Classification of Legal Text

Krithika Iyer (ksiyer)

Abstract

This project explores the feasibility of utilizing NLP techniques for the classification of legal opinions. Legal opinions and texts consist of long running sentences (or long spans of text) that do not conform to standard English linguistic or grammar patterns. NLP techniques and algorithms developed for processing standard linguistic sentences and documents often produce incorrect results while processing legal documents. Precedent is an important guiding principle in the legal domain. This work is part of my larger effort to automate analysis of legal opinions, which holds much promise towards developing systems that provide equal access to law. Automated systems allow lawyers to build cases more quickly, such as aiding overworked public defenders and provide access to the law to a greater number of people (such as people reading up on the law to defend themselves).

Introduction

The reasoning behind particular legal opinions and rulings are often cited in other legal cases to further the legal theory put forward by attorneys. The bedrock of the legal system are the precedents upon which disposition of legal disputes are based. Every legal opinion and ruling needs to be classified and cataloged so that they can be accessed during searches. Classification of legal opinions and rulings is a very time consuming and tedious manual task. Current NLP based approaches towards classification of legal documents are often guided by the annotations by human experts. Human experts create predefined categories

Related Work

Classification of legal documents is a relatively new field and many of the related research are new and in progress. A few efforts aimed at modifying the attention mechanism of Transformer based neural nets are discussed under the Future Work subsection. Classification of legal dockets using SVM is detailed in [1]. Early efforts aimed at classifying legal text described in [2, 3, 4]. Efforts aimed at classifying medical documents [5] provide some guidance for designing systems aimed at classifying legal documents. Knowledge graph based approaches have also been used to classify legal documents [6].

Approach

The reasoning behind legal decisions can be quite complex. For example, the famous *Roe V. Wade* (1973) case that legalized abortion in the US may naturally fit into human expert curated categories like medical, reproductive rights, and abortion. However an analysis of the reasoning behind the decision would put the decision in the “Privacy” category. To capture such subtleties in reasoning, the project will utilize both unsupervised and supervised learning during the classification of legal opinions. This project utilizes a Latent Dirichlet Allocation (LDA) model to extract topics in a given legal opinion and then utilize logistic regression (LR) to align and classify LDA models of the document to the human curated categories. The unsupervised LDA based system will be compared with (i) document vector embedding and (ii) Bidirectional Encoder Representations from Transformers (BERT) based neural networks.

Dataset and Features

The Supreme Court Database at the Washington University (http://scdb.wustl.edu/) contains the decisions and about 200 pieces of additional metadata about every case decided by the US supreme court between 1791 and 2018 terms). The Textacy package provides access to approximately 8,200 decisions from this database [7]. The textacy data also contains the manually curated classification codes for *issue* and *issue_area.* The classification codes are described in detail at [8].

