DS 8008 - NLP Project

Topic Modeling of Canadian Regulations

Eisha Patel
Chaitra Hosmani

Overview

1. **Introduction**
2. **Related Work**
3. **Methodology**
   Data Preparation ~ Building the LDA Model ~ Visualizing LDA Topics ~ Document Grouping ~ Time Evolution of Topics
4. **Results**
5. **Conclusion**

Overview

1. **Introduction**
2. **Related Work**
3. **Methodology**
   Data Preparation ~ Building the LDA Model ~ Visualizing LDA Topics ~ Document Grouping ~ Time Evolution of Topics
4. **Results**
5. **Conclusion**

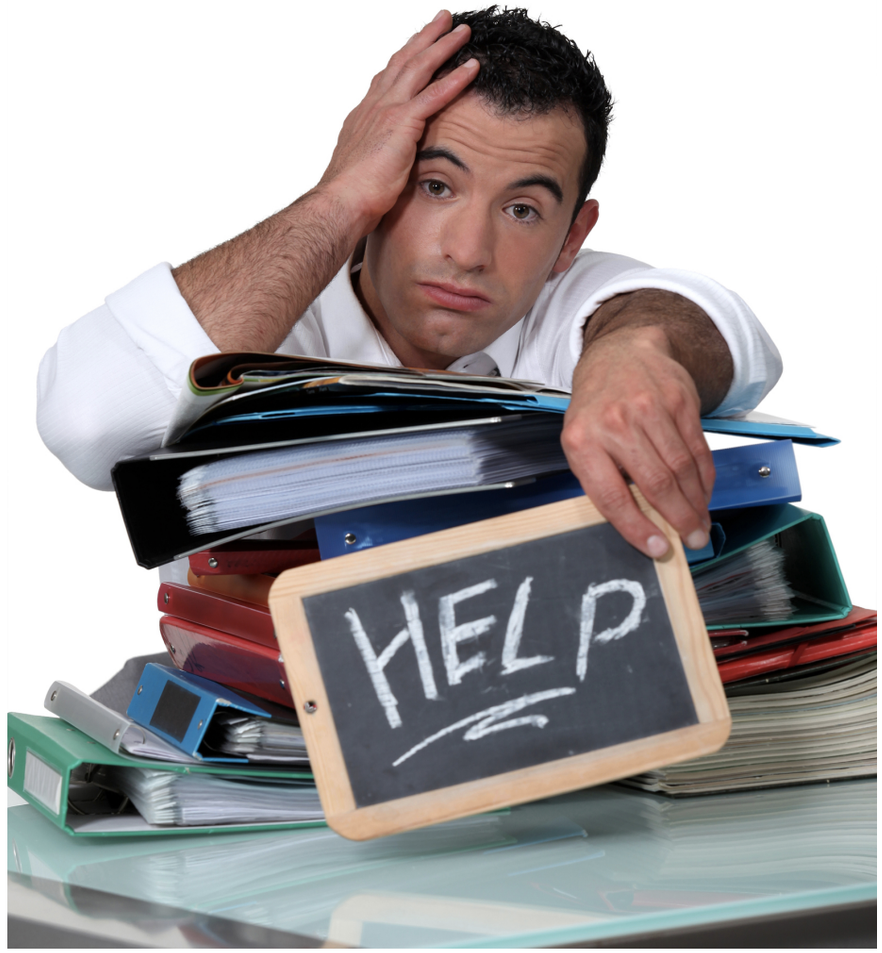
Introduction

It's a challenge for the Legal Community to go thru thousands of Legal documents to find what they need.

How can we make this less painful?

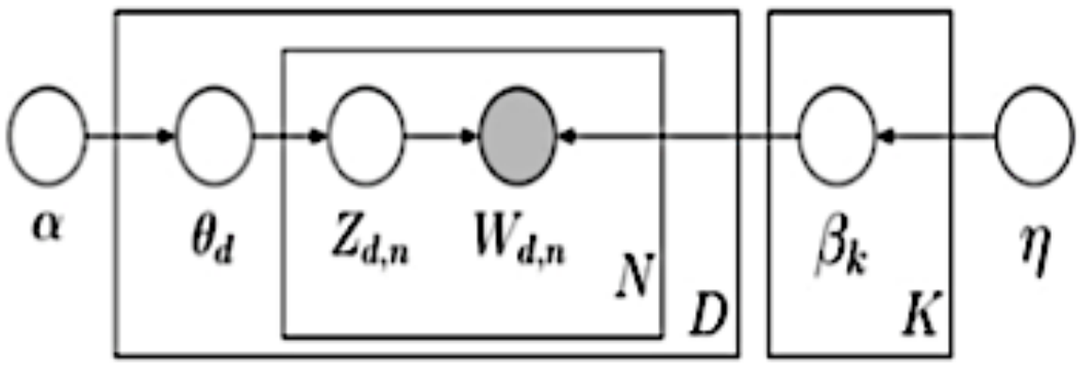
Problem Statement:

- How to simplify understanding of regulation documents?
- How can we analyse the evolution of regulations over time?



