetsförsäkringen och därmed underlätta för den som befinner sig mellan två anställningar?	Mr Speaker! The prime minister has today shared his somewhat unapprehensive view of the Swedish job market. If you, on the contrary, examine the latest workforce review by SCB, the failures of the government lead by Moderaterna are clearly illustrated. Unemployment levels are increasing from already high levels. Long term unemployment is consolidating. More and more people are getting further away from the job market. What the unemployed are asking for are two things - an active employment policy that makes their time in unemployment as short as possible and an opportunity to support themselves and their families meanwhile. The government, lead by the prime minister, is responsible for the hopeless situation that more and more people are suffering from. Previously, the prime minister has referred those affected by unemployment to their parents, their partners or the economic support provided by their municipalities. Is this still the prime minister's answers to the growing group that are now placed outside of the job market? Or could the prime minister, here and now, give notice about increasing unemployment benefits, and thereby ease the situation for those in between jobs?
Topics: Labor policy	

Annicka Engblom (S) on environmental policies (2012-10-25)	
Herr talman! Det övergripande målet för miljöpolitiken är ju att till nästa generation kunna lämna över en miljömässigt bättre värld än den vi lever i nu. På åhörarläktaren i dag har vi många företrädare för nästa generation, bland annat två tolvåriga flickor som liksom många barn i den åldern funderar mycket över sin framtid, och då i synnerhet miljön. De funderar och bekymrar sig väldigt mycket över bland annat utsläpp från trafiken, tillståndet i Östersjön, som de bor och verkar vid, och om hur vi ska jobba tillsammans med andra länder för att kunna förbättra miljön i hela världen. Det är sådant som rör sig i tolvåringars vardag. Jag vill ta tillfället i akt att bli talesman för dem i dag och å deras vägnar ställa frågan till statsministern om hur Sveriges regering jobbar för att förbättra deras framtida miljö. De undrar hur utsläppen ska minskas och hur övergödningen i Östersjön ska bekämpas.	Mr Speaker! The general goal of environmental policies is to be able to hand over an environmentally better world to coming generations. Listening today are representatives of the next generation. Among others, two twelve year old girls who, as many children in their age do, are thinking about the future, and about the environment in particular. Among other things, they are worried about traffic pollution, the Baltic Sea, which they live by, and how we are going to collaborate with other countries to improve the environment in the whole world. These are the sort of things that twelve year olds think about. I want to take the opportunity to be their spokesperson today and on their behalf pose a question to the prime minister about how the Swedish government are working to improve their future environment. They are wondering how pollution is going to be reduced and how the overfertilization of the Baltic Sea is going to be stopped.
Topics: Climate	

Kent Ekeroth (SD) on youth punishments (2012-10-23)	
Herr talman! Det är en något annorlunda gång i denna interpellationsdebatt än vad jag är van vid. När det gäller att redan från början förhindra brott har vi flera förslag på området. Den invandringspolitik som vi i dag bedriver bidrar enligt alla undersökningar till att öka brottsligheten i det här landet. Till och med när man tar hänsyn till socioekonomiska faktorer skapar den invandringspolitik vi för i dag de facto fler brott än vi hade haft om den inte funnits. Hade vi haft en begränsad invandring, och dessutom inte fört det som kallas integrationspolitik eller mångkulturell politik, skulle exempelvis de hedersrelaterade brotten, brotten mot kvinnors rätt till sin egen kropp och sexualitet, inte begås av de invandrargrupper som kommer hit. Vi skulle ha sluppit de övergrepp som tjejer i många fall blir utsatta för eftersom man har en annan uppfattning om tjejers värde, en uppfattning man har med sig från utlandet när man kommer till Sverige. Hade vi tagit emot en begränsad mängd invandrare och dessutom assimilerat dem i den svenska kulturen och det svenska tänkesättet skulle vi ha vunnit mycket och kunnat bekämpa den sortens brottslighet. Som vi konstaterade tidigare är andelen återfall bland dem som döms till sluten ungdomsvård 78 procent, och bland dem som sitter i fängelse är andelen också mycket hög. Jag tror att man skulle kunna komma undan många av återfallen och den återupprepade brottslighet som människor utsätts för varje dag om man dömde till mycket hårdare straff. Då skulle förövarna inte få chansen till återfall i första taget. Hur ger det rätt signal till ungdomar, brottslingar eller som i det här fallet våldsbrottslingar i landet i dag att det är fel att göra det när de som i Kortedalafallet kom ut? De fick inte en enda dag bakom lås och bom, utan de kom ut direkt. Hur signalerar det att samhället inte accepterar att sex personer står och hoppar på en medvetslös 61-årig sjukpensionär? Men dagens straffnivåer skickas inga sådana signaler. Det måste vi göra någonting åt.	Mr Speaker! The jargon of this interpellation debate is somewhat different than I am used to. We have several suggestions concerning the prevention of crime in this country. The current immigration policies are, according to all studies, contributing to the increase of crime in this country. Even when we take socio-economic factors into consideration, the current immigration policy is creating more crime than we would have had if it had not existed. If we had restricted immigration, and if we would not have pursued integration politics or multicultural politics, honor related crimes, the crimes on women's right to their own body and sexuality, would not be committed by the groups of immigrants that are coming here. We would not have to deal with the assault that girls in many cases are being exposed to because immigrants have another perception about the value of girls - values imported from other countries. If we would have accepted a limited number of immigrants and assimilated them into Swedish culture and the Swedish way of mind, a lot would have been won and we could have fought that kind of crime. As previously stated, the proportion of relapses among convicted juveniles is 78 percent and among those in jail, the proportion is also very high. I think that a lot of those relapses and the reappearing crime that people are exposed to every day could be escaped if we had more severe punishments. The perpetrators would not get the chance to relapse. How can we send the right signals to youths, criminals, or in this case violent criminals, that it is wrong to to do it when those in the Kortedala case got out? They did not get one single day behind bars - they got out immediately. How does this signal to society that they do not accept that six people are ambushing an unconscious 61 year old with an early retirement pension. With today's punishments, no such signals are being sent. We have to do something about that.
Topics: Crime, youth, immigration	

The gold standard exhibits 14 topics. These topics are listed in table 3.1. An intuition of the gold standard, which is in line with the output of LDA and χ^2 tests, is that a document can exhibit multiple topics. For example, an entry about taxation of air traffic can thus be labeled with topics such as *fiscal policies* as well as *air travel* and/or *transport*.

Gold standard topics
Cash handling
Climate
Crime
Data privacy
Defense
Defense music
European Union
Export
Finances
Foreign affairs
Health care
Immigration
Labor policy
Youth

Table 3.1: The manually created topics extracted from the parliament test set week.

It is, naturally, preferable to have a standardized and fully automatic evaluation approach. As discussed in section 2.5.3, it is not at all obvious how such an evaluation should be designed. It is therefore crucial to perform manual evaluations in order to draw conclusions about necessary characteristics of an adequate evaluation.

4 Experiments

4.1 Experimental design LDA

Splitting the corpus into a training set and a testing set is a common step in evaluation of the performance of a learning algorithms. For supervised learning algorithms, the model is trained on the training set and evaluated on the test set using a gold standard. For unsupervised learning such as topic modeling with LDA, the procedure is not as clear cut. It includes training the model on a training set and then manually inspect the topics generated by the training set and the topic assignments on the test set.

The corpora used in the LDA experiments have, thus, been divided into a training and a testing set. The testing subset of each corpus was randomly selected. A summary of the corpora can be found in tables 4.1, 4.2 and 4.3.

Parliament corpus

	Training set	*Test set*
Time period	Feb '12-Feb'13 (excl. week 43)	Week 43 '12
n documents	7077	193
n words	7 million	64,000
Avg.doc. length	1029	332

Table 4.1: Descriptive statistics of the parliament corpora.

