Exploring the Use of Topic Analysis in Latvian Legal Documents

Rinalds Vīksna[1], Marite Kirikova[1][0000-0002-1678-9523], and Daiga Kiopa[2]

[1]Department of Artificial Intelligence and Systems Engineering,
Riga Technical University, Latvia
[2] Lursoft, Latvia
rinaldsviksna@gmail.com; Marite.Kirikova@rtu.lv;
daiga@lursoft.lv

Abstract. The large number of legislative documents produced every day makes it difficult to follow each and every document. However, it is important for enterprises to comply with all current legislative acts. In this paper we demonstrate the application of different topic analysis algorithms and stop word filtering approaches to the corpus of legal texts of the Republic of Latvia. This is done for the purpose of supporting the discovery of expressive and meaningful legal topics and marking respective documents according to those topics. Topic models produced in this work are intended to be used as an aid for experts, enabling faster document browsing possibilities.

Keywords: Topic Analysis - Legal Analysis - Information Retrieval

1 Introduction

Every enterprise must conform to and comply with current regulatory acts. Moreover, some types of regulations may be used as blueprints for business process models [1]. Legislative documents may be laws issued by parliament, regulations issued by the Cabinet of Ministers, Municipalities or other institutions, as well as industry standards, various contracts and other documents [2]. Many regulations are related to others, being either an update of an earlier regulation or depending on or being implemented by other regulations. Keeping track of the changing regulatory environment requires significant time and effort.

In this paper we envision a solution that may help to save effort by the overview and summarization of various topics within the Latvian law domain. The goal of this paper is to explore the application of different topic analysis algorithms and stop word filtering approaches on the corpus of legal texts of the Republic of Latvia. For the demonstration we use three common topic analysis algorithms, briefly introduced in

Section 2. The paper presents the research in progress that is a part of more extensive research activity, the aim of which is to find core topics in Latvian legislation, as well as to identify, for further exploration, a method for automated document tagging with salient topics.

The paper is organized as follows. Section 2 discusses the problem domain and available topic analysis algorithms. Section 3 shows data preparation steps and the results of stop word removal. Section 4 discusses the results obtained and Section 5 provides brief conclusions.

2 Related Work and Background

Topic analysis (often called "topic modeling" or "topic detection") is a text-mining [3] technique for soft clustering (where each document has a probability distribution over all the clusters) of documents according to distribution of terms that occur in the text body. O'Neill et al. [4] used topic analysis to summarize and visualize British legislation to find useful topics and terms. Wyner et al. applied topic analysis to profile and extract arguments from legal cases [5]. Soria et al. applied topic analysis to annotate each paragraph in Italian law texts with semantic information [6]. Sulea et al. explore the use of text classification methods in [7], however, here, the use of text classification methods requires that documents have known labels. The results may differ depending on the language used. In this work we address regulatory (legal) documents in Latvian with the purpose of supporting legal document handling activities by experts.

Topic analysis is an unsupervised learning method, which produces a number of topics, which each consist of related terms and their respective weights. Topic analysis is most often done using either Latent Semantic Analysis (LSI), Latent Dirichlet allocation (LDA) [8] or its variant - Hierarchical Dirichlet Process (HDP) algorithms [9], [3]. These algorithms are briefly described below and, being the most popular ones, were used in the experiments reported in Section 4.

2.1 Latent Semantic Indexing

Latent Semantic Analysis, also called LSI, is a method for extracting and representing the contextual usage meaning of words by statistical computations applied to a large corpus of text. The text corpus is viewed as a set of term tf-idf weights, where tf is a term frequency in the given text, and idf is an inverse document frequency. To this term-document matrix singular value decomposition (SVD) is applied. In SVD a rectangular matrix is decomposed into the product of three other matrices in order to find a lower rank approximation of the term-document matrix [10]. LSI is implemented using gensim[1] python library.

