TOPIC EXTRACTION AND EVOLUTION OF LEGAL DOCUMENTS USING LATENT DIRICHLET ALLOCATION

EISHA PATEL, CHAITRA HOSMANI
DATA SCIENCE AND ANALYTICS
RYERSON UNIVERSITY - *2019*

Abstract

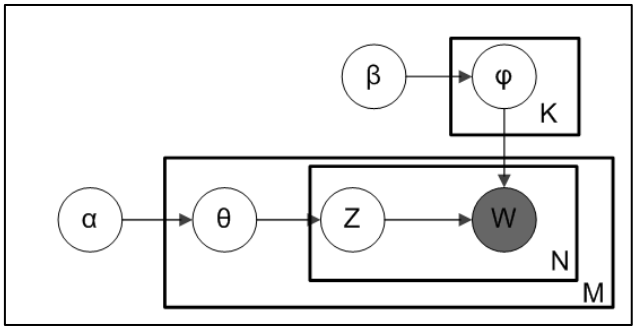
It's a challenge for the legal community to scavenge thru the abundance of legislative documents to access the information they need. There is a need for assistance in organizing and retrieving such documents. This paper attempts to address these concerns by using Latent Dirichlet Allocation (LDA) to extract topics within the text. Topics that are of similar nature can be clustered together and depicted using pyLDA visualizations. Documents can also be *tagged* based on their correlation to the various topics extracted.

Legal documents can change with time, leading to new policies and concerns. This paper also uses Dynamic Topic Modelling (DTM) to capture the changes in topics over time. Trends of evolution can be captured and displayed visually for interpretation.

1. Introduction

Topic Modelling is a set of unsupervised machine learning techniques that aim to discover a set of abstract themes (or topics) from a large collection of documents. The Latent Dirichlet Allocation (LDA) algorithm is one such technique. It is a continuous multivariate probabilistic model where each topic is a distribution over a fixed vocabulary [3]. Each topic contains each word from the dictionary, but in varying probabilities. Hence, the terms most relevant to a topic will occur with higher probabilities.

The fundamental concept of the LDA model is that each document can be described by a distribution of topics and each topic can be described by a distribution of words [3]. Figure A shows the plate diagrams of a typical LDA model [5].



α is the per-document topic distributions,
β is the per-topic word distribution,
θ is the topic distribution for document *m,*
φ is the word distribution for topic *k,*
z is the topic for the *n*-th word in document *m*, and
w is the specific word

Figure A: Plate Diagram of the LDA Model

Dynamic Topic Modeling (DTM) is an extension of the LDA in the sense that it attempts to capture the evolution of topics given a time-sequentially organized corpus of documents [4]. The goal is to identify newly emerging topics or even detect the evolution of topic vocabulary over time.

2. Related Work

[8] Rubayyi and Khalid discuss the strengths and limitations of various topic-modeling techniques including Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA), and Correlated Topic Model (CTM). For our research purposes, the LDA is simple to implement and provides sufficient results. This paper also discusses how time can be factored into topic modeling. TOT (Topic Over Time) discovers topics of time localization by considering word patterns. In DTM (Dynamic Topic Modeling), the documents are fed sequentially in time slices and a topic at time *"t"* evolves into topic at time *"t+1"*.

[9] This paper discusses how the coherence score can be used to find the optimal number of topics and increase the human interpretability of topics. It sheds light on various methods of evaluating topic coherence. The author talks about calculating the Pointwise Mutual Information (PMI) between each pair of words, where each word's proximity is calculated with every other word. We will be implementing this technique in our paper as well.

[4] This paper is particularly interesting because Ravi Kumar and Raghuveer are also using the LDA to classify a set of legal documents. They developed a topic based clustering model, capable of grouping the legal judgments into different clusters. The technique measures the cosine similarity between documents and topics and groups them based on the highest cosine value.

3. Dataset

The Justice Canada FTP server has a database of over 2600 Regulations, which are essentially a form of law with generic rules. Each regulation is made under the authority of an Enabling Act and covers a wide range of policies including food safety, import/export, and environmental protection. The server is updated regularly to add new documents and all files are publically available in XML format. A parsing function was created in Python to retrieve these files in a data frame format.

4. Methodology

4.1 Data Preparation

There are several columns available in the dataset, which are irrelevant for the analysis. We short-listed to keep only 3 columns; *content, xrefxternal* and *registration_year*. The *content* is the actual regulation itself. The *xrefxternal* tells us the Act under which that particular regulation is enabled. *Registration_year* is the year in which the regulation became valid. For this analysis, we dropped all rows with missing *registration_year* because this attribute is vital for Dynamic Modeling of Topics. This left 1753 rows of raw data to work with.