Introduction

- Topic Modeling is a set of techniques that aim to discover and annotate large archives of documents with thematic information.
- Unsupervised machine learning technique for which the validation of accuracy is mainly subjective
- Latent Dirichlet Allocation (LDA) - a continuous multivariate probabilistic model



θ — A distribution of topics, one for each document
Z — k Topics for each document
β — distribution of words, one for each topic
D — All the data we have (i.e. corpus)
α — A parameter vector for each document (document — Topic distribution)
η — A parameter vector for each topic (topic — word distribution)

Overview

1. **Introduction**
2. **Related Work**
3. **Methodology**
   Data Preparation ~ Building the LDA Model ~ Visualizing LDA Topics ~ Document Grouping ~ Time Evolution of Topics
4. **Results**
5. **Conclusion**

Related Work

1. **A Survey of Topic Modeling in Text Mining** [1]

   - Different methodologies for topic modelling methods : LSA, pLSA, LDA and Correlated Topic Model (CTM).
   - Topic Over Time (TOT), Dynamic Topic Models (DTM)

2. **Automatic Evaluation of Topic Coherence** [2]

   - Coherence score evaluates the quality of a given topic, in terms of its coherence to a human
   - **Intrinsic** (UMass) measures to compare a word only to the preceding and succeeding words respectively, need ordered words
   - **Extrinsic** (UCI): every single word is paired with every other single word. Uses pointwise mutual information (PMI)

[1] https://thesai.org/Downloads/Volume6No1/Paper_21-A_Survey_of_Topic_Modeling_in_Text_Mining.pdf
[2] https://mimno.infosci.cornell.edu/info6150/readings/N10-1012.pdf

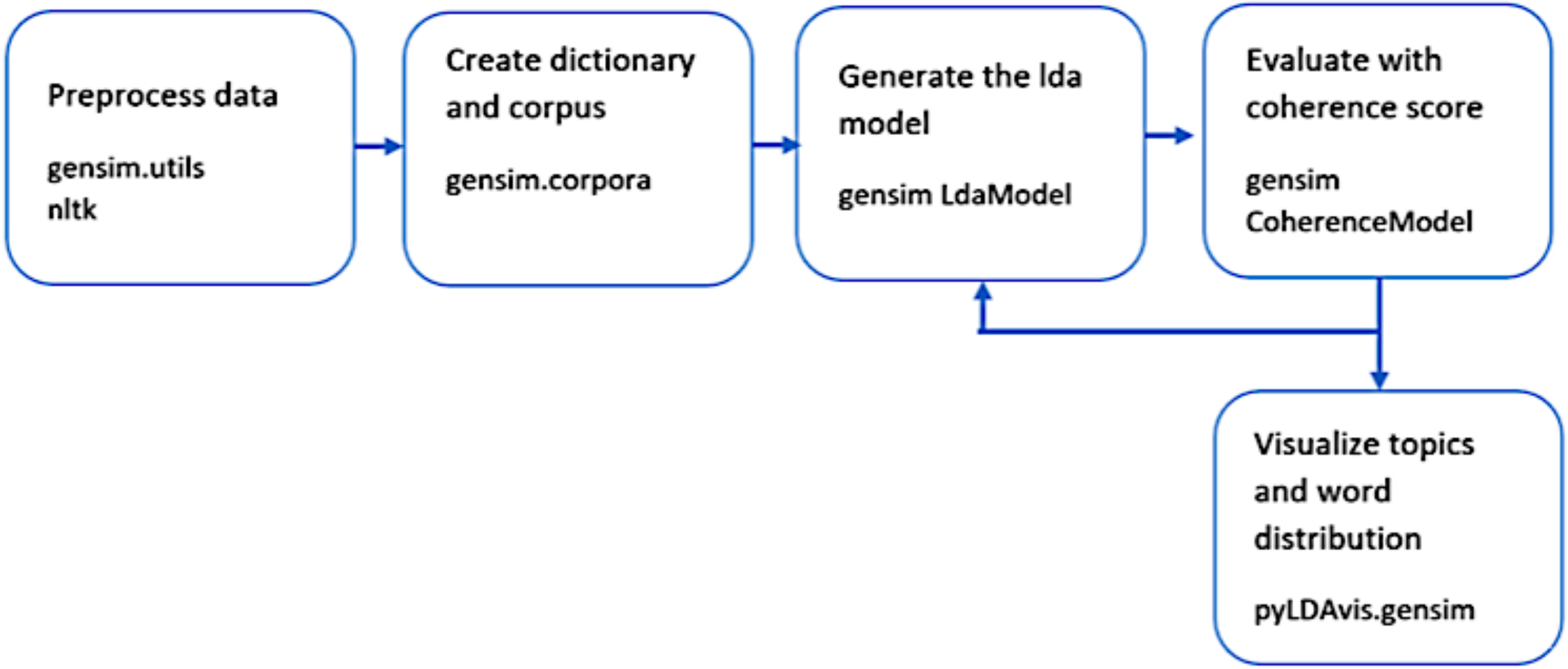
Overview

1. **Introduction**
2. **Related Work**
3. **Methodology**
   Data Preparation ~ Building the LDA Model ~ Visualizing LDA Topics ~ Document Grouping ~ Time Evolution of Topics
4. **Results**
5. **Conclusion**