[1] https://radimrehurek.com/gensim/

2.2 Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics where each topic is characterized by a distribution over words. The LDA algorithm is described in [11] and it is implemented in the sklearn python library[2] and in gensim. Term distinctiveness and saliency are used to evaluate generated topics. For a given word *w*, unconditional probability P(*w*) and probability P(*T*|*w*) that that given word was generated by latent topic *T* is computed. Probability P(*T*|*w*') that random word *w*' was generated by topic *T* is also computed. Distinctiveness of word *w* is then calculated as follows [12]:

$$\text{distinctivness (w)} = \sum_T P(T|w) log \frac{P(T|w)}{P(T)} \quad (1)$$

The above-presented equation describes how informative word *w* is for determining topic *T*. If a word occurs in all topics, observing the word tells little about the document's topic, and the word has little distinctiveness. The saliency of a word is defined as [12]:

$$saliency(w) = P(w) * distinctivness(w) \quad (2)$$

2.3 Hierarchical Dirichlet Process

Hierarchical Dirichlet Process (HDP) is a Bayesian nonparametric model for unsupervised analysis of grouped data. Documents are viewed as bags of words, which are drawn from a number of latent clusters or "topics", where "a topic" is modeled as a multinomial probability distribution on words from some basic vocabulary. Given a collection of documents, HDP finds latent clusters, without the need to specify the number of topics as a parameter [9]. HDP analysis requires multiple passes through all the data and therefore is poorly suited for massive and streaming data. Wang C. proposed Online Variational Inference for the HDP algorithm, which requires only one pass through data and is significantly faster [9] and is implemented in gensim library.

3 Corpus and Data Preparation and Analysis

In this paper, we use the corpus of legal acts from http://likumi.lv/ - a website of legal acts that ensures free access to systematized (consolidated) legal acts of the Republic of Latvia. Documents were downloaded as HTML documents which were kept for later metadata extraction (document information – issuer, status, adoption, end-of-validity date, related documents, etc.). Text contents of downloaded HTML documents were extracted and saved as plain text in UTF-8 format. Some of the documents contained mostly Russian or English text and were dropped from the corpus. In

[2] https://scikit-learn.org/stable/

total, over 50000 documents in the Latvian language were collected. With these documents, the experiments regarding stop word removal and different topic models were made.

In the remainder of this section, first we do exploratory data analysis, and assess the impact of stop word removal on the performance of clustering algorithms with LDA as an example (sub-Section 3.1), and in the second part (sub-Section 3.2) explore alternative clustering algorithms.

3.1 Experiments with Stop Word Removal Approaches

During text preprocessing boilerplate content (irrelevant text, ads) and generic Latvian stop words identified by Garkaje [13] were removed. After this step, the 10 most common words in the corpus were identified (Fig.1).

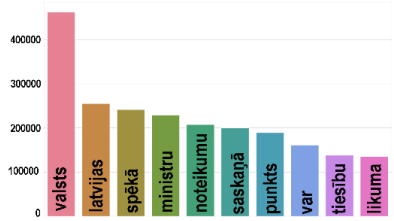

Fig. 1. Most common words in corpus.

As we see in **Fig. 1**, most common words ("state", "latvian", "in force", etc.) occur multiple times in most documents in the corpus and are not very informative, so in this context those words are stop words and were removed. To find stop words specific for this domain, a normalized tf-idf metric was used [14]. A tf-idf metric was calculated for each word as given in [14]:

$$\begin{aligned} tf\text{-}idf &= \mathrm{tf}(k_{\mathrm{norm}}) * idf(k) \\ tf(k_{norm}) &= -\log(\tfrac{TF}{U}) \\ idf(k) &= \log(\tfrac{N(doc)}{N(k)}) \end{aligned} \tag{3}$$

where TF – term frequency, is the number of times a certain word appears in this corpus; N(doc) – number of documents in the corpus; N(k) – number of documents containing term k, and U – total number of words in the corpus. A tf-idf was calculated for each word in the corpus, and 140 words with a tf-idf score of less than 9 were selected as stop words. This custom stop word list was combined with general Latvian stop words from Garkaje [13]: in total 450 stop words. After additional stop word filtering, documents contained relatively more informative words (see **Fig. 2**).

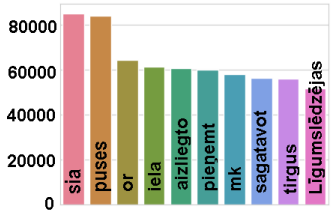

Fig. 2. Most common words in the corpus after stop words removal.