Social media corpus

	Training set	*Test set*
Time period	Oct '12	Nov '12
n documents	10 million	10 million
n words	36 million	36 million
Avg.doc. length	35.9	35.8

Table 4.2: Descriptive statistics of the social media corpora.

Apparent from tables in 4.1, 4.2 and 4.3 is that the corpora differ significantly in size from the parliament test set of 193 documents and approximately 64,000 words to the social media corpora in which each sub-corpus consists of approximately 10,000 documents and 36 million words. Further, the difference between the average number of words per document for the parliament training

Editorial media corpus

	Training set	*Test set*
Time period	August '12	Week 28 '12
n documents	67,000	38,000
n words	3 million	6 million
Avg.doc. length	81.6	88.7

Table 4.3: Descriptive statistics of the editorial media corpora.

and test set should also be noted; the average document length of the training set is 1029 words and the average document length of the test set is 332 words.

The above-mentioned features of the corpora apart, they are fully adequate for the experiments at hand. The parliament corpora is small-scaled in comparison to the editoral and social media corpora. However, this is in line with the nature of the given domains; it is an inherent quality of the domains from which the texts are collected.

In the experiments at hand, the parliament corpora have been collected from original sources and the editorial and the social media corpora have been patched together from multiple middlemen. An inevitable consequence of not collecting texts from original sources is that control of the texts - the selection, the collection and basic processing - is renounced. This means that there might be problems with the selection of sources, data loss and/or multiple other unknown difficulties. Implied by this is that the control of the corpora is not total and a full and detailed description of the content of the "patched" corpora cannot be given. It is therefore not possible to perfectly duplicate experiments conducted with these corpora. Nevertheless, this is an inherent feature in big data research. When the rule of thumb is quantity over quality and texts are collected from multiple elusive sources on the internet as opposed to a small, well-annotated and carefully chosen selection of language snippets of special interest to a certain line of research, the results are never perfectly reproducible. Nonetheless, it should be safe to say that if similar texts were collected, the outcome of experiments such as the ones conducted in this study, should be similar to those produced here.

4.1.1 LDA experiments

The following section describes the experimental scheme of the LDA topic modeling. These experiments test the tuning of the initial and basic model parameters. These parameters include the selection of the optimal number of topics as well as the number of iterations over the corpus. The significance of the optimal number of topics should be obvious. The number of iterations over the corpus is important to compute corpus likelihood. The number of iterations will, thus, make an impact on the final topics. On a detail level, these experiments aim to answer the following questions: *How many topics do we need to describe a given corpus? How many iterations over the corpus do we need to describe a given corpus?* The experiments also seek to answer the implied question of whether *we can successfully describe a corpus in terms of topics with*

LDA.

The experimental scheme of the LDA experiments can be found in table 4.4.

LDA experimental scheme

Experiment	*Topics*	*Iterations*
I	10	1000
II	20	1000
III	30	1000
IV	40	1000
V	50	1000
VI	100	1000
VII	500	1000
VIII	100	2000
IX	100	5000

Table 4.4: The experimental scheme of the LDA experiments. *Topics* refers to the number of topics defined in the experiment at hand. *Iterations* refers to the number of iterations through the corpus defined in the current experiment.

4.2 Temporal χ^2 experiments

As opposed to the LDA approach to topic modeling, where the procedure is divided into two steps (i.e. topic creation and subsequently document labeling), the χ^2 method is drastically different. Instead of topic creation and document labeling, χ^2 gives a list of important terms and associated terms for a given corpus. This means, for instance, that we cannot label documents as belonging to a certain topic. However, that is not a requirement of the planned system and therefore does not matter. In fact, it may be regarded as a strength of the χ^2 approach since an implication of this is that the algorithm is significantly less time consuming.

Table 4.5 displays the experimental scheme of the χ^2 experiments. As the χ^2 experiments, as opposed to the LDA experiments, are tested on all data sets (parliament, social media and news data), the experimental scheme is true for all corpora. This means that the four experiments displayed below will be performed for all corpora, effectively multiplying the four experiments by three to make twelve experiments.

The *experiment* column of 4.5 simply gives the name of the experiment. *Target*, however, refers to the target tag on which the experiments has been modeled. It is a query which denotes a period of time – for example, a week or a specific date.

Chi-squared experimental scheme

Experiment	*Target*
I	$\text{week}_{(i,(j))}$
II	$\text{date}_{(i,j)}$

Table 4.5: The experimental scheme of the chi-squared experiments. *Target* refers to the target tag, which may consist of a symbol representing a week or a date.

Please keep in mind that the χ^2 experiments are by nature exploratory as the χ^2 approach has never been previously applied to a topic modeling setting.

Further, the χ^2 experimental procedure requires no distinction between training and testing sets. Therefore, the full corpora (training and test set) has been used for these experiments.

5 Results of the topic modeling experiments

The results of the two different topic modeling methods are reported separately. Initially, the LDA experiments are accounted for. Secondly,the output of the temporal χ^2 is given. All results are displayed in Swedish alongside an English translation. Some notes and contextualizations are given to help make sense of the results fore the unoriented.

5.1 LDA results

LDA generates a set of n topics from the training set and subsequently uses these to assign topic tags to the documents of the test set. The results of the LDA experiments are divided into two sections. In section 5.1.1, the underlying topics discovered by the LDA algorithm under different conditions are presented. Section 5.2 introduces the tagging of the test set according to the experimental scheme presented in table 4.4.

5.1.1 Topic construction

The topic number effect

In the LDA model, topics are represented as probability distributions over words. This means that a topic is represented by an n number of words that all have a high probability of occurring in the context of that given topic. Displayed below are a selection of topics from some of the LDA experiments described in table 4.4.

For human readability reasons, three randomly selected topics from each experiment round are presented here – independently of how many topics are generated in the specific experiment. As previously mentioned, the tested parameters include number of topics and iterations through the corpus respectively.

Table 5.1 displays a sample of the underlying topics of the parliament training set according to an LDA model trained using 10 topics and 1000 iterations through the corpus. 1000 iterations is a standard value that will be used throughout the first round of experiments.

topic_i	**word distribution**
topic_{48}	storinvestera, formfråga, tuppfäktandet, gallfeber, svischar, fartvinden, måttligare, hojar, bockstyre *to make a big investment, matter of form, cock fighting, nuisance, swoosh, airspeed, more moderate, bikes, drop bar)*
topic_{13}	allenast, siffrornas, rubbet, fenestra, måttstockar, sippra, ungdomlig, lönsamheten, gluggen, präntade *(loneliest, the number's, the lot, fenestra, scales, seep, youthful, the profitability, opening, printed)*
topic_{72}	huvudsyssla, övernattningsmöjligheter, gårdsbaserad, jordbruken, distributionskanaler, hästsektorn, djurkontroll, förhandlingsvilja, djurfabriker, upphandlingskraven *(main occupation, possibilities to stay the night, farm based, the farms, channels of distributions, horse sector, animal control, will to negotiate, animal factories, procurement demand)*

Table 5.1: **Experiment I**: A small sample of the underlying topics of the parliament training set set according to LDA with 10 topics and 1000 iterations through the corpus.

The results of the experiment above cannot instantly be seen as a success, nor can they be seen as a failure. There is at least one coherent set of semantically related terms, which, as previously stated, serves as our definition of a topic in the context of this thesis. Consider topic_{72} that contains words such as *farm based*, *the farms*, *channels of distributions*, *horse sector*, *animal control* and *animal factories*. These words tell a distinct story about the topic's agricultural relatedness. Even the remaining words that are not immediately related to agriculture or animals - *main occupation* and *possibilities to stay the night* - can easily be found in a document about farms or animal husbandry.