Methodology: Data Preparation

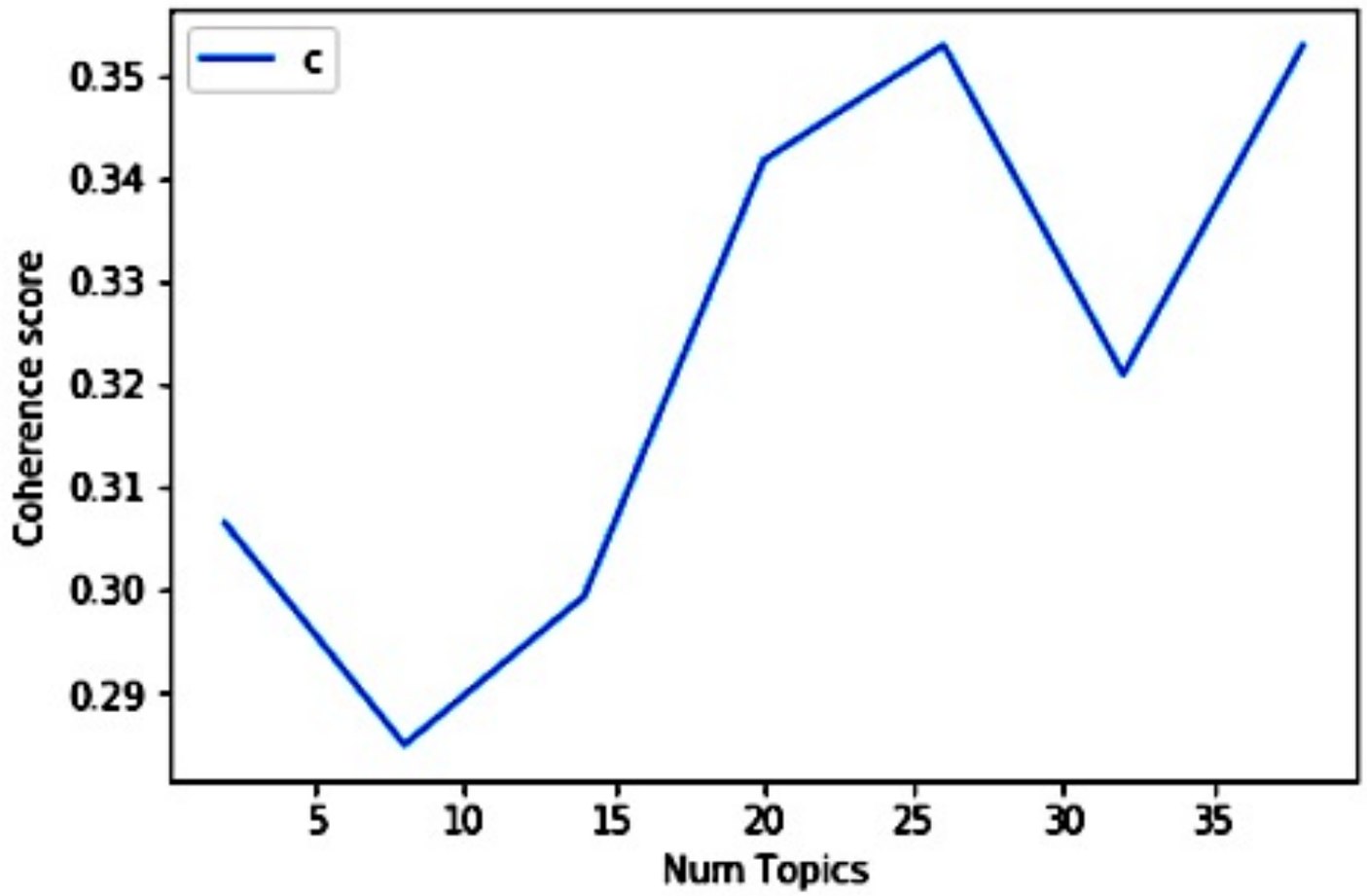
- Dataset contents:
  - `columns: [‘instrumentNumber','shorttitle','longtitle','modifiedyear', 'registrationyear',’consolidationyear','xrefxternal', ‘content']`
  - `rows: 2062`
- Extract the XML files using the code provided by Prof. Shariyar
- Reduce Dataset to required columns only
- Remove all rows with missing registration year
- Convert content column to list of documents
- Tokenize documents
- Remove stopwords
- Extract bigrams/trigrams
- Lemmatization

Methodology: Building the LDA Model



Methodology: Building the LDA Model

- **How many Topics are ideal?**
- Coherence score looks at the set of words generated in a topic and rates the interpretability of the topic
- We need to maximize this score