To evaluate topics created using different stop word selection, we used preassigned theme labels given in http://likumi.lv/ .[3] Documents belonging to the same theme should have similar content or describe similar topics and questions. It should be noted that each document in likumi.lv may belong to more than one theme. We used documents from 3 themes: "human rights", "banks, finance, budget" and "taxes and fees". Topic models were created (one – using generic stop word set, and another – using adapted stop word set) by which selected documents were then classified as belonging to particular topics. Document distribution by topics is shown in Fig. 3 and Fig. 4.

To assess the impact of domain-specific stop words removal, we implemented the LDA model using a general list of stop words (Fig. 3) and then compared it with the model generated using a domain-specific list of stop words (Fig. 4). One topic is found by both models: Topic 9 in Fig. 3 and topic 15 in Fig. 4 represent a document with significant English content – it is indicated by keywords in the English language. Other topics, although different, display some similarities (Table 1 and Table 2).

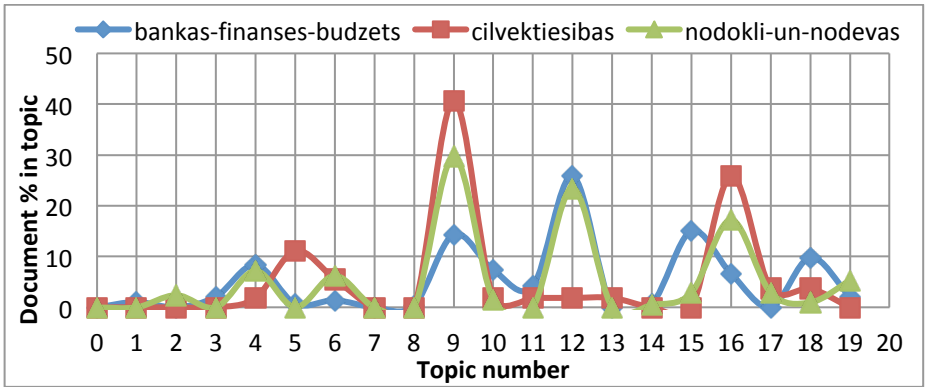

Fig. 3. Document distribution by topic using generic stop word list.

Table 1. Keywords of most prominent topics without filtering.

Topic	Keywords in Latvian	Keywords translated in English
# 4	Latvijas, republikas, padomes, eiropas,	Latvian, republic, councils, european

[3] https://likumi.lv/ta/tema

	gada	
# 5	satversmes, tiesas, likuma, gada, panta	constitutions, courts, in law, year, article
# 9	or, shall, be, for, by, article	or, shall, be, for, by, article
# 12	finanšu, panta, likuma, gada, labuma	financial, article, in law, year, benefit
# 15	puses, pants, līgumslēdzējas, saskaņā, spēkā	sides, article, contracting, under, in force
# 16	punkts, valsts, nr, personas, noteikumu	point, country, no, persons, regulations
# 18	valsts, izglītības, gadā, darba, gada	country, education, in year, work, year

As we see in Table 1, many keywords are present in multiple topics and are not representative (“year”, “in law”). Topics generated using the second model (Table 2) contain representative words (i.e. “Cadaster”, “convention”), which indicate that this topic talks about real estate (“Cadaster”), or international treaties (“convention”).

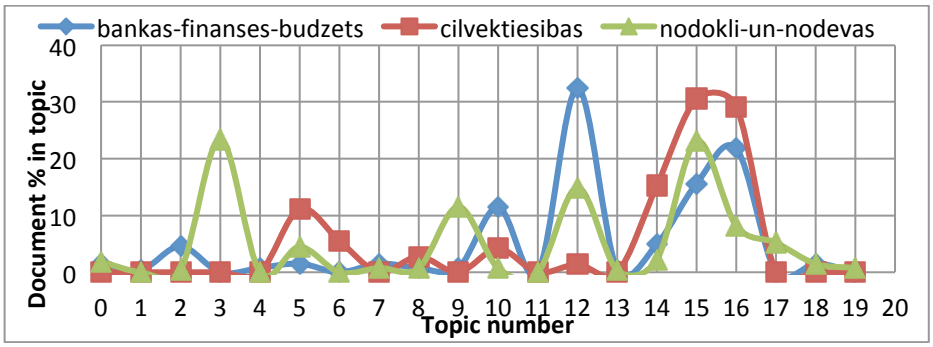

Fig. 4. Document distribution by topic using adapted stop word list.

