Project Proposal

Summarization and Citation Class Classification in Legal Case Reports

Group Number: 7

Group Member Names:
Nithish Kumar Senthil Kumar
Rishikesh Ramesh
Nagul Pandian Chinnasamy Ramkumar
Kunal Ahirrao

Objective:

This project focuses on addressing two tasks using legal case reports:

1. **Automatic Summarization of Legal Cases**: We aim to develop a model that can summarize complex legal cases, capturing key legal principles and decisions. This task will aid legal professionals by providing concise, meaningful summaries, helping them understand long and complicated cases quickly.

2. **Citation Class Classification**: The second task is to classify citation relationships between legal cases (e.g., "applied," "cited," "followed"). This will help in understanding the type of legal precedent a case sets or follows in relation to other cases.

The overall objective is to streamline the analysis of legal documents, making it easier to access critical information. Both tasks will be executed using part of the dataset, which is feasible within the computational limits of Google Colab. The models will be fine-tuned versions of pre-existing models, ensuring that even partial datasets will yield reliable results.

Data Set Description:

This dataset contains legal case reports from the **Federal Court of Australia (FCA)** between 2006 and 2009. The cases were downloaded from AustLII, a free access platform to Australian legal information. The dataset was created for experiments in summarization and citation classification. Each document includes multiple elements such as full text, catchphrases (summarizing key legal points), and detailed citation information.

Dataset Link: https://archive.ics.uci.edu/dataset/239/legal+case+reports

- **Dataset Components**:

    1. **Fulltext**: Includes full legal documents along with a list of catchphrases for each case.
    2. **Citations Summ**: Contains citation elements including catchphrases and sentences cited from or by other cases, as well as references to legislation.
    3. **Citations Class**: Lists citations to older cases with labels (e.g., applied, distinguished, cited) that indicate the relationship between cases.

Interesting Data Characteristics:

1. **Non-extractive Catchphrases**: The catchphrases provided in the dataset are not directly extractive from the text, meaning that they represent abstract summaries rather than verbatim excerpts.

2. **Wide Range of Citation Classes**: Citations in this dataset include various legal relationships, such as "applied," "cited," and "distinguished," offering rich classification opportunities.

3. **Variable Document Length**: The word count for documents varies widely, with some documents having very large chunks of text that may need to be trimmed for analysis.

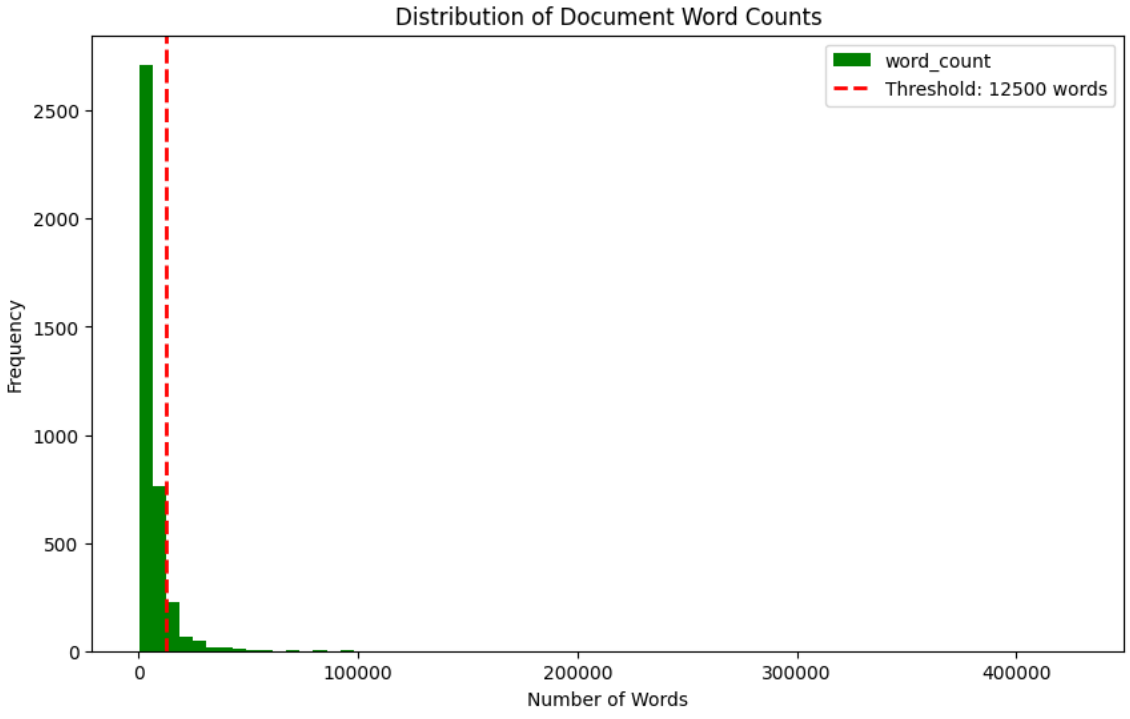
Preliminary Data Exploration:

Summary Statistics:

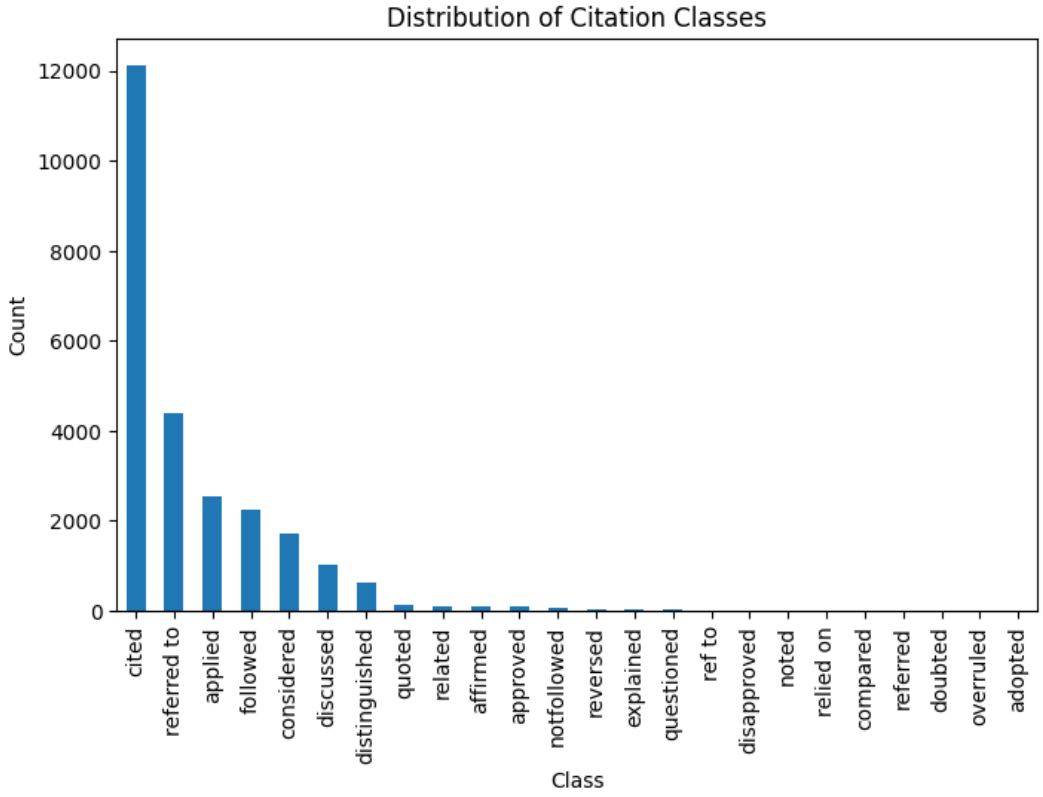
- The dataset contains 3,890 legal cases with more than 1.8 million catchphrases in total.
- There are 25,256 citation records across the dataset, each classified under different legal relationships.

Visualizations:

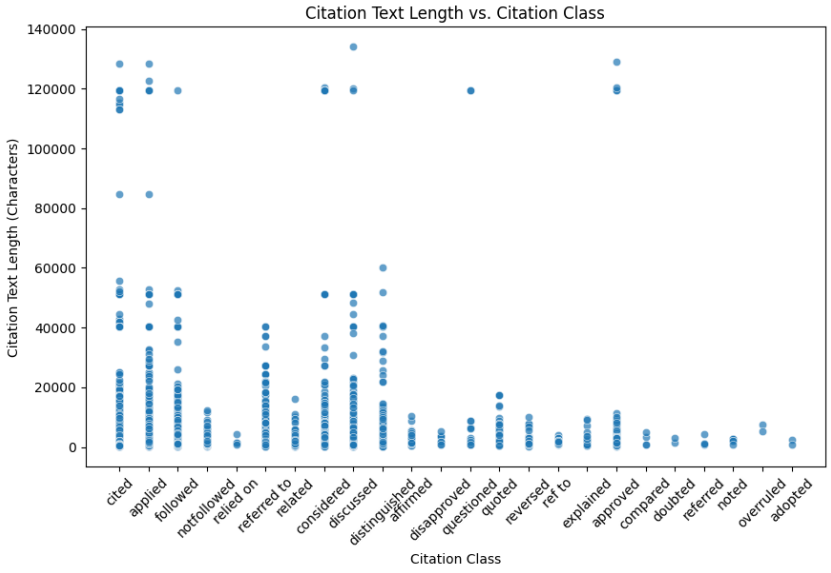
- **Document Word Count Distribution**: Most documents have a word count under 12,500, with a small fraction containing much larger text chunks.



- **Citation Class Distribution**: The distribution of citation classes is dominated by common citation relationships like "cited" and "referred to," while other categories like "approved" or "reversed" are less frequent.



- **Citation Text Length vs. Class**: Citations like "cited" tend to have shorter text lengths, while more intricate relationships such as "explained" or "reversed" feature longer citation texts.



Proposed Data Exploration:

Instead of focusing on extensive data exploration, we will analyze the dataset in ways directly related to our tasks:

1. **Summarization of Legal Cases**: We'll explore sentence-to-catchphrase relationships, identifying patterns between the input text and catchphrases.

2. **Citation Class Analysis**: We'll investigate the characteristics of citation text for different classes, looking at features like text length, complexity, and context.

Proposed Predictions:

For each of the two tasks, we will consider the following approaches:

1. **Summarization Task**:
   We may experiment with transformer-based models such as **BART** or **T5** for generating abstractive summaries of the legal documents. The summaries will be compared with the provided catchphrases to assess the model's performance.

2. **Citation Class Classification**:
   We will explore fine-tuning **LegalBERT** or **BERT** to classify citation relationships between cases. The citation text will be the input, while the citation class (e.g., "applied," "cited," "distinguished") will be the target output. Given the complexity of legal language, fine-tuning pretrained models on this specific dataset is a promising approach.

Note: We will be utilizing only a **subset** of the dataset due to computational constraints on Google Colab, but this will not impact the overall results because the models will be fine-tuned on top of existing pre-trained models.