On a less positive note, there are some rather peculiar terms starred in the topics above – topic_{72} is, unfortunately, not representative of all the topics. The other topics do not contain words that are as closely related to each other as the words in topic_{72}. See for example topic_{48} that holds only a couple of words that seem to be related: *bikes* and *drop bars*, as well as *swoosh* and *airspeed*. Further, it seems unlikely that a topic in which bike-related words are prominent is important enough to describe a tenth of the Swedish parliamentary activity during an entire year.

It is obvious that the text is not lemmatized or stemmed. The effect of this choice will be discussed later on.

In experiment II, the predefined number of topics is set to 20. The number of iterations through the corpus is static at 1000 passes. The results, i.e. the

sample of ten topics, can be viewed in table 5.2.

topic_i	**word distribution**
topic_2	industribransch, genusfrågor, inplanterat, språkgränser, förtroligt, sanslöst, samarbetsvillig, tillstånd, vindkraftsprojekt, framtidsutveckling *(branch of industry, gender issues, planted, linguistic boundaries, confidentially, outrageous, co-operative, permission, wind power project, future development)*
topic_9	etnifiering, underutnyttjad, stockholmsborna, kloten, mångbottnad, tåglinjen, diskrimineringsombudsman, utgiftsramen, profeten, återuppleva *(ethnification, under exploited, the citizens of stockholm, the spheres, multilayered, train line, discrimination ombudsman, expense frame, the prophet, re-live)*
topic_{18}	publiksport, idrottslyftspengarna, förespegla, normalgraden, jobbskatten, livstidsutvisning, omsorgsfrågor, majsen, kapats, testikel *(crowd-drawing sport, sporting grant money, hold out the prospect of, normal degree, income tax, lifetime expulsion, welfare issues, the corn, crosscut, testicle)*

Table 5.2: **Experiment II**: A small sample of the underlying topics of the parliament training set according to LDA with 20 topics and 1000 iterations through the corpus.

There are no obvious observable sets of semantically related terms in the results of experiment II – especially not in line with those of the agricultural topic in experiment I. However, most words have some political significance. These conclusions are true also for experiment III, IV and V. The results of these experiments will, therefore, not be discussed here. Samples can, however, be found in the appendix.

The results of experiment VI can be found in table 5.3.

topic_i	**word distribution**
topic_{40}	utdelningsmöjligheter, felstegen, terrorhandling, erfoderligt, piggade, ättestupa, klimattryck, upptaxerar, lastbilschaufförerna, diplomat *(profit possibilities, misdemeanour, action of terror, required, sharped, ättestupa*[1]*, climate stress, up value, truckers, diplomat))*
topic_2	pulsåder, banhållningen, kurera, företagarna, infrastrukturplan, förskottera, vän, pendlarkortet, tokyobörsen, ineffektivare *(artery, track maintenance, cure, entrepreneurs, infrastructure plan, advance, friend, commuter ticket, tokyo market, less effective)*
topic_{76}	oomstritt, nioårig, ingångscertifikat, sexårig, primärskola, urvalsgrundande, dn-artikeln, beskrivande, konkurrenskrav, sekundärskolesystem *(indisputably, nine year old, entry level certificate, base of selection, primary school, dn*[2] *article, descriptive, competition demand, secondary school system)*

Table 5.3: **Experiment VI**: A small sample of the underlying topics of the parliament training set according to LDA with 100 topics and 1000 iterations through the corpus.

As visible in table 5.3, the results of experiment VI are similar to previous experiments – most of the words that occur have a political significance. There are also, in fact, some examples of semantic relatedness, for example *primärskola* (primary school) and *sekundärskola* (secondary school) in topic_{72}. It is, however, not possible to attribute an obvious and general label to any of the topics.

A sample of the results of experiment VII is presented in table 5.1.1.

topic_i	**word distribution**
topic_{243}	fartygsägarna, damanaki, tillskrivit, eu-vatten, fiskemissionären, eu-fartyg, brysselarbetet, sektorsstöd, utvecklingsbrasch, underlagsmaterial *(ship owners, damanaki, attributed, eu waters, fishing missionary, eu ships, the work in brussels, sector support, developing branch, decision basis)*
topic_{136}	orustjord, medborgarfokus, blanketтflora, kunnigare, arkiverat, partipolitiken, varghonan, hemtjänstens, sandmagasin, korruptionshanteringen *(the soil of orust[3], citizen focus, form flora, more knowledgeable, partisan politics, female wolf, home care, sand repositories, corruption handling)*
topic_{498}	anpassningarna, amount, förutsedda, framemot, uppvärmningssidan, antagandeperiod, aauöverskottet, självtillräcklighet, aausystemet *(adaptations, amount, predicted, towards, heating side, the aau surplus, self-sufficiancy, the auu system)*

Table 5.4: **Experiment VII**: A small sample of the underlying topics of the parliament training set set according to LDA with 500 topics and 1000 iterations through the corpus.

Once again, a semantically coherent set of words appears in topic_{243}. It contains words such as *fartygsägarna, eu-vatten, fiskemissionären, sektorsstöd and eu-fartyg* – all fishing-related words. The remaining topics, however, shows no such coherency. It should be obvious to most that the output of the experiment displayed in table 5.1.1 shows no improvement but neither does it show deterioration.

In conclusion, there does not seem to be an optimal number of predefined topics (for this specific data set). As the results are far from promising in combination with the fact that the LDA algorithm is very time consuming, no further experiments on varying the number of predefined topics will be performed.

Next, experiments to test the effect of varying the number of iterations through the corpus will be performed. To do this, a value to represent *the optimal number of predefined topics* for the data set is necessary. As there is no obviously no optimal number, the *least objectionable* number will be used. According to previous experiments, this number should be 10. However, this is unreasonable

as ten topics cannot possibly describe one full year of parliamentary activity. Therefore, we will use 100 as the predefined number of topics. The standard value of the predefined number of topics during the next round of experiments with LDA will be 100.

As previously mentioned, another important parameter of the LDA model is the number of passes through the corpus that the algorithm makes. In the previous experiments, 1000 iterations has been used as the standard number of passes. As can be viewed in table 4.4, the following round of experiments will evaluate the output of the LDA algorithm on the parliament data set when setting the number of iterations through the corpus to 2000 and 5000 respectively.

Table 5.1.1 displays a sample of the ten first topics when 100 is the number of predefined topics and 2000 is the number of iterations through the corpus.

topic_i	**word distribution**
topic_{29}	byggutbildningar, detaljplaneprocess, radera, tvåveckorsperiod, utarbetning, smågrisdödligheten, festivaler, kondomer, chicagokonvention, fniss *(educations in construction, detailed plan process, delete, two week period, worn-out, piglet mortality, festivals, condoms, the chicago convention, giggle)*
topic_{86}	förstasidorna, pådriven, eurokurser, valutaexperiment, centraliseringens, emuländernas, bortblåsta, anfallen, växelkursen, samhällsinflytande *(first pages, actuated, euro rate, currency experiment, centralization, emu countries, vanished, the attacks, the exchange rate, societal influence)*
topic_5	krönikör, dogmernas, fiskebudget, breddande, produktionsmetoderna, t-shirts, finansinstitutionerna, spektualitionskapitalism, jympadojor, facebookandet *(chronicler, the dogmas, fishing budget, broadening, means of production, t-shirts, financial institutes, speculation capitalism, sneakers, facebooking)*

Table 5.5: A sample of the underlying topics of the parliament training set set according to LDA with 100 topics and 2000 iterations through the corpus.

