# Summarizing Privacy Policy with NLP

## Introduction

The primary goal of this project is to develop a robust algorithm capable of summarizing privacy policy texts, such as terms of service agreements, into concise, human-readable summaries. Given the length and complexity of these documents, users often forgo reading them, leading to a lack of informed consent. This project addresses this challenge by leveraging natural language processing (NLP) techniques to produce summaries that retain essential information while simplifying the content.

This notebook documents the end-to-end process, from data collection and cleaning to modeling and evaluation, emphasizing the importance of systematic preprocessing and thoughtful experimentation. Effective early-stage organization and exploration of data lay the groundwork for accurate, efficient summarization models.

## Objectives

The objective of this project is to develop an NLP-based summarization model for privacy policy texts. The model aims to generate summaries that achieve a high level of abstraction, compression, and semantic accuracy. Through experimentation with various approaches and evaluation metrics, the project seeks to identify methods that balance computational efficiency with linguistic precision and readability.

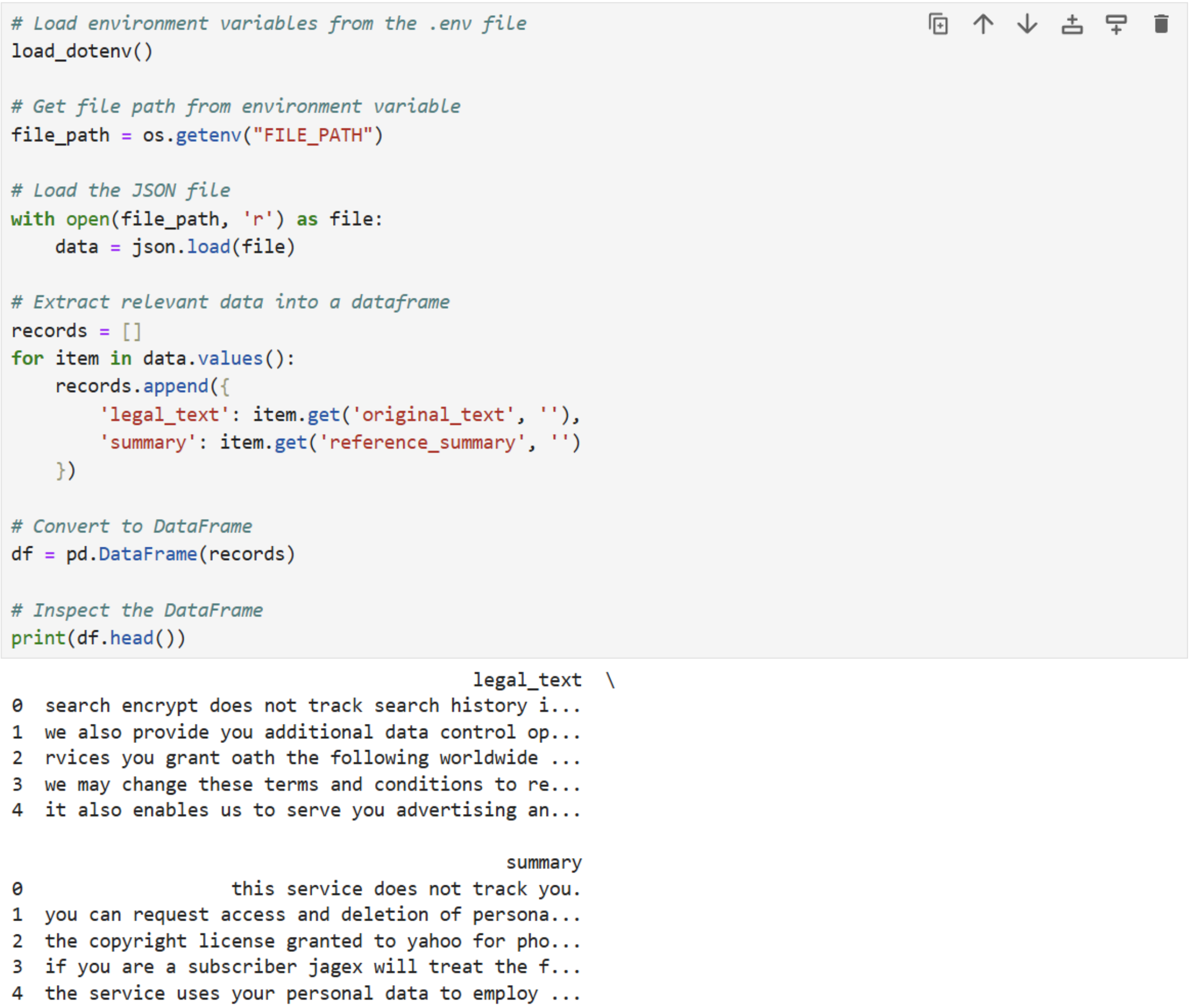
## About the dataset

This project utilizes a dataset of **421 pairs** of legal text snippets and their corresponding plain English summaries, sourced from: https://aclanthology.org/W19-2201/ Abstract Unilateral legal contracts, such as terms of service, play a substantial role in modern digital life. However, few read these documents before accepting the terms within, as they are too long and the language too complicated. We propose the task of summarizing such legal documents in plain English, which would enable users to have a better understanding of the terms they are