Analysis And Summarization Of Legal Case Reports

Eseoghene Emuraye *Data Science*
Sharmin Kantharia *Data Science*
Hemanth Koganti *Data Science*

Introduction

Legal documents are difficult to parse through for knowledge and application in future cases due to their extensive theory. Hence, it is imperative to generate the main ideas by analyzing them. Manually classifying these legal documents in a large repository is a tremendous task. Therefore, we must find easier ways for summarization, using catchphrases and citations present in them. Through this project, we aim to analyze different legal case reports by grouping them based on common citations mentioned in these cases and summarizing them. This presents insights into the types of cases in each category and their differences.

Scope of Project

The scope of the project covers data preprocessing, clustering of legal case reports based on similar citations, and multi-document summarization of documents in different clusters. Generated summaries are evaluated against reference summaries to determine the quality of generated summaries.

Dataset

The dataset was obtained from the Machine Learning Repository of the University of California Irvine (UCI), California. The corpus contains legal cases from the Federal High Court of Australia (FCA). The corpus contains 3890 full-text decisions from the Federal High Court of Australia (FCA) from 2006 – 2009. The legal case disputes are mostly civil cases, with minor criminal cases. The dataset is presented in an XML format.

Background

Automatic summarization is the technique of reducing a text document with a system to create a summary that retains the most important information of the original document. A review of the literature revealed two methods of text summarization, which are abstractive and extractive. For the summarization

of the legal case reports, we will be adopting the extractive summarization technique. Extractive summarization involves the summaries produced by extracting sentences from a source text. Furthermore, we will be applying multi-document summarization to generate summaries based on different clusters of legal case reports. The summary helps to quickly get familiarized with the information contained in a larger cluster of documents. To cluster the similar documents together, we assume that similar cases will have similar citations quoted, hence, clustering cases of cases will be based on similar citations. We shall implement the Infomap algorithm, a famous network clustering algorithm, to cluster the cases based on citations.

Methodology

A. ***Text Preprocessing:*** Each legal case report is an XML file that consists of 2 kinds of data – Catchphrases and Sentences. BeautifulSoup library was used to extract the catchphrases and sentences and stored in dictionaries and lists. Cleaned and tokenized catchphrases and sentences were saved as pickle files.

B. ***Clustering:*** The citations document is used to extract the citations in each case report and stored in a directed graph data structure. Case report clusters are generated based on similar citations using the Infomap algorithm. The resulting clusters are stored in pickle files.

C. ***Summarization:*** The pickle files storing the clusters and the reference summaries (catchphrases of respective case reports are taken as input for this step. Cases in each cluster were merged and summaries were generated for the entire cluster-based using the TextRank algorithm from the Gensim package. The summaries are stored for evaluation.

D. ***Evaluation:*** Recall-Oriented Understudy for Gisting Evaluation or ROUGE-N Score is used to evaluate the summarization of texts. The Recall and Precision with respect to the ROUGE Score are calculated as follows:

$$\text{Recall} = \frac{Number\ of\ overlapping\ words}{Total\ words\ in\ reference\ summary} \qquad \text{Precision} = \frac{Number\ of\ overlapping\ words}{Total\ words\ in\ generated\ summary}$$

Results

- 281 clusters were generated with cluster 1 having 618 cases while several clusters had just a case. Clusters with more than 10 cases per cluster were selected for case summarization. 39 clusters representing 51.56% (2006) of all cases were fed to the TextRank algorithm for summarization. Summaries were generated for all 39 clusters except for cluster 37.
- The recall in the context of ROUGE refers to how much of the reference summary is captured by the generated summary, while precision refers to how much of the generated summary is relevant. The recall and precision for monograms (ROUGE-1) and bigrams (ROUGE-2) were plotted for all clusters. The same recall and precision trends are observed when the summaries are evaluated as monograms and bigrams.
- For ROUGE-1, the maximum recall score was recorded as 0.8813 for cluster 1, cluster 9 coming closely behind with a score of 0.8756. The minimum recall score was 0 at cluster 37 since no summary was generated, cluster 31 comes close at 0.5217. Generally, the summaries had low precision scores with a maximum at cluster 46 (0.3525) and a minimum at cluster 4 (0.0149).

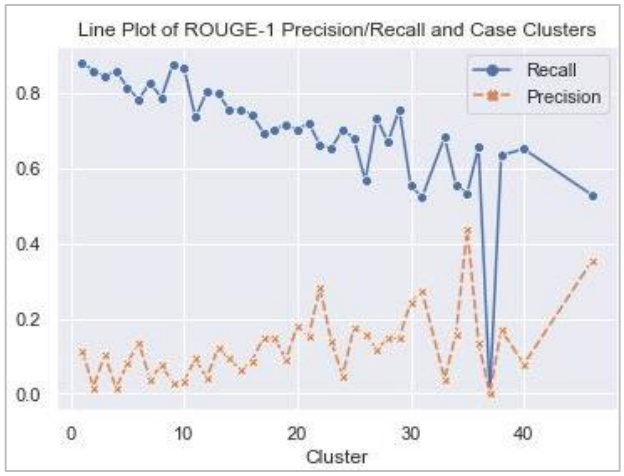

Figure 2: Recall and Precision for ROGUE-1

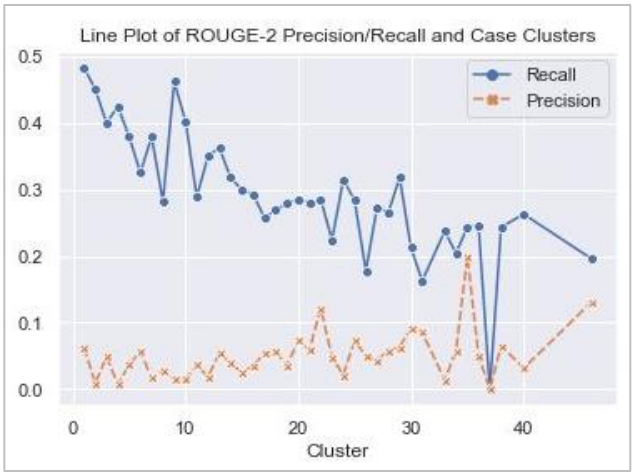

Figure 1: Recall and Precision for ROGUE-1

Challenges

1. The documents present in some clusters were large and thereby increased the computation time.
2. Converting clustered cases to the appropriate data type to fit in the summarizer in the gensim package was a hindrance.
3. It was difficult to ascertain the reason why cluster 37, despite having 16 cases, did not generate any summary when fed into the summarizer.

Conclusion

A great deal of work has been done to generate the summaries for each cluster. Judging by an overall recall score greater than 0.5, it can be understood that the extractive summarization techniques, give a good representation of cases based on similar citations. Through the technologies implemented in this project, it can be understood that given citations, we can generate clusters and analyze the texts in each cluster. Hence, similar concepts can be applied to legal cases in any constitution.

Future Work

Legal texts have characteristics different from news articles and other texts, specifically in the vocabulary and ambiguity.

- Given more time, it would be interesting to identify the topics labels present in different clusters of legal case reports by applying the Latent Dirichlet Allocation (LDA) algorithm to generate different topic labels per cluster allowing comparison across clusters to check for similarities or differences.
- There is also a possibility to explore and apply abstractive summarization techniques to generate summaries for the same dataset and compare the results of the 2 summarizations methods.
- Legal jargon can also be extracted for better understanding the language of a case, using keywords generated in through the TextRank algorithm.