Table 5.1.1 shows no observable improvements or deteriorations. Topic_{86} contains some financially related terms.

topic_i	**word distribution**
topic_{56}	produktionsföretagen, reseavdraget, sätts, bränslepris, styrfråga, betalsystem, vinstniåverna, bränslepriserna, valpskatt, metalliskt *(production corporations, travel deduction, is placed, price of fuel, management issue, paying system, profit levels, the fuel prices, financial transaction tax, metallic)*
topic_{24}	varghonan, ryggskott, huvudskyddsombud, grusats, idrottsklubb, dominanta, småriken, jasidan, doctor, mask *(the female wolf, lumbago, head safety representative, dashed, sports club, dominant, small countries, yes side, doctor, mask)*
topic_{57}	svältkatastrofen, personkult, militariserade, frammana, vapenframställning, kärnvapensprängningar, oberäkneliga, importembargo, importlagstiftningen, ärvas *(famine disaster, personality cult, militarized, conjure, weapon manufacturing, erratic, import embargo, import legislation, inherit)*

Table 5.6: A sample of the underlying topics of the parliament training set set according to LDA with 100 topics and 5000 iterations through the corpus.

In conclusion: an increased number of iterations through the corpus does not seem to improve the results of the LDA algorithm.

Parameter variation effects, in short

To summarize the LDA experiments on the parliament data, we can say that the results are unsatisfactory at best. The least unsatisfactory results are produced in experiment VI where the pre-defined number of topics is set to 100 and the number of iterations through the corpus is set to 1000. To evaluate the LDA inference, i.e. the tagging of documents, the topics generated by this experiment will be used.

5.2 Topic inference

Consider the following speech from the Swedish parliament:

This document is included in the parliament test set. It is properly assigned

Mikael Damberg (SD) on unemployment (2012-10-25)	
Herr talman! De som lyssnade kunde notera att statsministern inte svarade på frågan. Han svarade inte på frågan om regeringen står fast vid den prognos som innebär att arbetslösheten ska sjunka ganska dramatiskt de kommande åren. Det är hela fundamentet som regeringen har byggt sin budgetprognos och sin budget på. Då kan jag förstå varför regeringen valde bort stora insatser för att bekämpa arbetslösheten, valde bort stora insatser för att bekämpa ungdomsarbetslösheten och valde bort att införa en ungdomsgaranti. Man tror att arbetslösheten de kommande åren kommer att sjunka dramatiskt. Man är helt ensam om den bedömningen jämfört med andra ekonomiska bedömare, som Konjunkturinstitutet och Riksbanken. Alla andra drar en annan slutsats än den regeringen drar. Då blir slutsatsen att statsministern mindre än fyra veckor efter att han har lagt sin budget på riksdagens bord inte ens kan säga att de prognoser han lade fast i budgeten håller. Det är dags att införa nya åtgärder och större insatser, som vi har föreslagit i budget, för att bekämpa arbetslösheten. Kan vi räkna med det? (Applåder)	Mr Speaker! The listeners could note that the prime minister did not answer the question. He did not answer the question if the government are sticking to the prognosis that means that unemployment will, quite dramatically, sink in the near future. That is the foundation that the government has built their budget prognosis and their budget on. If this is the case, I can understand why the government chose to decrease support to fight unemployment and youth unemployment and to chose not to instantiate a youth guaranty. There is a belief that unemployment will decrease dramatically but this belief is not shared by other economical analysts, like Konjunkturinstitutet och riksbanken. Everyone else has come to other conclusions. The conclusion is thus that the prime minister, less than four weeks after he presented his budget on the parliament table, cannot even say that the prognoses he introduced holds in the budget. It is time to introduce the suggestions, that we have presented in the budget, to fight unemployment. Is this something we can count on? (Applause)

the topic LABOR POLICY, by multiple independent manual annotators in the gold standard. The LDA model with the "best" parameter settings judges that the topics which best represents the current document are $\text{topics}_{1,11,15,90}$. These topics are, in turn, represented by the following term distributions: topic_1: *fix, appendix, international security assistance force command, facades, turf, cutting device, managment level, the staffing companies, up sides, construction suggestion*; topic_{11}: *gmo technique, monarchy, climate guarantees, climat guarantee, big corporations, social security policies, reasons to be cautious, biodiversity, research conditions, mosebacke*; topic_{15}: *fish sticks, joyful leap, disability issues, the indigent sweden, defrosted, damaged, the regulator, scientists, freezer, tove*[4]; topic_{90}: *surround, health care, the health care, new car taxation, leverage, tightened, the meeting, huge fee, backed off, the volume*. What do these topics say about the topicality of the document? The conclusive answer is *not enough to draw any conclusions about the document at hand*. It should be obvious from this small manual evaluation

[4] A Swedish first name.

of the LDA output in relation to the gold standard that the output of the LDA model is borderline nonsense and that it cannot be mapped to the manually annotated documents on a semantic level, on a structural level or on a user friendly one (see 2.5.3). The output of the LDA algorithm is clearly lacking.

The observant reader should be aware that humanly created - relevant - topics of the gold standard are at best difficult to match and compare to the output given by the LDA model. At worst, they are impossible to match and compare. It is especially difficult if one would like do to this using automatic methods. For such a task to be possible, a title or label that represents the topic in question is crucial. Such a difficult task likely requires external resources, however. Consider the following terms from a randomly sampled topic[5]: *uranium access*, *social tourism*, *obligation to report*, *european union administrative authority*, *corruption handling*, *construction union*, *pine tank*, *surveillance cameras*, *domestic services*, *sand repository*. Is there a semantic centroid among these terms that can efficiently represent the entirety of the topic? Most likely, no. There are no obvious semantic connections between *access to uranium*, *social tourism* and *sand repositories*. Most likely, there are no substantial semantic connections at all between these terms. A computer might find hidden structures. Chances are that it could carry out the task in a better and a more meaningful way than a human. However, it seems unlikely and the results of the LDA model are in no way good enough to justify more work. In addition, this lies outside of the scope of the project, i.e. topic modeling – not topic labeling. Therefore, a manual evaluation of the results of the LDA experiments will be performed. This will include determining a semantic centroid among the terms of the topics *or* manually label the topic with a descriptive word. This will, in turn, be compared to the manually created gold standard topics.

The results of the experiments displayed here exclusively include parliament corpora. These are an essential and non-negotiable part of the project. Since the LDA model clearly does not perform well on this data, other types of media will not be considered in the present study. That is: no experiments with the LDA model are performed on news or social media corpora due to the unsatisfactory results of the LDA experiments on the parliament corpus.[6]

5.3 Results of the χ^2 experiments

The χ^2 topic modeling results of the parliament corpus is accounted for in section 5.3.1; the results of the social media corpus in section 5.3.2 and the news media corpus in section 5.3.3. The results are consistently presented as a table that includes terms, associated terms and notes about the aforementioned experiments. The notes serve to guide the reader through Swedish political discourse. The χ^2 results are subsequently manually compared to a gold standard *when a gold standard is available*. When no gold standard is available, the results are manually evaluated.

[5] Topic_5 from experiment VII.

[6] In personal communications between Jussi Karlgren and Bruce Croft, Bruce Croft has allegedly stated that LDA generally performs very poorly with texts collected from social media.

5.3.1 Temporal χ^2 parliament results

The results of the temporal χ^2 experiments on the parliament corpus will be manually evaluated against the parliament activity during the week from which the test set was collected[7], i.e. the gold standard.

The gold standard, as previously mentioned, is comprised of all documents produced in the Swedish parliament during week 43, 2012. As the gold standard is constructed for this one coherent week, only the results of the topic modeling of week 43 are evaluated against the gold standard. The topic modeling experiments of single days are evaluated manually.

