Topic Modelling of Legal Documents via LEGAL-BERT

Raquel Silveira[1], Carlos Gustavo O. Fernandes[2], João A. Monteiro Neto[3], Vasco Furtado[4], and José Ernesto Pimentel Filho[5]

[1]Federal Institute of Education, Science and Technology of Ceará, Tianguá, Ceará, Brazil
[2]University of Fortaleza and Banco do Nordeste do Brasil S.A, Fortaleza, Ceará, Brazil
[3]University of Fortaleza Law School, and FUNCAP, Fortaleza, Ceará, Brazil
[4]University of Fortaleza, Fortaleza, Ceará, Brazil
[5]Federal University of Paraíba, João Pessoa, Paraíba, and FUNCAP, Fortaleza, Ceará, Brazil

DPDI - DIRETORIA DE PESQUISA, DESENVOLVIMENTO E INOVAÇÃO

Schedule

- Introduction
- Related Works
- Methodology
  - Data Collection
  - Topic Modelling
  - Evaluation
- Results and Analysis
- Conclusions

Introduction

- Topic Modeling applied to Natural Language Processing (NLP)
- Topics represent the theme or subject of the text
- Increasing the volume of legal information requires automatic processing
- Topic modeling may contribute to making efficient the analysis of legal documents

Introduction

- In this article we investigate stochastic topic modeling approaches for legal documents
- We used BERTopic (Marteen, 2020) with representation of legal documents according LEGAL-BERT (Chalkidis and Fergadiotis, 2020).
- We extended the representation of the document, with the insertion of text describing the USCode cited in the document.

Related Works

- Latent Dirichlet Allocation (LDA) has been used to model legal corpora.
- Araújo and Campos (2020) use LDA to model Extraordinary Resources received by the Supreme Court of Brazil.
- Neill et al., (2017) qualitatively evaluate the performance of topic models (LDA, LSA, NMF, HDP) in the British legislation.
- Remmits (2017) evaluates the use of the LDA in extracting precise and useful topics of Dutch case law.
- LDA has several weaknesses: require the number of topics, custom stop-word lists, stemming, lemmatization, and ignore the ordering and semantics of words.

Related Works

- Distributed representations are gaining popularity due to their ability to capture the semantics of words and documents.
- Thompson and Mimno (2020) use contextualized language models (BERT, GPT-2, and RoBERTa) and k-means algorithm to produce topics of Supreme Court of the United States legal opinions.
- Angelov (2020) developed Top2Vec, a model that uses semantic embeddings to find topic vectors.

Related Works

- LEGAL-BERT (Chalkidis and Fergadiotis, 2020) has the property of capturing the characteristics of the language for the legal domain.
- BERTopic is a topic modeling technique that uses HDBSCAN, and class-based TF-IDF (c-TF-IDF) to allow easy interpretable topics (Marteen, 2020).
- We are not aware of publications examining the topic modeling of legal documents considering the representation of the document from language models of the legal context.