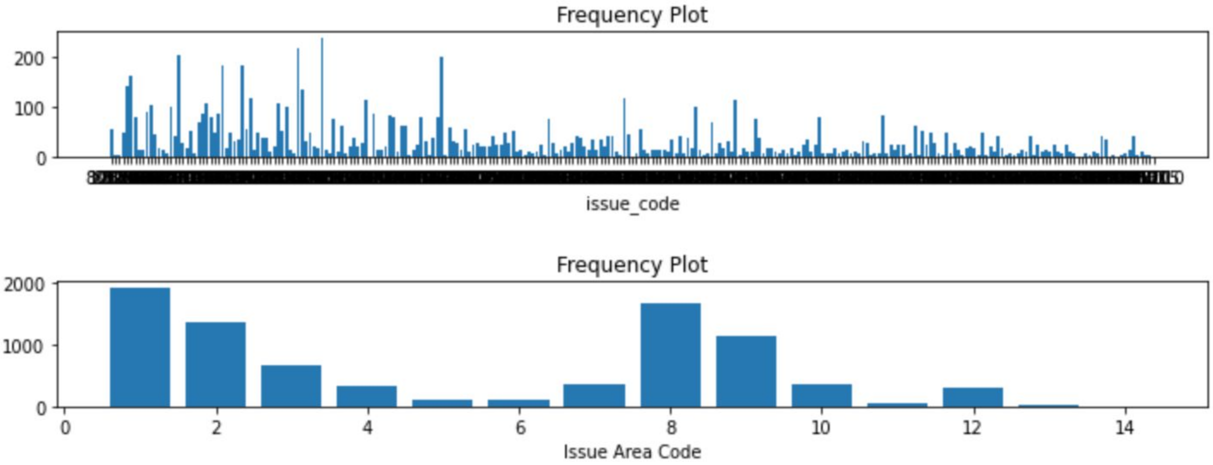

Figure 1. Expert classification of the 8,100 US Supreme Court decisions used in this project.

Methods

The reasoning behind legal decisions can be quite complex. For example, the famous *Roe V. Wade* (1973) case that legalized abortion in the US may naturally fit into human expert curated categories like medical, reproductive rights, and abortion. However an analysis of the reasoning behind the decision would put the decision in the “Privacy” category. To capture such subtleties in reasoning, the project will utilize both unsupervised and supervised learning during the classification of legal opinions. This project utilizes a Latent Dirichlet Allocation (LDA) model to extract topics in a given legal opinion and then utilize logistic regression (LR) to align and classify LDA models of the document to the human curated categories. The Gensim package was used for extracting LDA topics [9]. The unsupervised LDA based system will be compared with (i) document vector embedding and (ii) Bidirectional Encoder Representations from Transformers (BERT) based neural networks. Huggingface library [10] was utilized for BERT based experiments.

Experiments

Three distinct sets of experiments were carried out:

1. LDA and Logistic Regression: Unsupervised LDA probabilistic approach is used to generate a series of latent topics in the corpus of legal decisions. Each decision is represented by a term frequency-inverse document frequency (tf-idf) matrix. The gensim package was used to generate the LDA model. LR was used to classify decisions into one of 279 issue_code or 15 issue areas.

2. Doc2Vec and Logistic Regression: Doc2vec or document vectors map semantic meaning from a variable length document to a fixed size vector. This part of the experiment is aimed at testing if the nuances of the reasoning behind a decision can be captured in a fixed-sized vector.

3. BERT Neural Nets: Transformers incorporating attention have demonstrated impressive results in the NLP domain. This explores whether they can deliver similar results in the legal domain, especially with transfer learning from structured / standard text to non-standard legal text.

Results and Discussion

The LDA based model is considered a baseline model. The classification accuracy for the 279 issue_code is around 0.133. When only the issue_areas (15 categories) are considered, the classification accuracy jumps to 0.47. The details of the results for each the 279 issue_code and 15 issue_areas may be viewed in the URL to the Google Collaboratory notebook listened in Appendix A. The 279 issue code is analyzed in the first part and the 15 issue_areas are addressed in the second part of the notebook.

Model	15-Labels	279 Labels
LDA + LR	0.13	0.47
Doc2Vec + LR	0.48	0.63

Table 1. Classification accuracy.

The paragraphs embeddings extracted by Doc2Vec is used in the second experiment. In the initial phase the entire legal opinion document is used as input to the model. The embedding vectors extracted by the Doc2Vec model used by LR for further classification of the legal opinions. While this approach improves the overall accuracy (compare to baseline LDA + LR model (as shown in Table 1)), more detailed results for each category is shown in Table 2. The results clearly indicate the difficulties involved in the classification of legal documents. Large collections of documents do not necessarily lead to better results. Table 2 shows several instances of large documents (categories 4, 9, and 100) producing moderate results and also instances of small documents (categories 5, 6, 11) producing good results.