Methodology: Building the LDA Model

- What can we expect our topics to look like?
- Extract all unique Acts from the 'xrefxternal' using regular expression
- List of top 19 Acts with frequency of occurrence

AERONAUTICS ACT	122
FINANCIAL ADMINISTRATION ACT	81
AGRICULTURAL PRODUCTS MARKETING ACT	71
BANK ACT	69
FOREIGN MISSIONS AND INTERNATIONAL ORGANIZATIONS ACT	49
INSURANCE COMPANIES ACT	45
PUBLIC SERVICE EMPLOYMENT ACT	41
APPROPRIATION ACT	39
COOPERATIVE CREDIT ASSOCIATIONS ACT	36
EXPORT AND IMPORT PERMITS ACT	36
CANADA CONSUMER PRODUCT SAFETY ACT	36
TRUST AND LOAN COMPANIES ACT	35
CUSTOMS ACT	33
CANADIAN ENVIRONMENTAL PROTECTION ACT	30
CANADA SHIPPING ACT	26
CANADA TRANSPORTATION ACT	24
EXCISE TAX ACT	22
SPECIES AT RISK ACT	22
FOOD AND DRUGS ACT	19

Methodology: Building the LDA Model

- The Gensim LDA model take 2 inputs

  1) Dictionary of words

  2) Term-Document frequency matrix (ie. corpus)

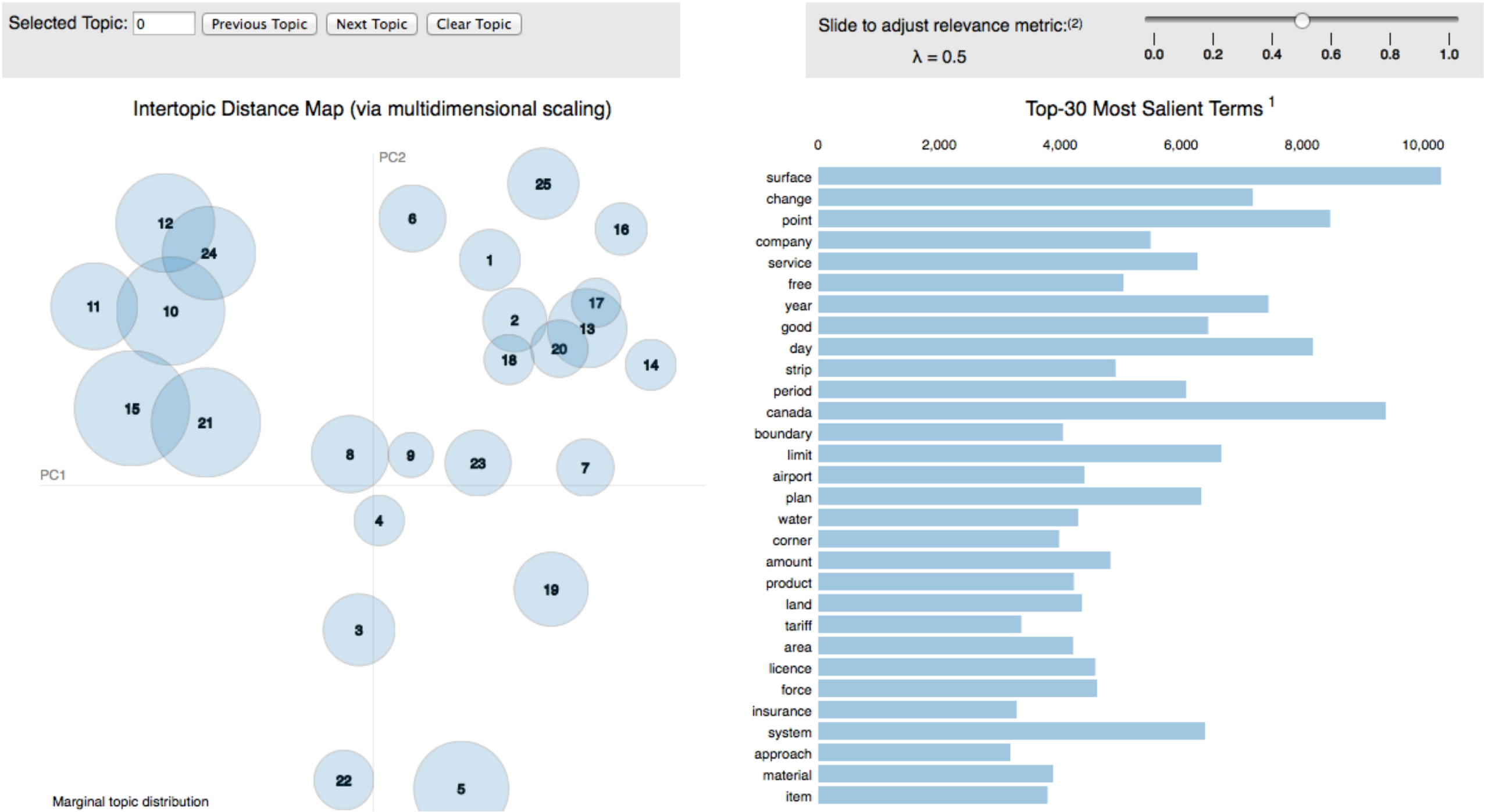
```
lda_model = gensim.models.ldamodel.LdaModel(corpus = corpus,
                                            id2word = dictionary,
                                            num_topics = number_of_topics,
                                            random_state = 150,
                                            update_every = 1,
                                            chunksize = 75,
                                            passes = 50,
                                            alpha ='auto',
                                            per_word_topics=True)
```