Table 2. Keywords of most prominent topics with filtering.

Topic	Keywords in Latvian	Keywords in English
# 3	valstī, līgumslēdzējas, nodokļiem, state, līgumslēdzējā	in the country, contracting, taxes, state, contracting
# 5	likumu, persona, daļā, tiesas, redakcijā	law, a person, part, courts, version
# 9	kadastra, nekustamā, eiro, nodokļa, īpašumu	cadaster, real, euro, tax, property
# 10	atbalsta, izmaksas, programmas, ietvaros, sadarbības	supports, costs, programs, within, cooperation
# 12	kapitāla, ieguldījumu, tirgus, pārskata, apdrošināšanas	capital, investment, market, review, insurance
# 14	vēlēšanu, domes, pilsētas, komisija, pārvaldes dienesta	election, city council, cities, commission, administration, service
# 15	or, shall, be, for, by, article	or, shall, be, for, by, article
# 16	puses, līgumslēdzējas, puse, konvencijas, teritorijā	sides, contracting, sides, convention, territory

The LDA model using adapted stop word list created more meaningful topics, as it contained more meaningful words, which tell us more about its content. Furthermore, as most of the more popular words were labeled as stop words, the model was able to classify more tax related documents into very expressive topics 3, 5, 9, 12 in Fig. 4, while the model with just generic stop word filtering applied, created more broad topics as # 4, 6 and 12 in Fig. 3. Both models classified documents into multiple topics, some of which corresponded to topics assigned to those documents in http://likumi.lv/. However, most topics were different; for instance, topics containing English words, or terms related to international treaties. This shows that topics that are found using topic analysis give insights into data and offer different classification schemes.

3.2 Topic Models and Their Evaluation

To evaluate the performance of different topic analysis algorithms, we built topic models using LDA, HDP and LSI algorithms, and visualized topics found using gensim visualization tool. HDP algorithm does not need to know the number of topics as it is able to determine the number of topics automatically. In this case it has found 150 topics, which is the maximum allowed by gensim implementation. The first topic (see Fig. 5 left) contains 80% of tokens (words), the second topic – 16% of tokens, the third – 2.5% of tokens, and the rest of the topics contain less than 1.5% of tokens.

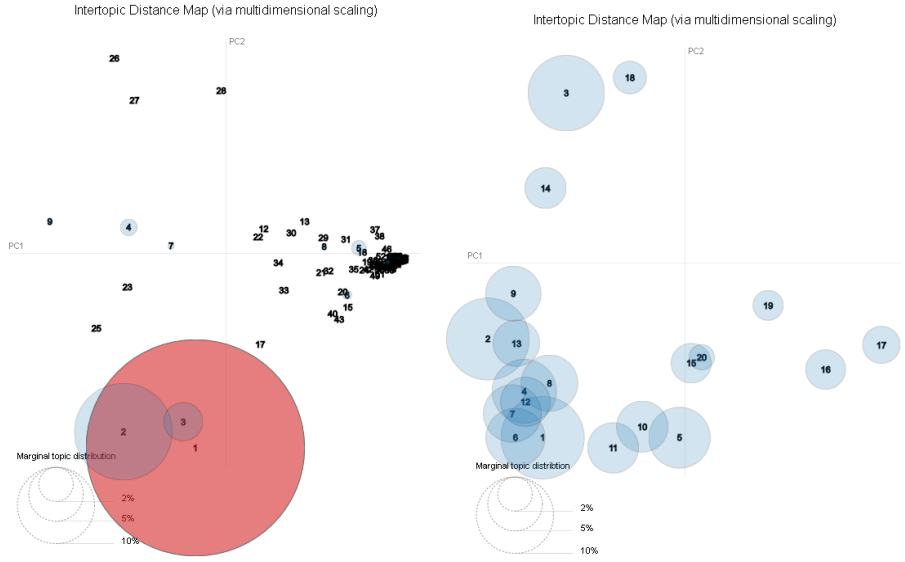

Fig. 5. Topics found using HDP (left) and LDA (right).