Related Works

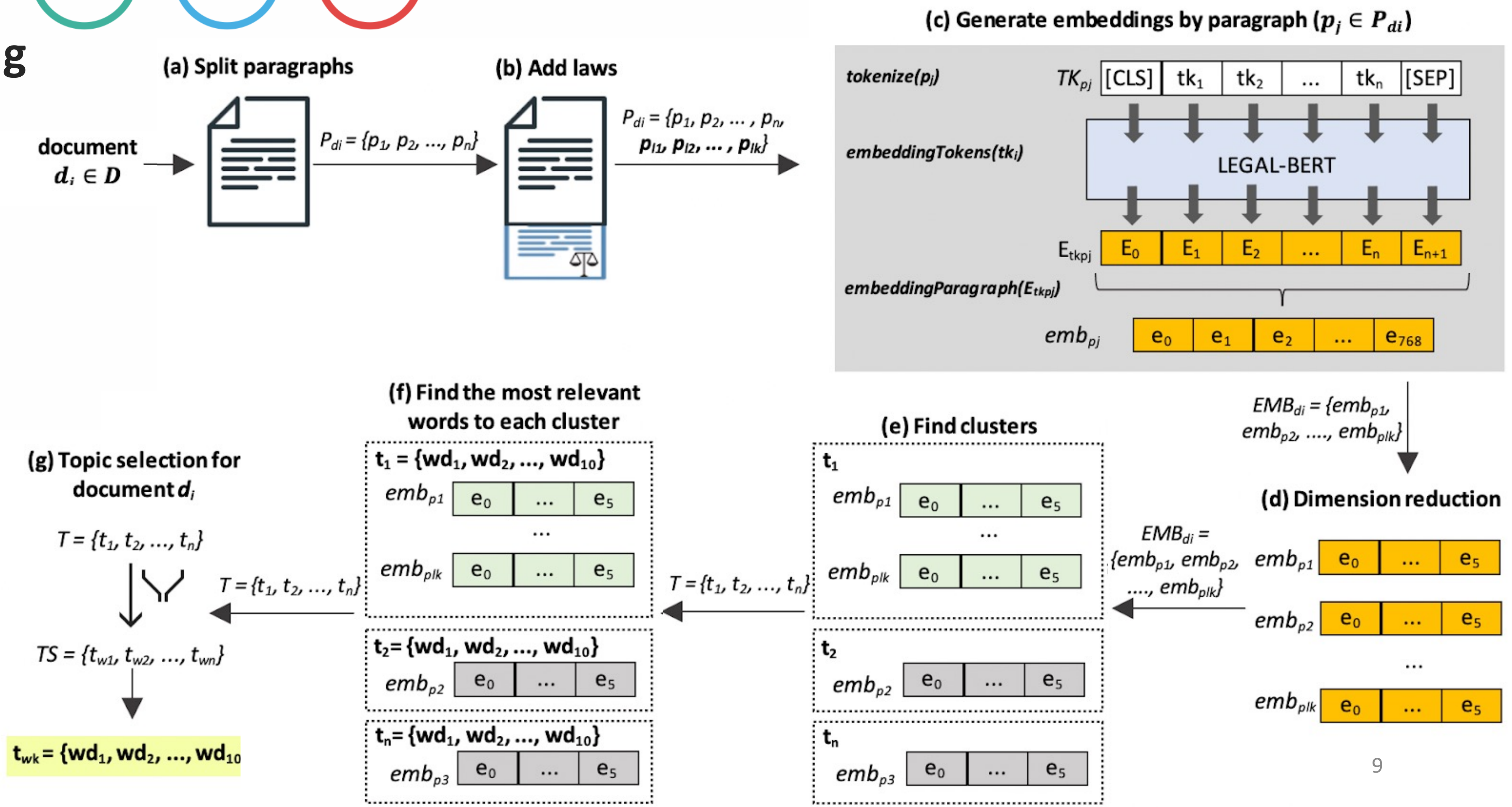
Data Collection

- Set of legal documents from the Cornell Legal Information (Cornell LII)'s repository of Historic US Supreme Court Decisions representing the list of landmark court decisions* in the United States.
- 314 legal cases were selected randomly.
  - Each cases classified division and subdivision
    - For example, Individual Rights (discrimination based races, sex, abortion, Freedom), Criminal Law (capital punishment), First/Second Amendments

* Landmark court decisions in the United States substantially change the interpretation of existing law

Methology

Topic Modelling



Methology

Evaluation

- Two variations of the approach were evaluated:

  (1) the document is represented only by the paragraphs of document

  (2) the document is represented by the paragraphs of document and the text of the laws cited in the document.

- We carry out a qualitative assessment under the criterion of interpretability.
- Two experts in the legal field performed a manual inspection on the set of words most representative of the topics selected by the model.

Results and Analysis

- In approximately 8% of the documents the approach fails to model the topics (representation of the document only with the paragraphs that compose it)
- By expanding the representation of the document with the insertion of the text of the cited laws, only 5% of the documents had no topics modeled
- **Qualitative evaluation:** 84.6% of the topics selected by the model correspond to the main theme of the document

Table 1
Topics modeled to legal documents.

ID	Division	Subdivision	Topics
D1	Criminal law	Capital punishment	death, execution, risk, id, injection, penalty, pain, lethal, punishment, protocol
D2	Criminal law	Detainment of terrorism suspects	court, jurisdiction, habeas, states, united, united states, courts, district, eisentrager, writ
D3	Equal Protection Clause	Passengers and Interstate Commerce	statute, interstate, state, commerce, court, passengers, led, states, sct, virginia
D4	Federal Native American law	Federal Native American law	indian, non indians, jurisdiction, non indian, try, courts, congress, tribes, indian tribes, try nonindians
D5	First Amendment rights	Amish	amish', 'education', 'children', 'religious', 'school', 'life', 'state', 'child', 'parents', 'compulsory
D6	First Amendment rights	Freedom of speech and of the press	sct, states, united states, present, led2d, danger, present danger, clear present
D7	Individual rights	End of life	new, suicide, treatment, medical, health, sct, york, new york, ann, patients
D8	Intellectual Property	Copyright/Patents	copyright, work, facts, original, works, protection, originality, act, author, telephone
D9	Tax Law	Federalism	direct, constitution, tax, taxes, apportioned, apportionment, cases, rule, present, indirect
D10	Women's rights	Birth control and abortion	abortion, procedure, state, fetus, court, medical, law, statute, dx, id

Results and Analysis

A word cloud of top-30 words extracted by model of a legal document dealing with "capital punishment".

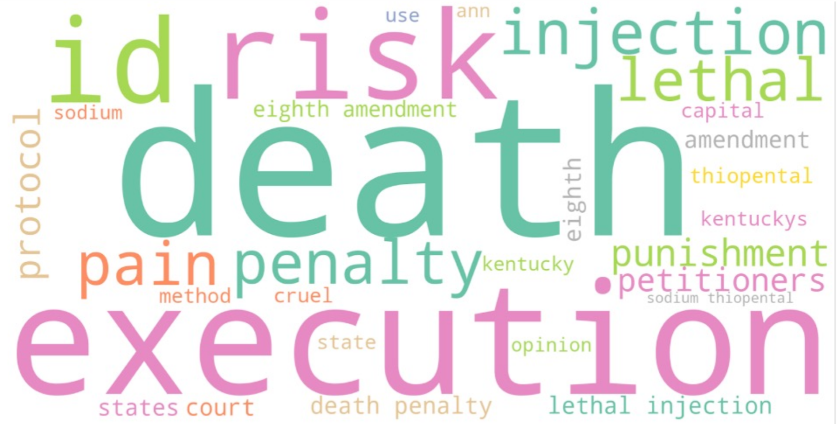


Figure 1: Most relevant words for the topic, according to c-TF-IDF.

Conclusions

- We propose the use of BERTopic to build thematic models of legal documents.
- We represent the text contextually from the LEGAL-BERT and provide information about the laws mentioned in the document.
- From a qualitative assessment, the approach reveals topics consistent with the document's theme.
- This preliminary approach can be used as a baseline for future works.

Conclusions

Future Works

- Explore different strategies for choosing the topics of a document
- Quantitatively evaluate the interpretability and coherence of the topics
- Compare the proposed approach with other approaches of the state of the art.
- Extend the approach to clustering documents according to the modeled topics.

Thank you!

DPDI - DIRETORIA DE PESQUISA, DESENVOLVIMENTO E INOVAÇÃO