Methodology: Building the LDA Model

Topic 7	Topic 8	Topic 14	Topic 22	Topic 24
free	product	service	boundary	amount
tariff	cannabis	public	water	pay
good	originate	group	area	fee
item	origin	period	land	payment
custom	manufacture	employ	direction	payable
duty	sale	employee	activity	interest
import	container	employment	measure	respect
January	display	employer	zone	rate
sor	exporter	cease	plan	day
Canada	package	department	northwest_terri	sor

file:///Users/eishapatel/Desktop/DS%20-%208008/DS%208008%20-%20project/vis25.html

Methodology: Visualizing LDA Topics

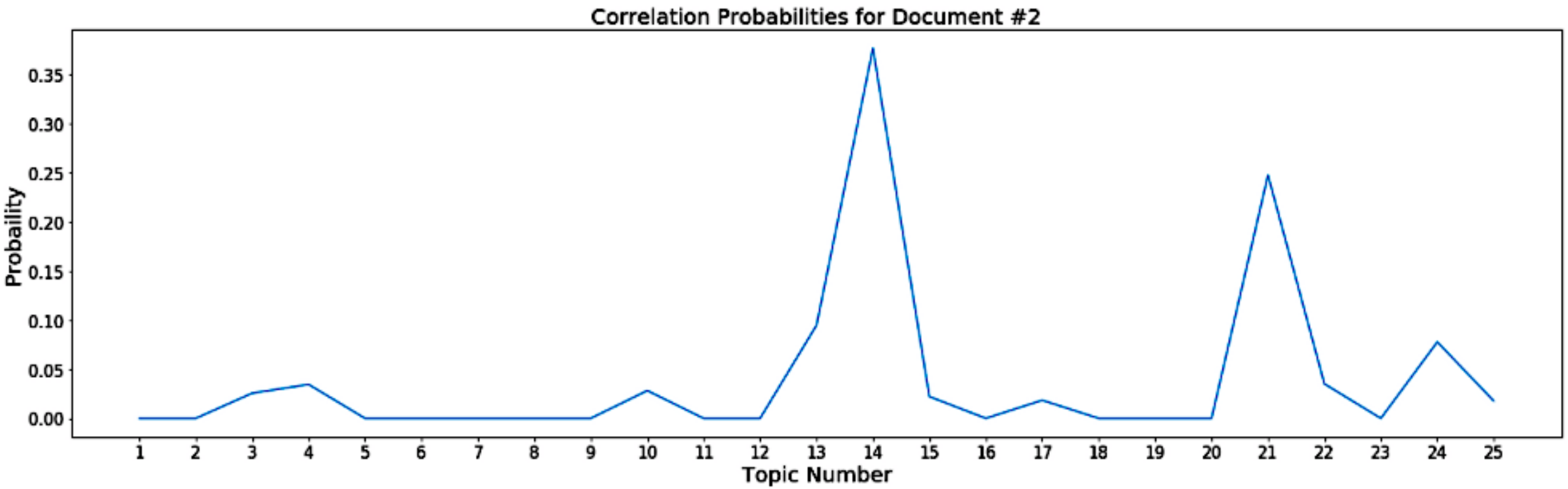


Methodology: Document Grouping

- Each document can be described by a list of topics and correlation values that relates the topics to the document.
- The list of topics and correlation is unique among documents and can also be considered the *signature* of that document.
- Since one document can be highly correlated to multiple topics, we can create soft clusters of documents
- Example: document 1 and it's probabilities

	document	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	. . . 25
0	0	0.049088	0.066063	0.041108	0.023798	0.031143	0.070530	0.048153	0.063373	0.026528	0.053886	