Table 5.7 displays the results of the temporal χ^2 topic modeling of week 43, 2012. This corresponds to experiment I of the χ^2 experimental scheme in table 4.5. Some notes have been included to help guide the reader.

term (notes)
romsom (Member of Parliament)
lööf (Minister for Enterprise)
blekinge (Province)
panaxia (security company)
säffle (city)
kronoberg's (county)
european union
ekeroth (Member of Parliament)
adolfsson (Member of Parliament)
norman (Minister for Financial Markets)
carina (first name)
kronoberg (county)
swedec (Swedish EOD and Demining Center)
spain
elgestam (Member of Parliament)
kent (first name)
näringsminister annie (Minister for Enterprise)
statsminister fredrik (Prime Minister)
justitieminister beatrice (Minister for Justice)

Table 5.7: Displays the χ^2 topic modeling results of the parliament week 43, 2012.

The majority of the terms in the current table are names. More specifically, they are names of various members of the parliament. Note that ministers are also members of parliament. Lists of names of current and historical members of the Swedish Parliament can be found online. Using this resource, it is a trivial

[7] Week 43, 2012

task to filter out the names of members of the parliament from the parliament corpora and re-perform the topic modeling of the parliament test set week. The results corresponding to table 5.7 with all names removed are displayed in table 5.8.

term (notes and translations)
blekinge (province)
eksjö (municipality)
pcb (Polychlorinated Biphenyl)
panaxia (company)
säffle (city)
european union board
storbrittanien (United Kingdom)
eu (European Union)
danmark (Denmark)
mali
swedec (Swedish EOD and Demining Center)
kronoberg (county)
ecb (European Central Bank)
arbetslöshet (unemployment)
swedec's (company)
spanien (Spain)
katalonien (Catalonia)
sanktionsutredningar (sanctioned investigations)
försvarsmusik (defense music)

Table 5.8: Displays the χ^2 topic modeling results of the parliament week 43, 2012 with names of members of the parliament excluded.

This small alteration of the corpus gives significantly improved results.

The key terms of table 5.8 are not exactly overlapping with the gold standard. Nonetheless, they demonstrate many similarities. The most obvious are the key terms *european union* and *defense music*, which are verbatim to topics in the gold standard. However, there are other key terms in the temporal χ^2 output that are obvious to the informed reader. One example of this is the topic *panaxia*. Panaxia is a Swedish security company that, for a long period of time, was primarily engaged in the transportation of currency. It is thus easy to relate the key term *panaxia* to the gold standard topic *cash handling*. Another key term *kronoberg*, referring to the Swedish county Kronobergs län, is easily relatable to the gold standard topic *labor policy* as Kronobergs län is a part of Sweden where unemployment is very high.

Present in the temporal χ^2 output are also broader concepts such as *unemployment* that corresponds nicely with the broader gold standard topic *labor policy*. The same goes for the key terms *Spain* and *Catalonia*. These terms naturally fall into the very general topic *foreign affairs*.

The second round of the experiment tests the capability of the temporal χ^2 approach on smaller periods of times in the parliament, namely days as opposed to weeks as were tested in the previous experiment. As there is no gold standard present to evaluate this entity of time, experiments include two dates. This will hopefully help estimate the performance of the method. Dates have been randomly selected from week 43, 2012. Table 5.9 displays the results of experiment II, the temporal χ^2 topic modeling during the 25th of October, 2012.

term (notes and translations)
kalmar (city)
pcb (Polychlorinated Biphenyl)
panaxia (security company)
ssu (Swedish Social Democratic Youth League)
spanien (Spain)
eu-val (european parliament election)
volvo (car manufacturer)
katalonien (Catalonia)
eu (European Union)

Table 5.9: Temporal χ^2 topic modeling results of parliament on October 25th, 2012.

When decreasing the time window to include days instead of weeks, the temporal χ^2 topic modeling includes some key terms that occurred in the previous experiment (for example, *Spain*, *pcb* and *Panaxia*) but also some terms that are unique to the date at hand (for example, *Swedish Social Democratic Youth League*, *Volvo* and *European Parliament election*).

In table 5.10, the results of an additional χ^2 topic modeling experiment are displayed. The data of this experiment includes all documents produced in the parliament on the 25th of October, 2012.

term (notes and translation)
eu (European Union)
danmark (Denmark)

Table 5.10: Temporal χ^2 topic modeling results of parliament on October 26th, 2012.

The results displayed in table 5.10 might seem odd upon first inspection. However, very few documents were produced by the parliament on this specific day. In fact, the only documents collected from this day concerns Denmark and the European Union. Therefore, the output given by χ^2 from this particular date is in fact sound.

In summary, the results of the χ^2 topic modeling with parliament texts are excellent.

5.3.2 χ^2 social media results

As there is no gold standard available for the social media corpus, some additional experiments were performed with the social media corpora. An increased amount of experiment results will most likely give a better overview of the performance of the temporal χ^2 topic modeling algorithm. The present experiments include topic modeling of two weeks and two individual days.

Displayed in table 5.11 and 5.12 are the results of experiment I as described in table 4.5. Some notes about the results: entries that include names of non-public and semi-public individuals have been excluded to protect the privacy of these individuals. If the names of individuals are somehow especially relevant to the results, the entries have been included but the names of the non- or semi-public individuals have been encoded as $\mathtt{name}_i$. Surnames, however, are included in the table as it is largely impossible to trace these to specific individuals. Swedish names that might be unknown to the non-Swedish speaking reader are explicitly stated as names for clarity. Other excluded items are URLs. Some notes are included within parentheses to help orient the reader.

Table 5.11 displays the most important terms and their associated terms in social media during week 43 in October 2012.

term (notes and translation)
vaknasverige (swedenwakeup)
nuvoryn (diet pill)
savankotcecha (singer-songerwriter)
rögletimrå (ice hockey teams)
sats (clause)
bokbål (book burning)
bemöta (refute)

Table 5.11: Displays the χ^2 topic modeling results of social media week 43 in October, 2012.

Five entries have been excluded from the current table for the reasons stated above. Four of these entries are sports related. Out of these four, three of them concern video games. The many names in the current table are in most cases names of users of communities. Some key terms form obvious topics. The most salient example of this is sports. This, however, might not be obvious to the non-Swedish reader. The term *rögletimrå*, for example, is an automatic compound of the words *Rögle* and *Timrå*, referring to *Rögle BK* and *Timrå IK* respectively. The compound was created during the preprocessing of the texts at hand. Rögle and Timrå refer to Swedish ice hockey teams. The teams met in a match in October 2012, denoted as *Rögle-Timrå*. Further, an episode of the popular television show *Glee* aired in October 2012. This particular episode contained a song written by singer-songwriter Savon Kotecha, occurring in the

table as *savonkotecha* and *savon kotecha*. According to the current experiment, other topics of the week at hand include *book burning* and (grammatical) *clauses*.

Table 5.12 display the results of the χ^2 topic modeling of week 43 in October, 2012.

term (notes and translation)
youtube
management
korttidskontrakt (short term contract)
företagsinformation (company information)
hctimrå (ice hockey team)

Table 5.12: Temporal χ^2 topic modeling results of social media week 43 in October, 2012

.

Six entries have been excluded from the current experiment. Two of these are URLs and four of them include names of non-public individuals. They are all sports related. Salient topics of the week are management. During this week, news of the management of the investment company Creades was prevailing. This might have given rise to the topic. Further, a short term contract was assigned to the Swedish ice hockey player Niklas 'Bäckis' Bäckström. The ice hockey team HC Timrå is also discussed.

