The George Washington University

Columbian College of Arts & Sciences

DATS 6312: NaturalLanguageProcessing

Professor AmirJafari,PhD

Team 3: Akhil Bharadwaj, Brunda Mariswamy & Chirag Lakhanpal

Legal Document Summarization and Analysis

Project Proposal

Problem Selection and Rationale

The problem our group has selected is the summarization and analysis of legal documents. The legal industry is inundated with large volumes of complex documents such as contracts, court rulings, and legislation. These documents are often lengthy and difficult for nonexperts to understand. Efficiently summarizing and analyzing these texts can significantly benefit legal professionals and the general public by saving time, reducing costs, and enhancing understanding. We chose this problem because there is a growing demand for tools that can simplify the consumption of legal information, and we believe that Natural Language Processing (NLP) offers promising solutions.

Database/Dataset

For this project, we will use the Caselaw Access Project (CAP), European Court of Human Rights (ECHR) Cases, the EURLex dataset, and consolidated datasets from Hugging Face. This dataset contains full-text case reports and is annotated, which makes it suitable for training and testing summarization, Q&A, and NameEntity extraction tasks.

NLP Methods

We plan to employ a combination of classical NLP methods and deep learning models. Initially, we will start with extractive summarization using classical algorithms like TFIDF to establish a baseline. Progressing from there, we aim to customize a transformer-based model, specifically BERT (Bidirectional Encoder Representations from Transformers), for abstractive summarization.

Packages

The NLP packages we are planning to use include

1. NLTK For tokenization, stopwords removal, and other text preprocessing steps.
2. Unstructured for text preprocessing
3. spaCy For advanced linguistic features like part-of-speech tagging and named entity recognition, which are crucial for understanding legal texts.
4. Transformers by Hugging Face For implementing and customizing BERT models.
5. Scikitlearn For classical algorithms and metrics.

NLP Tasks

The NLP tasks we will work on are

1. Text summarization (both extractive and abstractive)
2. Named entity recognition to identify legal entities
3. Sentiment analysis to gauge the tone of the document
4. Text classification to categorize documents into areas of law
5. Q&A of legal documents (If time permits)
6. Similar Topic (If time permits)

Performance Metrics

We will judge the performance of the model using a variety of metrics, including

1. ROUGE (RecallOriented Understudy for Gisting Evaluation) for summarization quality.
2. F1 score for named entity recognition accuracy.
3. Accuracy and confusion matrix for text classification tasks.

Project Schedule (Tentative)

- Week 1 Data collection, cleaning, preprocessing, and implementation of baseline extractive summarization models.
- Week 2 Development and training of BERT-based abstractive summarization models, and Finetuning models and implementing additional NLP tasks (NER, sentiment analysis, classification).
- Week 3 Evaluation of models using selected metrics.
- Week 4 Finalizing report and preparing a presentation of our findings.