Methodology: Document Grouping



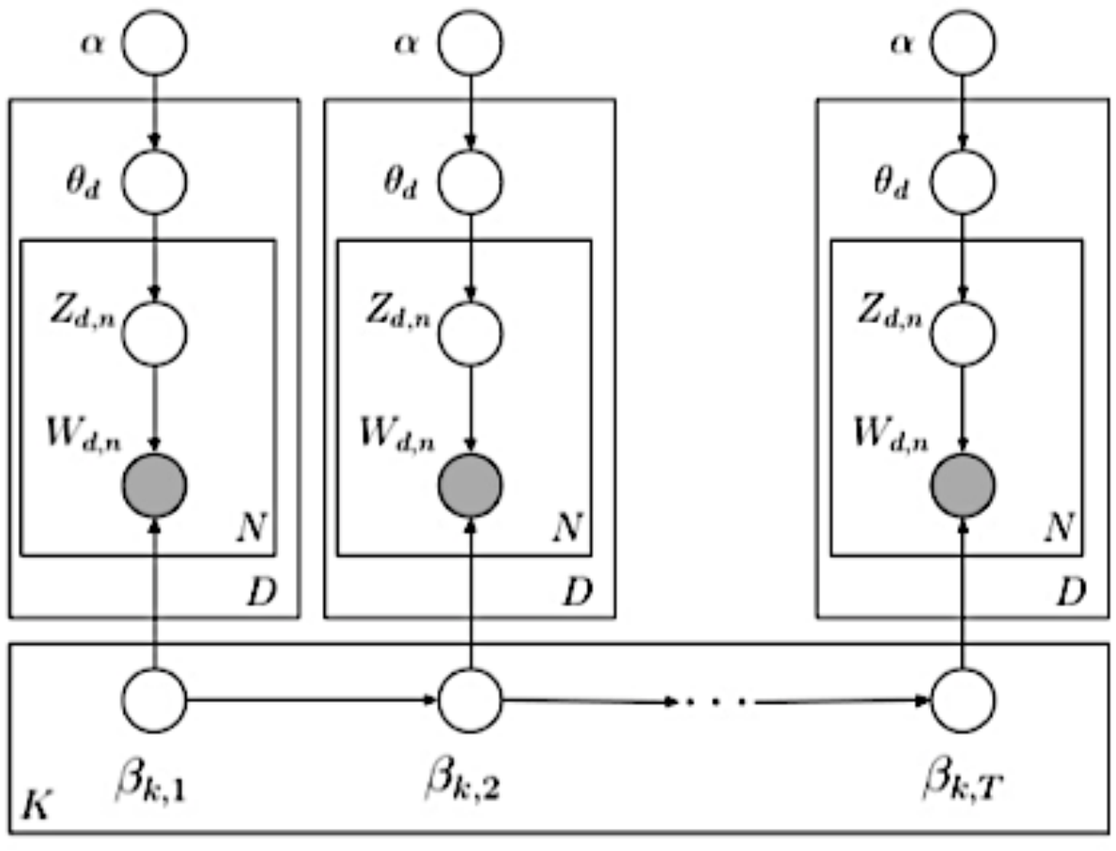
.

PUBLIC SERVICE LABOUR RELATIONS ACT'

```
['public', 'service', 'labour_relation', 'designate', 'staff', 'public', 'fund',
  'canadian', 'force', 'separate', 'employer', 'purpose', 'excellency', 'governor',
  'general', 'council', 'recommendation', 'minister', 'national_defence', 'pursuant',
  'public', 'service', 'staff_relation', 'designate', 'staff', 'public', 'fund',
  'canadian', 'force', 'separate', 'employer', 'purpose', 'public', 'service',
  'staff_relation', 'provision', 'reference', 'context_requir', 'reference', 'public',
  'service', 'labour_relation', 'provision', 'define', 'statutory_instrument',
  'parliament', 'provision', 'refer', 'read', 'reference', 'federal', 'public',
  'sector_labour_relation']
```

Methodology: Time Evolution of Topics

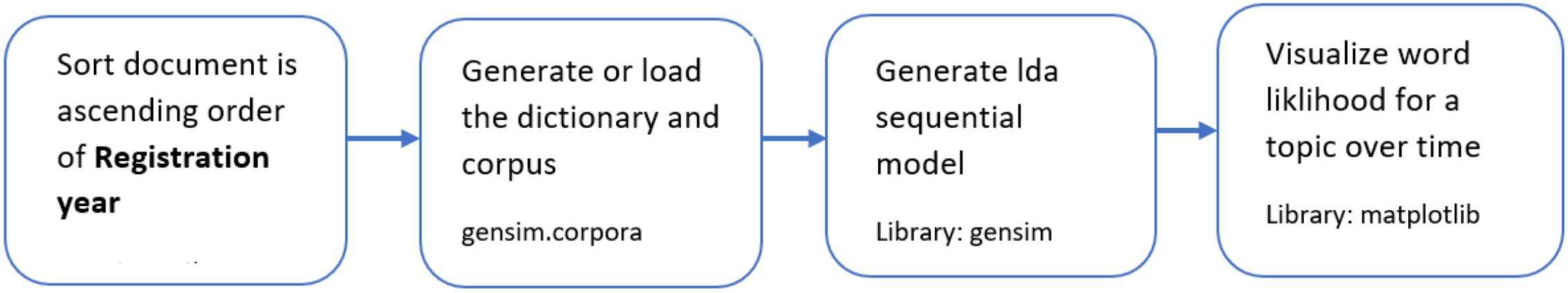
Dynamic Topic Modelling



- Captures the evolution of topics in a sequentially organized corpus of documents
- Documents are grouped by time slice (e.g.: years): documents order matter
- Exploration of a large document collection in new way