Label	Precision	Recall	F-1	# Docs

0 - None	0.00	0.00	0.00	5
1 - Criminal Procedure	0.80	0.78	0.79	302
2 - Civil Rights	0.59	0.58	0.59	200
3 - First Amendment	0.69	0.77	0.73	94
4 - Due Process	0.35	0.47	0.40	51
5 - Privacy	0.61	0.52	0.56	21
6 - Attorneys	0.48	0.71	0.57	14
7 - Unions	0.61	0.63	0.62	60
8 - Economic Activity	0.77	0.64	0.70	276
9 - Judicial Power	0.46	0.53	0.49	186
10 - Federalism	0.42	0.43	0.42	72
11 - Interstate Relations	0.50	0.64	0.56	11
12 - Federal Taxation	0.83	0.75	0.79	53
13 - Miscellaneous	0.00	0.00	0.00	2
14 - Private Action	0	0	0	0

Table 2. Classification Results by (manually curated) Category of Legal Opinions

The Doc2Vec model embeddings were also utilized as search indices in the latent embedding space. The vectors derived from Roe v. Wade decision was used to search for additional opinions in the corpus that are similar to it. The results are shown in Figure 2.

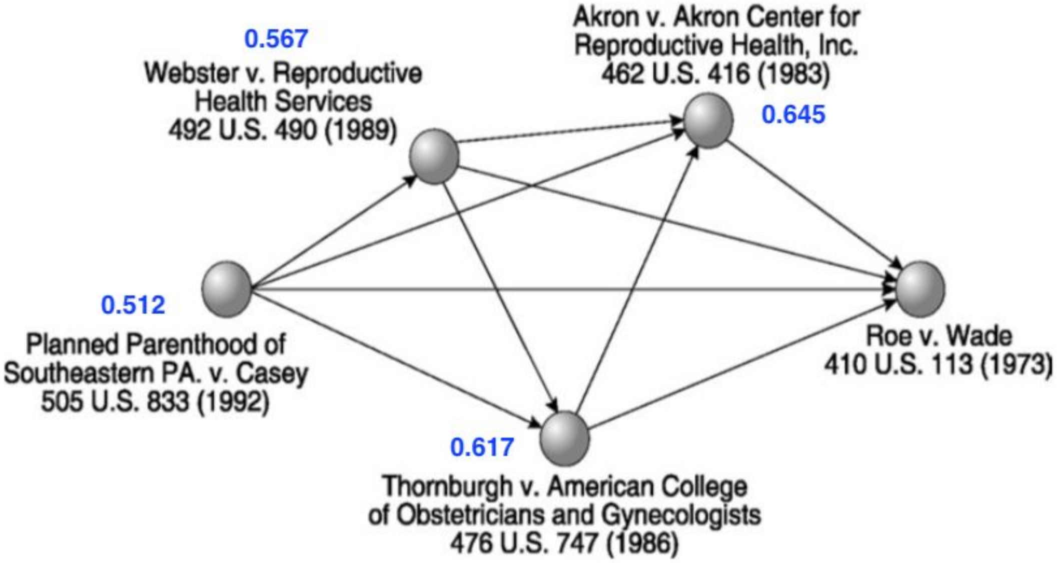
0.567
Webster v. Reproductive Health Services
492 U.S. 490 (1989)
Akron v. Akron Center for Reproductive Health, Inc.
462 U.S. 416 (1983)
0.645
0.512
Planned Parenthood of Southeastern PA. v. Casey
505 U.S. 833 (1992)
Roe v. Wade
410 U.S. 113 (1973)
0.617
Thornburgh v. American College of Obstetricians and Gynecologists
476 U.S. 747 (1986)

Figure 2. Similarity scores between major abortion-related US Supreme Court decisions. Scored calculated based on doc2Vec embeddings of gensim package. (Original image taken from [11]).

Our experiments with Transformer based neural nets ran into memory limitations. Even after upgrading to paid version of Google Collaboratory and utilizing the high-RAM version, the model ran out of memory. The notebook used for BERT based experiments is also listed in Appendix A.