A closer examination of a subset of table 5.12 can be found in table 5.13, which displays the most important terms in social media on October 29th, 2012. The chosen dates have been randomly sampled from the test set.

term (notes and translation)
illafy
workoutoutfit
ayda
anginas
dietpills
dreamzz
wedding
mom
lose
remö y
twothousand
stöddemonstration (support demonstration)
fng

Table 5.13: Displays the χ^2 topic modeling results of social media day August 29th, 2012.

Four entries were excluded from the experiment as they include names of non-public individuals. Initially, it should be noted that these results are more noisy and less clear cut than the previous experiments that included a larger time window. For example, it is not at all clear what *lose*, *twothousand* or *fng* refers to. The only clear and useful topic of this particular day is the Pussy Riot support demonstration, which took place in October, 2012.

Table 5.14 contains the most important terms of August 23rd in social media, according to the corpora at hand.

term (notes and translation)
källström (ice hockey player)
tunnelbanebiljett (subway token)
försöker (tries)
gfga
sökförslag (search suggestions)
nbspbr
borgensman (warrantor)
spärrvakt (subway gateman)
bengals (sports team)

Table 5.14: Displays the χ^2 topic modeling results of social media day August 23rd, 2012.

Five entries were excluded due to privacy issues. All of these included sports, mostly video games. This particular day is dominated by sports-themed key terms. *Källström* most likely refers to football player Kim Källström. *Bengals* refers to the Cincinnati Bengals, an American football team. Other important terms of August 23rd, 2012 refer to public transportation. Subway tokens and subway gate men are discussed.

In summary, the results of the social media χ^2 topic modeling are more difficult to make sense of than the parliament topic modeling. It is not, however, impossible. Some distinct topics can be noted. On the other hand, the key terms contain a great deal of noise originating from, for example, spam.

5.3.3 χ^2 editorial media results

As with the social media corpora, no gold standard is available for the news media corpora. Additional tests will thus be performed to evaluate the performance of the χ^2 model on the news media corpora. The experiments will include topic modeling of two different weeks and two separate dates respectively. Table 5.15 displays the results of week 28, Month, 2012.

term (notes and translation)
taikon (romani name)
exeotech (company)
vildagliptin (drug)
everolimus (drug)
novartis (company)
envirologic (company)
kivisaari (name)
mirren (name)
dalakraft (company)
karaterally
c2sat (company)
tajik (name)
hcl (company)
cefour (company)
volta (overturn)
glu (company)
capillum (company)

Table 5.15: Displays the χ^2 topic modeling results of editorial media week 28, 2012.

The results of the temporal χ^2 topic modeling of news media are notably richer than that of the parliament or social media experiments. The most distinct topic is that of drugs.

Evorolimus is an immunosuppressant drug that prevents the rejection of organ transplants and treatment of renal cell cancer and other tumors. *Vildagliptin* is an anti-diabetic drug. *Novartis* is a multinational pharmaceutical company.

Another distinct topic is business. *Exeotech*, *envirologic*, *dalakraft*, *c2sat*, *hcl*, *cefour*, *capillum* and *glu* all refer to established companies. *Kivisaari* is the name of the president of mobility services at Telia Sonera. It is possible to find the motivation for why these companies are listed as key terms in the results of the temporal χ^2 topic modeling experiments using online resources. Some of these motivations are listed here. Exotech was subject to a name change from C2SAT Holding to Exeotech at the end of July, 2012. In August, 2012 reports were made that the energy company Dalakraft was on the rise. HCL Technologies were during the very same period named the most innovative IT supplier in Scandinavia. Another key term in the results is *Taikon*, which is a traditional Romani surname. In August, 2012, a biography of the author Katarina Taikon was released and subsequently reviewed. Further, the key term *overturn* most likely refer to a news story in August, 2012 about an 82 year old man overturned his car and drowned in Årydssjön, outside of the Swedish town Kalmar.

Table **??** displays the results of the χ^2 topic modeling of week 25 in August,

	term (notes and translation)
2012.	oppohjola (company)
	teckningsoptioner (stock option rights)
	aktieanalys (stock analysis)
	rezidor (hotel)
	biogaia (company)
	pohjola (company)
	affär (business)
	lån (mortgage)
	reuteri (lacotbacillus reuteri, bacterium)
	lactobacillus (lactobacillus reuteri, bacterium)
	tigran (company)
	uponor (company)
	bohjälte (living hero)
	adtail (company)
	regent

Table 5.16: Displays the χ^2 topic modeling results of editorial media week 25, 2012.

One entry has been excluded from the table above. That entry included the name of a semi-public person.

A strong topic during the period at hand seems to be the stock market. There is also a key term *business* present in the week at hand.

OP-Pohjola is a commercial banking company whose profit in August 2012 was reported to be positive despite the recession. Reijo Karhinen is the Chief Executive Officer of the OP-Pohjola Group. Other company-related terms are *biogaia*, *rezidor*, *uponor*, *living hero* and *adtail*. *Living hero* refers to a public relations campaign conducted by Svensk Fastighetsförmedling. Adtail was reported bankrupt during the early fall of 2012.

Another important topic during this week seem to be that of *lactobacillus*. This is an extended topic as it is associated with terms such as *biogaia*. Lactobacillus, or lactobacillus reuteri protectis is a probiotic. Biogaia is a health care company that distributes this probiotic.

A related topic that relates closely to the field of medicine is *tigran*. Tigran PeriBrush is titanium brush for dental implants. A study comparing this particular brush with a similar brush manufactured by another company was published in August, 2012.

Table 5.17 displays the results of the χ^2 topic modeling of August 22nd, 2012.

term (notes and translation)
morphosys (company)
universeum (museum)
zubsolv (drug)
förövare (attacker)
suboxone (drug)
ruthberg (CEO of eWorks)
ework (company)
adidas (company)
uponors (company)
exait (company)
konsortiumadministration (consortium administration)
opiatberoende (opiate addiction)
uponor (company)
orexo (company)
kwh consortium
kirsebergs fängelse (prison of kirseberg)
millicom (company)

Table 5.17: Displays the χ^2 topic modeling results of news media day August 22nd, 2012.

The most dominant news media topics on August 22nd, 2012 are related to the broader categories drugs and business. In the drug category falls key terms such as *zubsolv*, *suboxone* and *orexo*. Orexo is an Uppsala-based pharmaceutical company that manufactures, among other things, the drug Zubsolve, which is a dissolvable tablet for the treatment of opioid dependency. Suboxone is also a drug that treats opioid addiction. This information may be extracted from the key terms in the table 5.17 as they hold terms and associated terms such as *opiate addiction zubsolv* and *opiate addiction*.

The conceptual business category is comprised of terms such as *morphosys*, *ruthberg*, *ework*, *adidas*, *exane*, *exait*, *consortium administration*, *uponor*, *kwh consortium* and *millicom*. These all refer to companies or their management.

The remaining topics are restricted to *universeum*, which refers to a museum in Gothenburg, *attacker*, and *kirseberg's prison*.

In August, 2012, an announcement from the museum Universeum arrived to editorial media announcing that Universeum had been forced to let go of staff. During the same month, another news piece about the prison of Kirseberg was also topical – the news that inmates were no longer allowed to play cards. These seminal events are most likely the roots of the occurrences of these experimental topics. Key terms such as *attacker*, however, are too general to be traced to a single seminal event. Nonetheless, it is apparent that it concerns a current crime.

term (notes)	associated terms (notes)
grammer (Kelsey Grammer, actor)	
företagsrekonstruktion (debt reconstructing)	
imperva (company)	
arbeit (german for *work*)	
kusträddarvm (coast rescue world cup)	
falkarna (the falcons)	
självmordsbombning (suicide bombing)	
aspera (company)	
ablynx (company)	
sipri (peace research institute)	
breivik (Anders Behring Breivik, mass murderer)	
nationalteater (national theater)	
kelsey (Kelsey Grammer, actor)	
karimova (Gulnora Islomovna Karimova, social entrepreneur)	
spårvagnstrafik (tramway traffic)	
brück (german economist)	
paceco (company)	
frei (german for *free*)	

Table 5.18: Displays the χ^2 topic modeling results of news media day August 21st, 2012.