Methodology: Time Evolution of Topics

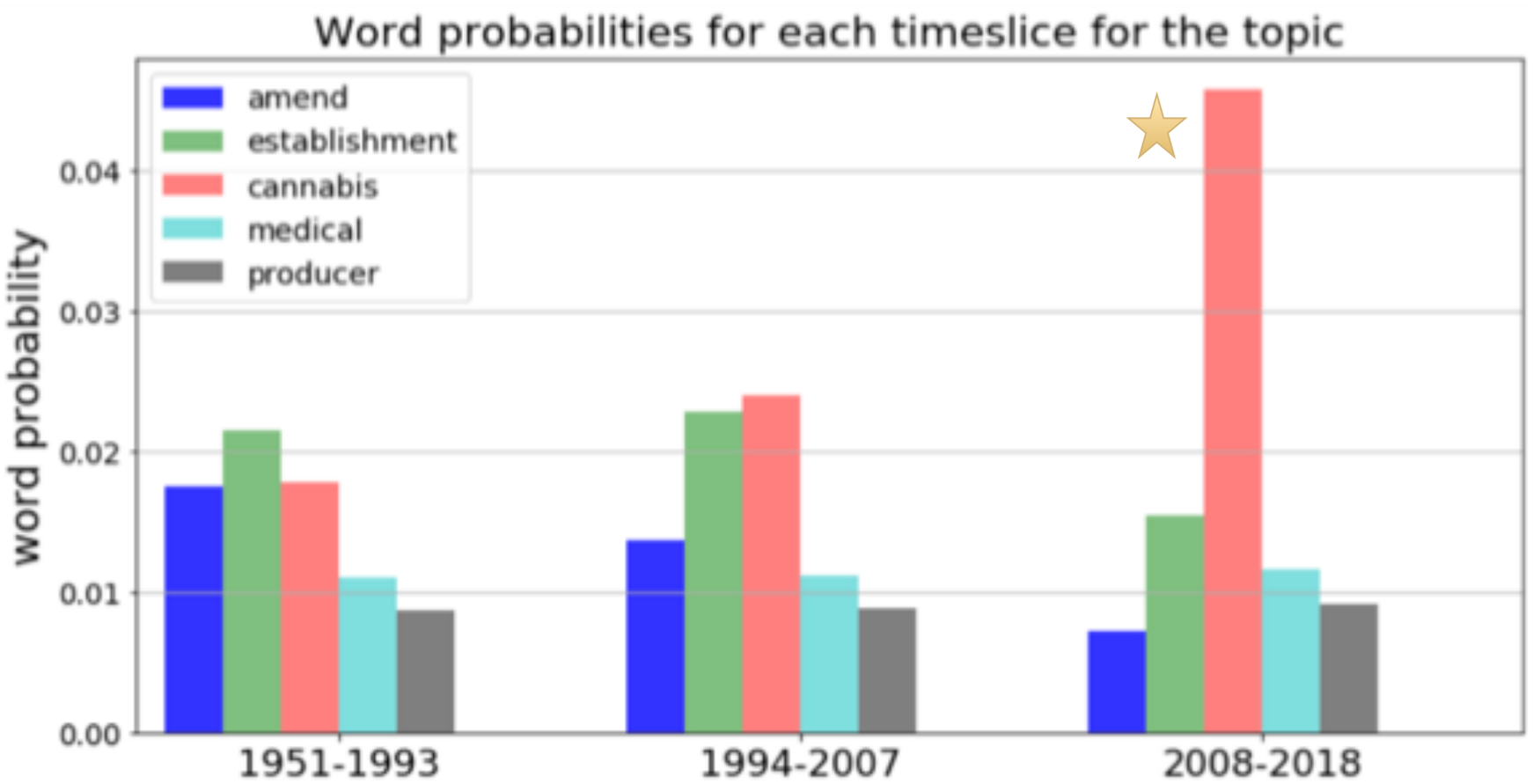
Dynamic Topic Modelling Approach



Experiment 1: Evolution with 3 time slices

Topic A

time
security
cannabis
establishment
amend
medical
producer
registration
blood
distribute