The *content* column needed to be preprocessed before the LDA algorithm could be applied to it. The *content* column can be converted to a list of documents. To begin, each document was tokenized and any punctuations and characters in the text were dropped. The tokenized words were further filtered to remove any stop words (see appendix for extended_stopwords.txt list). Legal documents contain a lot of legislative jargon, which do not contribute significant meaning to the text. Such terms were also filtered from the data to insure the topics generated are as interpretable as possible. Bigrams and Trigrams contribute to a meaningful phrase which otherwise would be broken down during tokenization. Phrases such as *"Royal Canadian Mounted Police"* and *"Northwest Territories"* were regrouped to preserve their meaning in analysis ahead. Furthermore, terms are reduced to their base form via lemmatization and the final dataset is passed thru a part-of-speech (POS) filter to preserve only nouns, adjectives, adverbs, and verbs.

4.2 Building the LDA Model

The main inputs of any LDA model are: 1) the dictionary of words and 2) the corpus (i.e. term-document frequency matrix). Once these 2 inputs are fed into the algorithm, a *"k"* number of topics will be generated. Deciding on the ideal number of topics before implementation is critical. Topics generated need to be interpretable and cover a reasonable range of policy types available in the dataset. The coherence score computes the sum of pairwise scores between the top frequently occurring "*n"*

words, used to describe a topic [1] (eq.2). For our analysis, we implemented the Extrinsic UCI Measure, which uses the Pointwise Mutual Information (PMI) function to evaluate the pairwise scores [1] (eq.1).

$$1.\ Coherence = \sum_{i<j} score(w_i, w_j)$$

$$2.\ score_{UCI}(w_i, w_j) = log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$$

The coherence score measures the interpretability of the topics generated and higher scores are generally favourable [1]. We can evaluate the coherence score for several values of *"x"* and determine the maximum. Figure B shows a maximum value occurring at 25 topics for our given dataset.

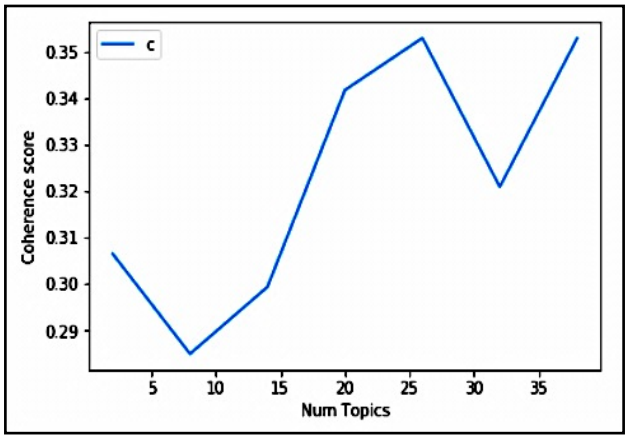


Figure B: Coherence Score versus Number of Topics

The LDA model generated a set of 25 fairly comprehendible topics. Figure C shows a subset of a few topics the LDA created. Having knowledge of various acts contained in the dataset allows us to see the resemblance between topics and acts. For example, topic 7 resembles the import/export permits act and topic 22 resembles the public service employment act.

Topic 7	Topic 8	Topic 14	Topic 24
free	product	service	amount
tariff	cannabis	public	pay
good	originate	group	fee
item	origin	period	payment
custom	manufacture	employ	payable
duty	sale	employee	interest
import	container	employment	respect
January	display	employer	rate
sor	exporter	cease	day
Canada	package	department	sor

Figure C: A sample of few topics generated via LDA Model. Words are arranged in order of highest probability of occurrence.