accepting. We propose an initial dataset of legal text snippets paired with summaries written in plain English. We verify the quality of these summaries manually, and show that they involve heavy abstraction, compression, and simplification. Initial experiments show that unsupervised extractive summarization methods do not perform well on this task due to the level of abstraction and style differences. We conclude with a call for resource and technique development for simplification and style transfer for legal language.

# Methodology

## Data Preprocessing

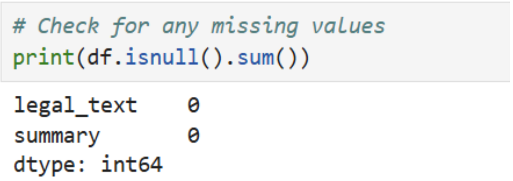
1. **DataFrame Creation:** The data is organized into a structured format to facilitate analysis and manipulation.



```python
# Load environment variables from the .env file
load_dotenv()

# Get file path from environment variable
file_path = os.getenv("FILE_PATH")

# Load the JSON file
with open(file_path, 'r') as file:
    data = json.load(file)

# Extract relevant data into a dataframe
records = []
for item in data.values():
    records.append({
        'legal_text': item.get('original_text', ''),
        'summary': item.get('reference_summary', '')
    })

# Convert to DataFrame
df = pd.DataFrame(records)

# Inspect the DataFrame
print(df.head())
```

```
                                          legal_text  \
0  search encrypt does not track search history i...
1  we also provide you additional data control op...
2  rvices you grant oath the following worldwide ...
3  we may change these terms and conditions to re...
4  it also enables us to serve you advertising an...

                                             summary
0                    this service does not track you.
1  you can request access and deletion of persona...
2  the copyright license granted to yahoo for pho...
3  if you are a subscriber jagex will treat the f...
4  the service uses your personal data to employ ...
```