The results of the topic modeling on August 21st display less business related terms. Nonetheless, there are business-related terms present. For example, *imperva*, *aspera*, *ablynx* and *paceco*.

More news events are present in the current results. For example, there are key terms present, such as *suicide bombing, the falcons, tramway traffick*, *breivik* and *debt reconstructing*, all of which have support in the editorial media during this time period.

There are for example entertainment-related news such as *grammer* and *kelsey* which both refer to the actor Kelsey Grammer who walked out in the middle of a talk show interview in August, 2012. Other entertainment oriented news includes the world life saving championships, i.e. *the coast rescue world cup*.

In summary, the temporal χ^2 topic modeling of news media are good and sound. However, the observant reader might have noticed that there are very few topics of political significance. The editorial media topics are highly dominated by corporate and medical topics. Very few of the topics at hand are related to politics.

5.3.4 Results overview

The aim of the current study is to perform topic modeling for three data channels: (1) parliament, (2) social media and (3) news media. Two methods were tested: LDA and temporal χ^2. With LDA, topic modeling was performed only for parliament data. This was due to the fact that the results were very discour-

aging. Temporal χ^2, on the other hand, performed topic modeling successfully for all three data channels.

The results from the current experiments clearly show that LDA is *not* an appropriate method for the purposes of this project. The poor results of the LDA will be accounted for and discussed in section 6.1. Temporal χ^2, however, is in fact a satisfactory method for the data channels at hand. This is discussed in section 6.2.

6 Discussion

What conclusions can we draw from the current project? What is to be learned?

6.1 Discussing the LDA results

As is clearly stated previously, the topics created by LDA for the parliament data are semantically incoherent at best, and nonsense at worst. The output of the LDA is unusable for the aims of the project at hand. The idea of using LDA in the development of a dynamic atlas of Swedish political discourse is abandoned.

Please note, however, that experiments with LDA were not performed for the social media and editorial media corpora, but only the parliament corpus. Therefore, we have no assessment of the performance of LDA on the social and editorial data channels. We do, however, have strong reasons to believe that the results would have been dramatically better for these data sources as the literature and previous research indicates this. Nonetheless, topics obtained from parliament data are an essential and non-disposable part of the project at hand which makes the use of LDA out of the question.

But what happened? Why are the results of the LDA on parliament data so discouraging?

The most obvious explanation of the poor performance of the LDA on parliament data is the topic distribution of the given data set. The term probabilities are maximized by dividing the terms into topics, i.e. clusters of co-occurring words. The Dirichlet of the topic proportions, however, encourage sparsity. This means that a document is "penalized" if it uses too many topics. This is not, however, a property from which the parliament texts suffer. Most often, they are short and well defined. They rarely concern more than one or two topics, within the scope of topics that LDA can manage. Therefore, we need to explore other explanations for the poor performance of the LDA topic modeling on parliament data.

Editorial media texts, on which LDA traditionally works very well, cover widely different topics – from environmental politics to crime and celebrity gossip. The linguistic configuration of these topics are also highly varied. That is, the stylistics of texts from these different topics are distinct. This might help boost the performance of the LDA topic modeling as it becomes easier to separate terms and the co-occurrence of terms in discrete clusters. The language of the parliament, however, is homogeneous and non-distinct over topics. The poor performance of LDA topic modeling on this particular data source might be due to this.

6.2 Discussing the temporal χ^2 results

The primary contribution of this thesis is the conclusion that temporal χ^2 statistics is - most likely - an excellent alternative to computationally expensive topic modeling algorithms such as LDA. I say *most likely* because we lack evidence as no formal evaluation of the algorithm and the results are currently in place. The results of the temporal χ^2 topic modeling of the parliament corpus is, in my opinion, more than satisfactory as they do in fact reflect the parliamentary activity.

The attentive reader might ask why, considering that the quality of the results of the temporal χ^2 is varying and often does not even contain terms of political significance. The answer is simple. The fact that the temporal χ^2 results do not output topics of political significance is not an algorithmic or methodological problem, but a problem of filtering – or lack thereof.

Before elaborating on this conclusion, let us revisit some of the disappointing results of the temporal χ^2 topic modeling.

For example, the results of the temporal χ^2 topic modeling of social media for the date August 29th, 2012 rendered very poor results. It cannot be concluded that these results might be due to something so trivial as a bad news day. A bad news day, in this case, refers to a day in which no big news items that attracted the coherent attention of the blogosphere were published.

It should also be noted that the results of the editorial media topic modeling are significantly richer than the topic modeling of the social media. This is most likely due to the fact that editorial media is more mainstream and, therefore, more coherent. It lies in the nature of editorial media. When one source of news reports a seminal event, other news sources are likely to pick up on it and publish more articles on that specific event. Editorial media is probably more coherent than the social media stream and the parliament text stream.

In addition to this, most of the editorial media topics are pharmaceutical or business-related. This is probably due to an overrepresentation of industry papers. This can be easily corrected by only including certain news sources into the editorial media stream.

It is, however, obvious that the study at hand is lacking in some aspects. This has been partially addressed previously. For example, it is easily noted that the social media corpus should have been filtered to exclude entries that do not specifically concern politics. In the current study, all available tweets and blog entries have been used. This was clearly a mistake since most tweets and blog entries are not, in fact, concerned about politics. The assumption behind this choice was that terms that are important enough during a given period of time to form topics, can in reality only be sparked by politics and similar seminal events. However, this assumption is clearly false. A subset of the data at hand - that actually concerns politics - should be used to re-run the temporal χ^2 experiments. This can easily be performed by sampling political entries from a larger set of social media data.

6.3 Tuning the topic model

Both the results of the LDA and the temporal χ^2 topic modeling clearly display *the absence of lemmatization and stemming*. The lack of this preprocessing step will inevitably lead to a loss in recall as the system will not read inflected versions of a single term as one word. However, in morphology lies information – information which a system including stemming or lemmatization may lose.

6.4 Evaluation

A large portion of this project has been spent thinking about evaluation. No suggestion of how a modern and pragmatic evaluation should be constructed can be made at this point in time. However, some guiding principles of such an evaluation have been laid out. These include, but are not limited to:

1. Reproducibility.
2. Transparency.
3. Portability.
4. Different types of user rationales require different types of evaluation.
5. Some factors of the planned system are linguistic; others are not.
6. Interaction is diverse and requires adapted metrics.

The small evaluation present in the present thesis obviously does not fulfill all of these criteria.

7 Conclusions

The aim of this thesis was to answer the following questions:

1. Is it possible to successfully perform topic detection (also called topic extraction) with LDA on the current set of parliament texts, social media texts and news media texts?
2. Is it possible to successfully perform topic detection (also called topic extraction) with χ^2 statistics on the current set of parliament texts, social media texts and news media texts?

These questions can be answered by stating that:

1. **No**, it is not possible to successfully perform topic detection with latent Dirichlet allocation on the current set of parliament texts, social media texts and news media texts.
2. **Probably.** It is most likely possible to successfully perform topic with χ^2 statistics (temporal χ^2, more specifically) on the current set of parliament texts, social media texts and news media texts. However, as this study lacks formal evaluation it cannot be established beyond doubt.

8 Lessons learned

1. **What is necessary for topic modeling of political discourse?** Most crucial is to *keep your data clean*: use a highly balanced corpus and filter it accurately.