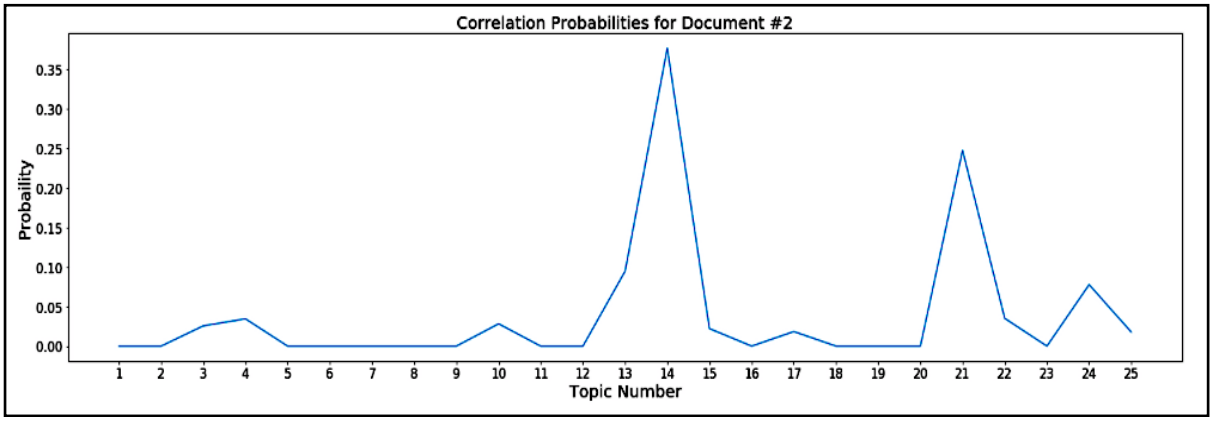


Figure D: Correlation probabilities of document 2 against each topic

4.3 Visualizing the LDA Topics

We can visually depict the 25 topics generated using pyLDA visualizations in Python [2]. Due to the size and interactive nature of the visual, we have attached an html link to the plot in the appendix. Each bubble in the visualization represents one of the 25 topics and bubble size represents the relative popularity of a topic in the dataset. We can hover over a topic to identify the top 30 relevant terms for that particular topic. The relevancy metric can be adjusted as well, where $\lambda = 0$ implies absolute specificity and $\lambda = 1$ implies looser specificity. The ideal value of λ can be established when we are able to see a semantic relationship between the top 30 terms. Distance perceived between bubbles shows the approximate semantic similarity between topics and hence can be used a means to cluster topics into groups. It's important to note that some terms occur in many topics. These terms are popular generic terms that do not contribute to the overall theme of a topic. We will see how these terms affect our model in the next section.

4.4 Document Tagging

Each document can be described by a list of topics and correlation values that relates the topics to the document [4]. The list of topics and correlation values is unique among documents and can also be considered the *signature* of that document [3]. The sum of all correlation probabilities between a document and all topics is equal to 1. Since one document can be highly correlated to multiple topics, we can create soft clusters of documents. Figure D shows the correlation probabilities of document 2 in the dataset against each topic. Document 2 falls under the public service and labour relations act. We can observe 2 distinct peaks at topic 14 with ~ 40% correlation and topic 21 with ~25% correlation. Topic 14 describes terms for labour laws and employee rights while topic 21 describes terms for automobiles and operation policies. Topic 21 does not resemble the content in document 2 very well. The ~25% similarity is being detected from the popular generic terms defining topic 21.

Tagging a document with the highest correlation probabilities can make document retrieval very efficient. For example, if a search were implemented for documents related to labour laws, document 2 would likely appear in the top results. But if a search were implemented for documents related to automobiles, document 2 would likely appear in the last few results.

4.5 Dynamic Topic Modeling

The DTM is a sequential extension of the LDA model. It requires one more parameter called the *time_slice,* which is a list containing the number of documents present in each time frame. The registration year of all documents in the dataset spans from 1951 to 2018. We can arrange the documents in chronological order of registration year and feed them into the *time_slice* parameter.

In the first experiment, we divided the documents into 3 time frames and ran the DTM for 25 topics. Below is a set of the top 10 unordered words for a single topic.

Time, Security, Cannabis, Establishment, Amend, Medical, Producer, Registration, Blood, Distribute

Based on the words above, we can predict that the topic resembles the following acts: Food and Drugs Act Controlled Drugs and Substances Act Cannabis Act/regulations, Cannabis Act/regulations, Tobacco and Vaping Products Act/regulations

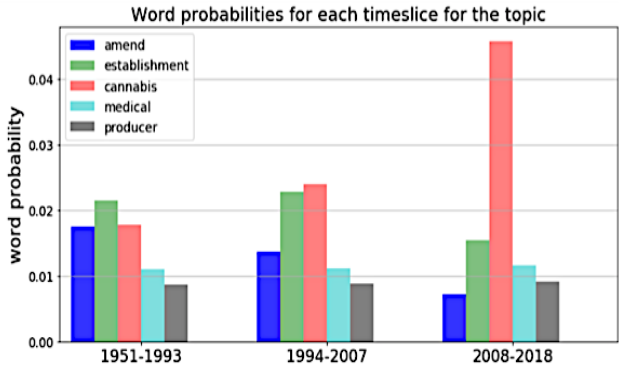


Figure E: Word probabilities of the top 5 words in 3 *time_slices*

It's interesting to see the dramatic increase in word probability of the term *"cannabis"*. In 2018 cannabis was legalized in Canada for both medical and recreational purposes. Hence, there are more policies referring to that particular term.

In the second experiment, we implemented a similar study, except we divided the documents into 10-year windows and ran the DTM for 25 topics. Below is a set of the top 10 unordered words for a single topic.