The LDA model, in comparison with the HDP model, is more balanced – the largest topic contains 11.7% of tokens and the smallest one contains 1.7% of tokens. It was not possible to visualize the LSI model, as it contains negative weights for terms, which are not supported by gensim. Therefore, models were evaluated using coherence metrics proposed by Röder et al. [15]. The results are shown in Table 3.

Table 3. Coherence measures of topic models.

Topic model	Coherence measure (u_mass)
HDP	-7.906688044302112
LDA (20 topics)	-7.7222343265180715
LSI	-9.482837750182188

In Table 3, the LDA 20 topic model has the highest coherence measure among the three models.

4 Conclusions

We explored the application of different topic analysis algorithms and stop word filtering approaches to the corpus of legal texts in the Latvian language. Domain-adapted stop word filtering improves topic models that are produced using the LDA model, yielding more expressive topics, which allow separating of topics into more distinctive groups. When the corpus contains multiple documents, stop word filtering methods, explored in this work, are applicable to other corpora and languages. Compared to the LSI and the HDP, the LDA algorithm produced a topic model which was shown to perform better than the alternatives. However, for topics generated by the

LDA to be of practical value, some further finetuning needs to be done, as, currently, topics from different dimensions are mixed – for instance, there was a topic for documents in English, and a topic for documents which are international treaties, and both topics encompass documents which talk about different themes such as civil rights and taxes.

Acknowledgments. The work on this paper is supported by ERAF research 1.2.1.1/18/A/003 project No. 1.9.

References

1. Kirikova, M., Buksa, I., Penicina, L.: Raw materials for business processes in cloud. Lect. Notes Bus. Inf. Process. 113 LNBIP, 241–254 (2012).
2. T.L.L.S. of Washington, Types of Legislative Documents. (2018).
3. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., Kochut, K.: A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. (2017).
4. O'Neill, J., Robin, C., O'Brien, L., Buitelaar, P.: An analysis of topic modelling for legislative texts. CEUR Workshop Proc. 2143, (2017).
5. Wyner, A., Mochales-Palau, R., Moens, M.F., Milward, D.: Approaches to text mining arguments from legal cases. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 6036 LNAI, 60–79 (2010).
6. Soria, C., Bartolini, R., Lenci, A., Montemagni, S., Pirrelli, V.: Automatic extraction of semantics in law documents. Proc. V Legis. XML Work. (February 2007). 253–266 (2007).
7. Sulea, O.M., Zampieri, M., Malmasi, S., Vela, M., Dinu, L.P., Van Genabith, J.: Exploring the use of text classification in the legal domain. CEUR Workshop Proc. 2143, (2017).
8. Mehdiyev, R., Nava, J., Sodhi, K., Acharya, S., Rana, A.I.: Topic subject creation using unsupervised learning for topic modeling. (2019).
9. Wang, C., Paisley, J., Blei, D.M.: Online variational inference for the hierarchical Dirichlet process. J. Mach. Learn. Res. 15, 752–760 (2011).
10. Olmos, R., León, J.A., Jorge-Botana, G., Escudero, I.: New algorithms assessing short summaries in expository texts using latent semantic analysis. Behav. Res. Methods. 41, 944–950 (2009).
11. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003).
12. Chuang, J., Manning, C.D., Heer, J.: Termite: Visualization techniques for assessing textual topic models. Proc. Work. Adv. Vis. Interfaces AVI. 74–77 (2012).
13. Garkaje, G., Zilgalve, E., Dargis, R.: Normalization and Automatized Sentiment Analysis of Contemporary Online Latvian Language. Front. Artif. Intell. Appl. 268, 83–86 (2014).
14. Tsz-Wai Lo, R., He, B., Ounis, I.: Automatically Building a Stopword List for an Information Retrieval System.
15. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. WSDM 2015 - Proc. 8th ACM Int. Conf. Web Search Data Min. 399–408 (2015).