Conclusion and Future Work

The results of this project clearly indicate that NLP techniques hold much promise in the analysis and classification of legal. The research also demonstrated the applicability of paragraph embeddings (Doc2Vec) as a search index for identifying similar opinions in a legal corpus. The limitations (large memory requirements) of neural networks were also demonstrated.

Document classification (especially documents with long sentences) is an area of fertile research [12]. Document classification relies less on syntactic structures than other NLP tasks such as language inference, Q&A answering, paraphrasing, etc. Leveraging the advantages of Transformer architectures for the classification of long documents is a very hot area. Recent advances in this area include:

- Longformer based on modified self-attention [13] ,
- Reformer The Efficient Transformer (replaces dot product attention with locality sensitive hashing) [14]

Both approaches may help alleviate the memory limitations confronted by this project and worth trying out in the future.

References

[1] R. Nallapati and C. D. Manning, “Legal docket-entry classification: Where machine learning stumbles,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing, ser. EMNLP ’08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 438–446. [Online]. Available: http://dl.acm.org/citation.cfm? id=1613715.1613771

[2] R. Bartolini, A. Lenci, S. Montemagni, V. Pirrelli, and C. Soria. Automatic classification and analysis of provisions in italian legal texts: a case study. In Proceedings of the Second International Workshop on Regulatory Ontologies, 2004.

[3] Jackson, Peter & Al-Kofahi, Khalid & Tyrrell, Alex & Vachher, Arun. (2003). Information extraction from case law and retrieval of prior cases. Artificial Intelligence. 150. 239-290. 10.1016/S0004-3702(03)00106-1.

[4] Wan, L., Papageorgiou, G., Seddon, M., & Bernardoni, M. (2019). Long-length Legal Document Classification.

[5] W.-H. Weng, K. B. Wagholikar, A. T. McCray, P. Szolovits, and H. C. Chueh, “Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach,” in BMC Medical Informatics and Decision Making, 2017.

[6] Cavar, Damir et al. “Law Analysis using Deep NLP and Knowledge Graphs.” (2018).

[7] https://github.com/chartbeat-labs/textacy

[8] http://scdb.wustl.edu/documentation.php?var=issueArea#norms
and http://scdb.wustl.edu/documentation.php?var=issue#norms

[9] https://radimrehurek.com/gensim/
[10] https://huggingface.co/transformers/model_doc/bert.html
[11] Fowler, J., Johnson, T., Spriggs, J., Jeon, S., & Wahlbeck, P. (2007). Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court. *Political Analysis, 15*(3), 324-346.
[12] Adhikari, A., Ram, A., Tang, R., & Lin, J. (2019). DocBERT: BERT for Document Classification. https://github.com/castorini/hedwig
[13] https://arxiv.org/abs/2004.05150
[15] https://arxiv.org/abs/2001.04451

Appendix A - List of Google Collaboratory Notebooks

Datasets
Pre-process / exploration of 8K decisions from the US Supreme Court:
https://colab.research.google.com/drive/1Ejxg-aMCPixeOAjLamu7MIPdsnY5DN1R

Models
The following has the LDA + LR model.
cs229_1a_LDA.ipynb
https://colab.research.google.com/drive/1kzLLFmkJRlpJgaMTX4Q5IjhtNnmOeVUy?usp=sharing

Document Embedding vectors used with Logistic Regression.
cs229_2a_DocVec.ipynb
https://colab.research.google.com/drive/1TRRFryj-OmJ1dXSW46Bs1i0mxfiY1I2K?usp=sharing

Search based on doc embedding vectors - gensim search.
cs229_2B_search_DocVec.ipynb
https://colab.research.google.com/drive/1Y67rNfpF8OPkxnQ9_ikT2kgKpCU-Wh9O?usp=sharing

BERT based document analysis. (At present runs out of memory)
cs229_3a_bertDoc.ipynb
https://colab.research.google.com/drive/1ZDr8fi1_SQxKEu1Wtps9uOyBGj_Bf4iF?usp=sharing