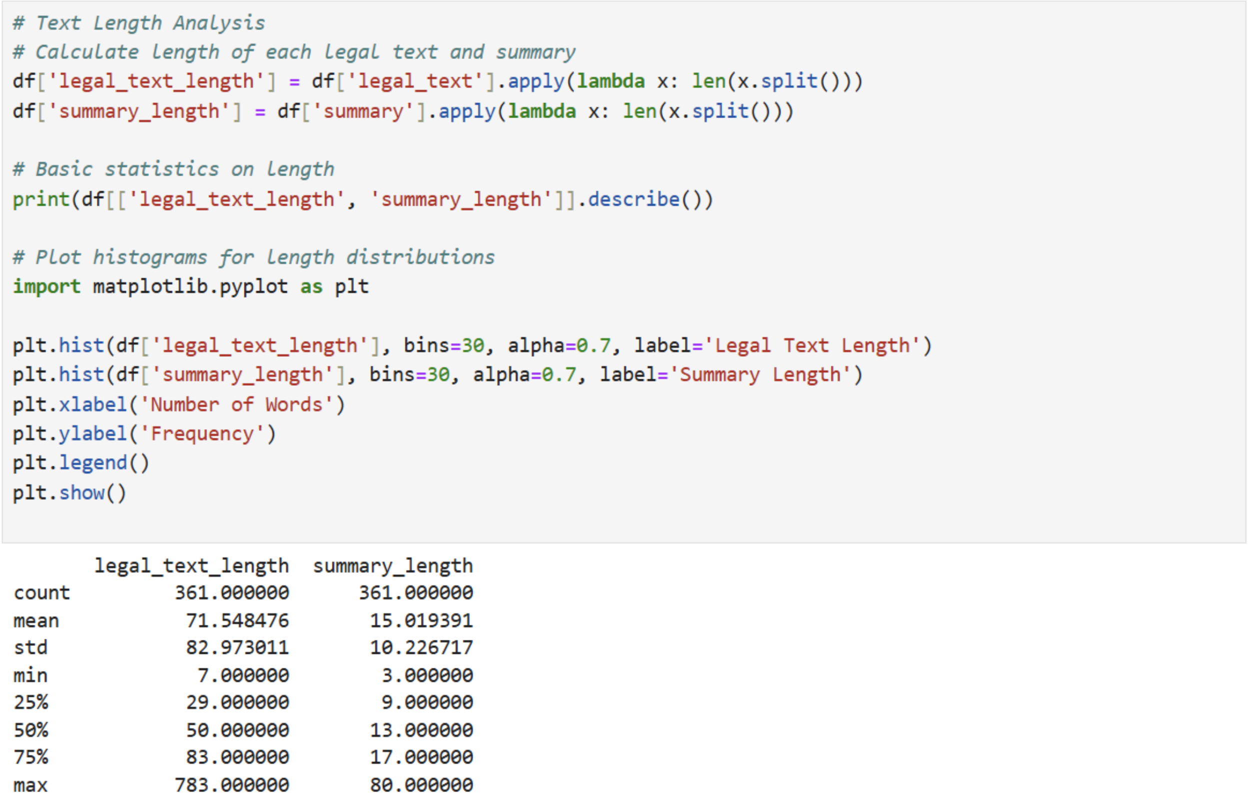
2. **Inspection and Cleaning:** Irregularities in text, such as incomplete sentences, extra spaces, or non-informative tokens, are addressed during cleaning. However, over-cleaning is avoided until a thorough exploratory analysis is conducted.



```
# Check for any missing values
print(df.isnull().sum())
```

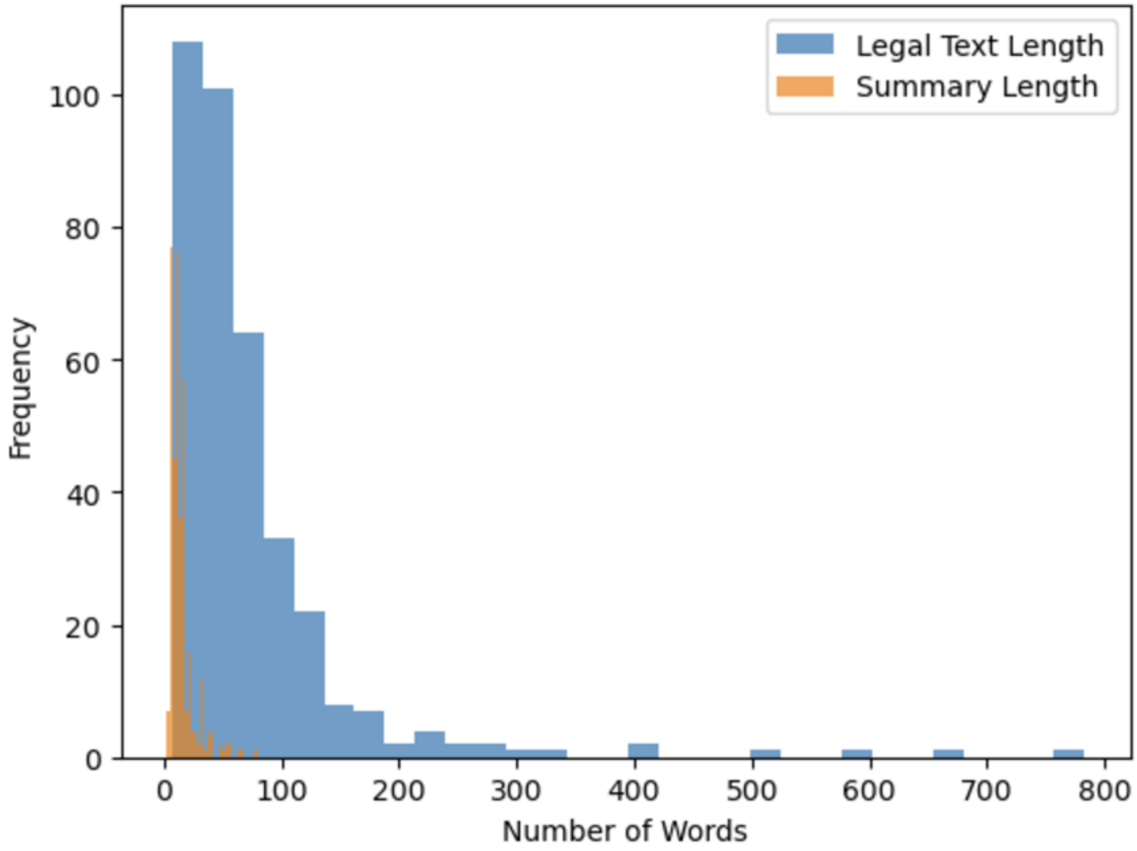
```
legal_text    0
summary       0
dtype: int64
```