2. **How does one comprehend political discourse?** Establish a register analysis of political discourse: get to know your data and process it accordingly.

3. **How does one perform practical language modeling?** A key term here is *practical*, so keep it simple. Do not use algorithms with near exponential time complexity and make sure the modeling is doable without terabytes of data.

4. **How does one evaluate information retrieval systems?** By constructing evaluations that are reproducible, transparent and portable, but that also takes other factors, such as rationale, media type and user interaction, into consideration.

Bibliography

Beaugrande, Robert-Alain de (1996). *Introduction to Text Linguistics*. Longman Linguistics Library. Longman. URL: http://books.google.se/books?id=276lKQEACAAJ.

Blei, David M. (2011). "Introduction to Probabilistic Topic Models". *Communications of the ACM*. URL: http://www.cs.princeton.edu/~blei/papers/Blei2011.pdf.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent dirichlet allocation". *J. Mach. Learn. Res.* 3, pp. 993–1022. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=944919.944937.

Brants, Thorsten, Francine Chen, and Ayman Farahat (2003). "A System for new event detection". In: *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '03. Toronto, Canada: ACM, pp. 330–337. ISBN: 1-58113-646-3. DOI: 10.1145/860435.860495. URL: http://dx.doi.org/10.1145/860435.860495.

Chen, Kuan-Yu, Luesak Luesukprasert, and Seng-cho Timothy Chou (2007). "Hot Topic Extraction Based on Timeline Analysis and Multidimensional Sentence Modeling". *IEEE Trans. Knowl. Data Eng.* 19.8, pp. 1016–1025. DOI: http://doi.ieeecomputersociety.org/10.1109/TKDE.2007.1040.

Cieri, Christopher, David Graff, Mark Liberman, Nii Martey, and Stephanie Strassel (2000). "Large, Multilingual, Broadcast News Corpora For Cooperative Research in Topic Detection And Tracking: The TDT-2 and TDT-3 Corpus Efforts". In: *In Proceedings of Language Resources and Evaluation Conference*.

Croft, Bruce W. and Jinxi Xu (1999). "Cluster-based language models for distributed retrieval". In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '99. Berkeley, California, USA: ACM, pp. 254–261. ISBN: 1-58113-096-1. DOI: 10.1145/312624.312687. URL: http://doi.acm.org/10.1145/312624.312687.

Dahllöf, Mats (2012). "Automatic prediction of gender, political affiliation, and age in Swedish politicians from the wording of their speeches - A comparative study of classifiability." *LLC* 27.2, pp. 139–153. URL: http://dblp.uni-trier.de/db/journals/lalc/lalc27.html#Dahllof12.

Ejerhed, Eva and Gunnel Källgren (1997). *Stockholm Umeå Corpus. Version 1.0*.

Fiscus, Jonathan G. and George R. Doddington (2002). "Topic Detection and Tracking". In: ed. by James Allan. Norwell, MA, USA: Kluwer Academic Publishers. Chap. Topic Detection and Tracking Evaluation Overview,

pp. 17–31. ISBN: 0-7923-7664-1. URL: http://dl.acm.org/citation.cfm?id=772260.772263.

Fung, Gabriel Pui Cheong, Jeffrey Xu Yu, Huan Liu, and Philip S. Yu (2007). "Time-dependent event hierarchy construction". In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '07. San Jose, California, USA: ACM, pp. 300–309. ISBN: 978-1-59593-609-7. DOI: 10.1145/1281192.1281227. URL: http://doi.acm.org/10.1145/1281192.1281227.

Griffiths, Thomas L., Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum (2005). "Integrating topics and syntax". In: *In Advances in Neural Information Processing Systems 17*. MIT Press, pp. 537–544.

Hoffman, Matthew D., David M. Blei, and Francis Bach (2010). "Online learning for latent dirichlet allocation". In: *In NIPS*.

Hulpus, Ioana, Conor Hayes, Marcel Karnstedt, and Derek Greene (2013). "Unsupervised Graph-Based Topic Labelling using DBpedia". In: *6th ACM International Conference on Web Search and Data Mining (WSDM'13)*.

Jurafsky, Daniel and James H. Martin (2008). *Speech and Language Processing (2nd Edition) (Prentice Hall Series in Artificial Intelligence)*. 2nd ed. Prentice Hall. ISBN: 0131873210.

Jurgens, David and Keith Stevens (2009). "Event detection in blogs using temporal random indexing". In: *Proceedings of the Workshop on Events in Emerging Text Types*. eETTs '09. Borovets, Bulgaria: Association for Computational Linguistics, pp. 9–16. ISBN: 978-954-452-011-3. URL: http://dl.acm.org/citation.cfm?id=1859650.1859652.

Kumaran, Giridhar and James Allen (2004). "Text classification and named entities for new event detection". In: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '04. Sheffield, United Kingdom: ACM, pp. 297–304. ISBN: 1-58113-881-4. DOI: 10.1145/1008992.1009044. URL: http://doi.acm.org/10.1145/1008992.1009044.

Lam, W., H. M. L. Meng, K. L. Wong, and J. C. H. Yen (2001). "Using contextual analysis for news event detection". *Int. J. Intell. Syst.* 16.4, pp. 525–546.

Li, Wentian (1992). "Random texts exhibit Zipf's-law-like word frequency distribution". *IEEE Transactions on Information Theory*, pp. 1842–1845.

Nallapati, Ramesh, Ao Feng, Fuchun Peng, and James Allan (2004). "Event threading within news topics". In: *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. CIKM '04. Washington, D.C., USA: ACM, pp. 446–453. ISBN: 1-58113-874-1. DOI: 10.1145/1031171.1031258. URL: http://doi.acm.org/10.1145/1031171.1031258.

Porteous, Ian, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling (2008). "Fast Collapsed Gibbs Sampling for Latent Dirichlet Allocation". In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '08. Las Vegas, Nevada, USA: ACM, pp. 569–577. ISBN: 978-1-60558-193-4. DOI: 10.1145/1401890.1401960. URL: http://doi.acm.org/10.1145/1401890.1401960.

Sahlgren, Magnus and Jussi Karlgren (2008). "Buzz Monitoring in Word Space". In: *Proceedings of the 1st European Conference on Intelligence and Security Informatics*. EuroISI '08. Esbjerg, Denmark: Springer-Verlag, pp. 73–84. ISBN: 978-3-540-89899-3. DOI: 10.1007/978-3-540-89900-6_10. URL: http://dx.doi.org/10.1007/978-3-540-89900-6_10.

Stoyanov, Veseline and Claire Cardie (2010). "Topic Identification for Fine-Grained Opinion Analysis." In: *COLING*. Ed. by Donia Scott and Hans Uszkoreit, pp. 817–824. ISBN: 978-1-905593-44-6. URL: http://dblp.uni-trier.de/db/conf/coling/coling2008.html#StoyanovC08.

Wallach, Hanna M. (2005). "Topic modeling: beyond bag-of-words". In: *NIPS 2005 Workshop on Bayesian Methods for Natural Language Processing*.

Wang, Canhui, Min Zhang, Liyun Ru, and Shaoping Ma (2008). "Automatic online news topic ranking using media focus and user attention based on aging theory". In: *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*. Napa Valley, California, USA: ACM, pp. 1033–1042. ISBN: 978-1-59593-991-3. DOI: 10.1145/1458082.1458219. URL: http://dx.doi.org/10.1145/1458082.1458219.

Wang, Chong, John William Paisley, and David M. Blei (2011). "Online Variational Inference for the Hierarchical Dirichlet Process". *Journal of Machine Learning Research - Proceedings Track* 15, pp. 752–760.