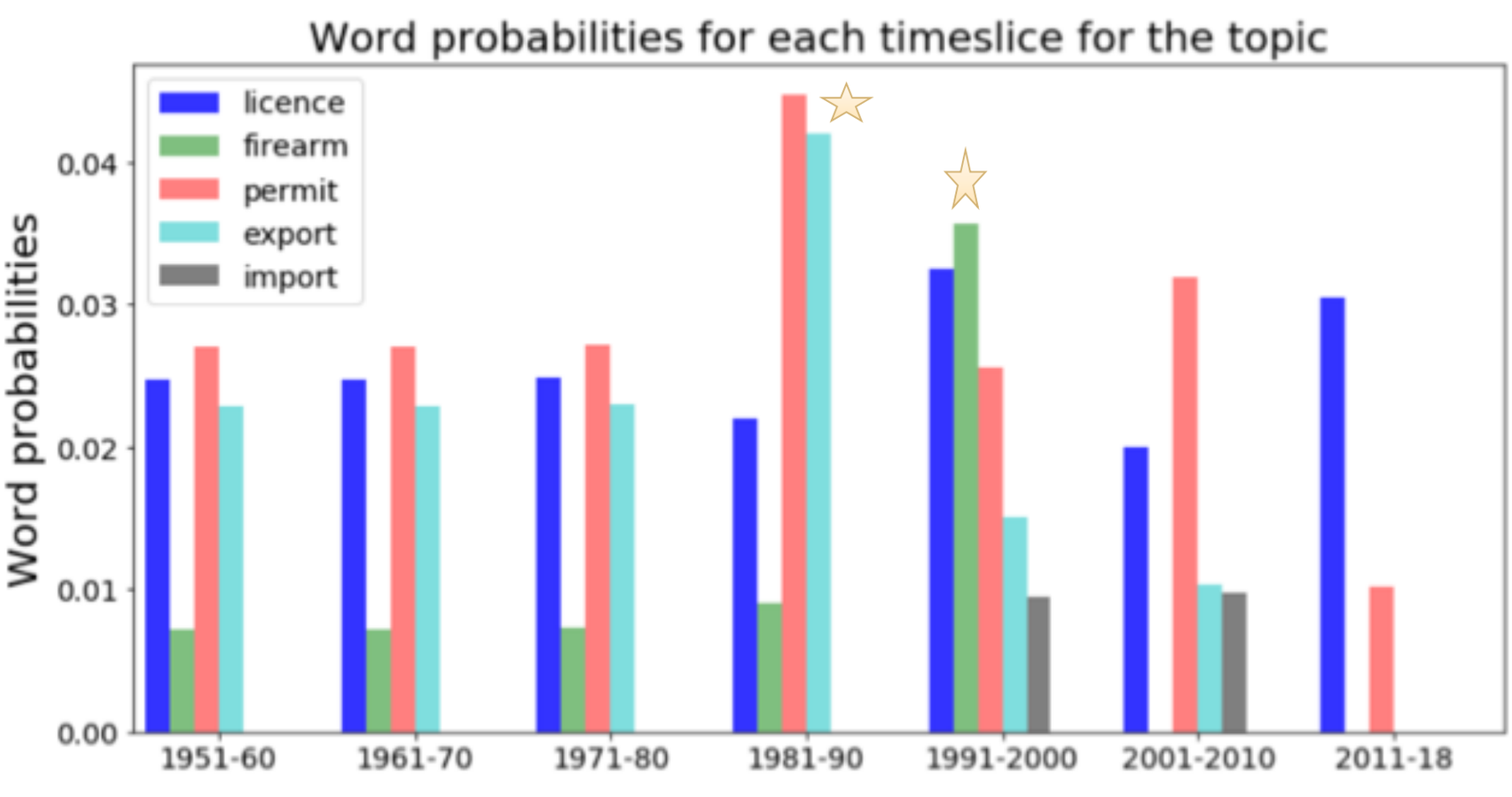


Food and Drugs Act Controlled Drugs and Substances Act Cannabis Act/regulations, Cannabis Act/regulations, Tobacco and Vaping Products Act/regulations

Experiment 2: Evolution with 10 years window

Topic B

firearm
licence
request
issue
permit
chemical
quantity
registration
import
export



Firearms Act and Regulations
Chemical Weapons Convention Implementation Act

Overview

1. **Introduction**
2. **Related Work**
3. **Methodology**
   Data Preparation ~ Building the LDA Model ~ Visualizing LDA Topics ~ Document Grouping ~ Time Evolution of Topics
4. **Results**
5. **Conclusion**

Results

- Coherence Scores can be used to determine how many topics to create
- Topics generated reflect the ‘Acts’ in the dataset
- pyLDAvis shows the clusters of topics
- 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 be used to ‘tag’ documents with possible topics
- Able to analyze evolution of some topics over time
- It did not capture new topic which might have emerged in between entire time period

Overview

1. **Introduction**
2. **Related Work**
3. **Methodology**
   Data Preparation ~ Building the LDA Model ~ Visualizing LDA Topics ~ Document Grouping ~ Time Evolution of Topics
4. **Results**
5. **Conclusion**

Conclusion

- Fine-Tuning the Corpus:
  - Remove words that occur in too many topics and do not contribute to the semantics of a topic
  - This will improve the correlation between a document and its relevant topics
  - It will also improve DTM analysis
- Implement other visualization techniques for DTM to observe trends

References

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

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

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

4. **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*

5. **Topic Modeling in Python with Gensim -** https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/

6. **Topic Modelling For Finding Similar Contracts -** https://medium.com/@dudsdu/topic-modelling-for-finding-similar-contracts-df00b3aea8b2

Thank You