Firearm, License, Request, Issue, Permit, Chemical, Quantity, Registration, Import, Export

Based on the words above, we can predict that the topic resembles the following acts: Firearms Act/ Regulations, Chemical Weapons Convention Implementation Act. It's interesting to see the fluctuation of term probabilities given finer time slices. For example, the probability of word "*permit*" was roughly constant in the first 3 time slices but suddenly increased from ~0.025 to ~0.045 in between the 1981-1990 time slice. This could reflect the rise of stricter gun control policies in Canada. We can see similar variations for words *"export"* and *"firearm"*. Another point to note is that the term "*import*" emerged later into the list of top words in between the 1999-2000 time slice. Perhaps, Canada is beginning to take in more imported artillery during this time frame and policies are being created to regulate their use.

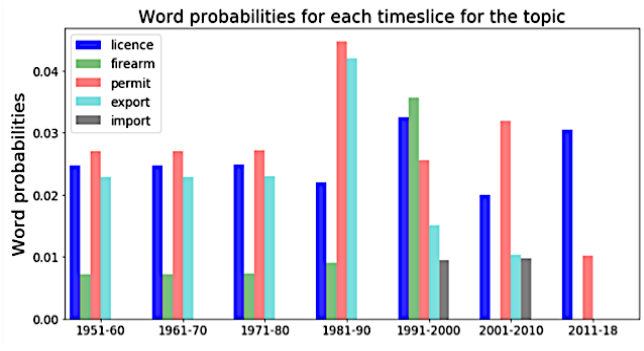


Figure F: Word probabilities of the top 5 words in 10 time slices.

5. Results

This section will be used to summarize the major findings of our research.

1. The coherence score can be used as metric to evaluate the interpretability of the topics generated via LDA model. Higher scores are generally favourable and it's ideal to select a value of *"k"* (i.e. number of topics) that maximizes the coherence score. We were able to achieve a high score of 0.35 with 25 topics for our analysis.
2. The LDA model is an unsupervised machine learning technique for which the validation of accuracy is mainly subjective. Hence, the interpreter needs to have an understanding of the dataset prior to generating topics. In this paper, we used our knowledge of the various types of acts in our dataset and attempted to find the resemblance of acts against the topics generated.
3. The pyLDA package in Python offers a powerful interactive tool to visualize the topics generated via LDA. We were able to use the tool to evaluate our topic content and identify clusters of topics.
4. The correlation between a document and topic is measured by correlation probability, but we can't group documents by these values unless they are significantly large. Correlation probabilities can instead be used to 'tag' documents with possible topics that reflect a document's content.
5. The DTM implementation allowed us to visualize the evolution of some topics over time. However, it did not capture new topics that may have emerged in mid-timeframe.

6. Conclusion and Remarks

pyLDA visualizations is an extremely powerful tool for depicting topics generated via LDA. The interactivity feature allows us to identify the flaws and improvements that need to be made. In particular, we can see which topics are weak based on the composition of vocabulary. Some topics were

mostly composed of popular generic terms that added little to no value to the overall topic semantics. As a future advancement, we would need to fine-tune our corpus to remove such terms. This would help us isolate topics more precisely. Well-defined topics lead to accurate tagging of documents. A fine-tuned corpus may also benefit the DTM model. We may detect new trends, which were otherwise lost in a sea of generic terms.

7. References

[1.] **Topic Coherence To Evaluate Topic Models**, http://qpleple.com/topic-coherence-to-evaluate-topic-models/

[2.] **Machine Learning Plus: Topic Modeling in Python with Gensim**, https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/

[3.] **Medium: Topic Modelling for Finding Similar Contracts**, https://medium.com/@dudsdu/topic-modelling-for-finding-similar-contracts-

[4.] **Legal Documents Clustering and Summarization using Hierarchical Latent Dirichlet Allocation -** *Ravi Venkatesh - IAES International Journal of Artificial Intelligence (IJ-AI) – 2013*

[5.] **Latent Dirichlet Allocation -** *David M. Blei, Andrew Y. Ng and Michael I. Jordan. University of California, Berkeley - 2003*

[6.] **A heuristic approach to determine an appropriate number of topics in topic modelling -** *Weizhong Zhao-James Chen-Roger Perkins-Zhichao Liu-Weigong Ge-Yijun Ding-Wen Zou - BMC Bioinformatics – 2015*

[7.] **Sentiment-topic modeling in text mining -** *Chenghua Lin-Ebuka Ibeke-Adam Wyner-Frank Guerin - Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery – 2015*

[8.] **A Survey of Topic Modeling in Text Mining** - *Rubayyi Alghamdi - Khalid Alfalqi - International Journal of Advanced Computer Science and Applications - 2015*

[9.] **Automatic Evaluation of Topic Coherence -** *David Newman - Jey Han Lau - Karl Grieser - Timothy Baldwin - NICTA Victoria Research Laboratory Australia - 2012*

8. Appendix

All files below can be accessed on the following link:
https://github.com/eispat28/LDA-Topic-Modeling-Legal-Documents

1. StopWords.txt
2. Dataset
3. pyLDA visual
4. LDA code
5. DTM code