3. **Exploratory Data Analysis (EDA):** This step includes analyzing summary length ratios, identifying common patterns, and ensuring data consistency, with the goal of tailoring preprocessing to the unique challenges of legal text summarization.

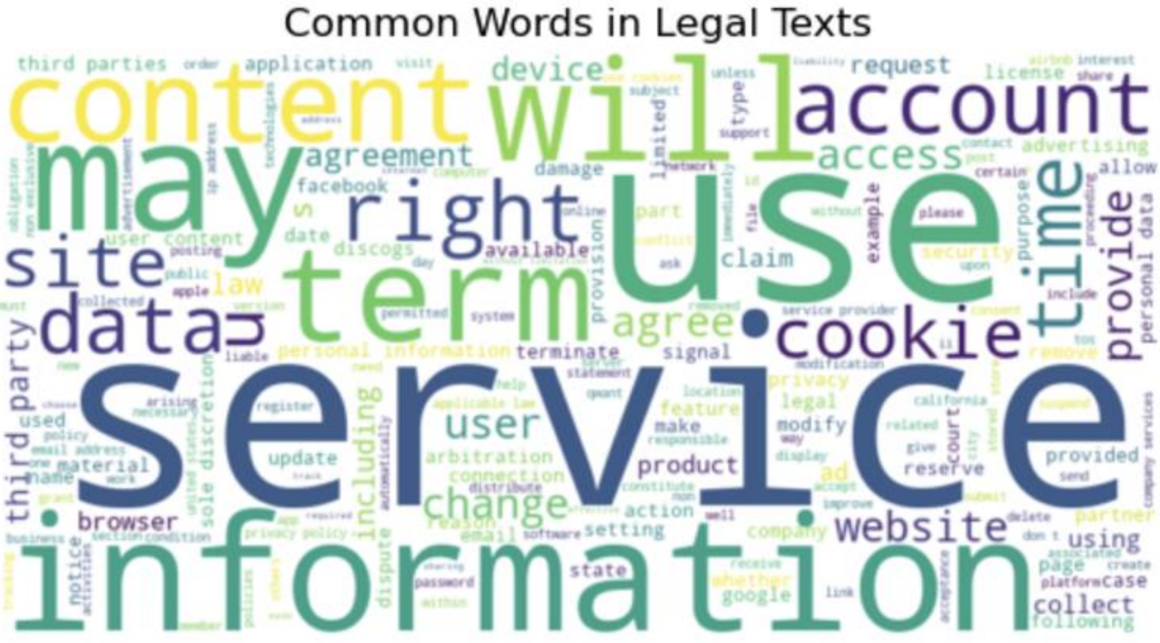


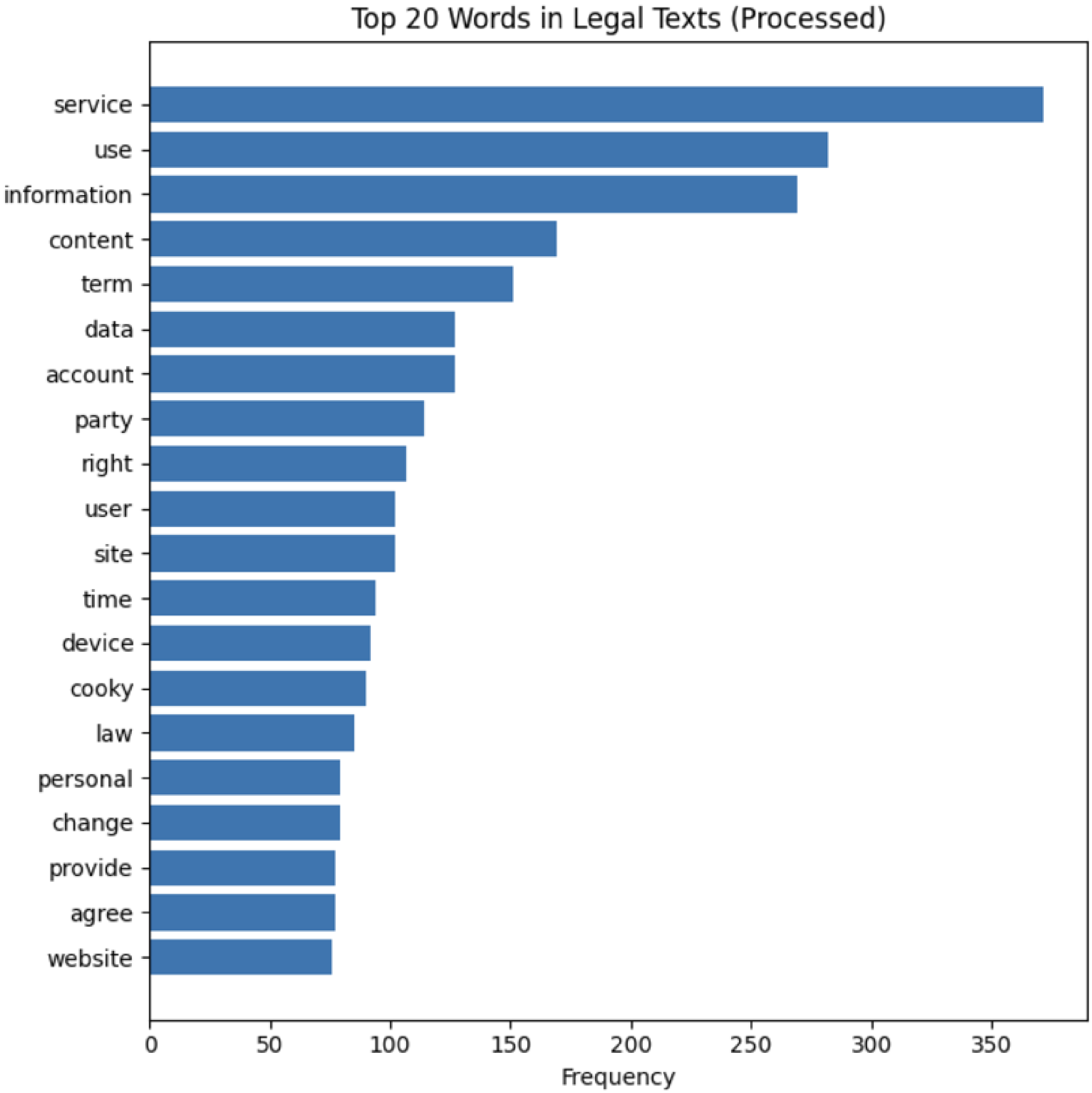
```
# Text Length Analysis
# Calculate length of each legal text and summary
df['legal_text_length'] = df['legal_text'].apply(lambda x: len(x.split()))
df['summary_length'] = df['summary'].apply(lambda x: len(x.split()))

# Basic statistics on length
print(df[['legal_text_length', 'summary_length']].describe())

# Plot histograms for length distributions
import matplotlib.pyplot as plt

plt.hist(df['legal_text_length'], bins=30, alpha=0.7, label='Legal Text Length')
plt.hist(df['summary_length'], bins=30, alpha=0.7, label='Summary Length')
plt.xlabel('Number of Words')
plt.ylabel('Frequency')
plt.legend()
plt.show()
```

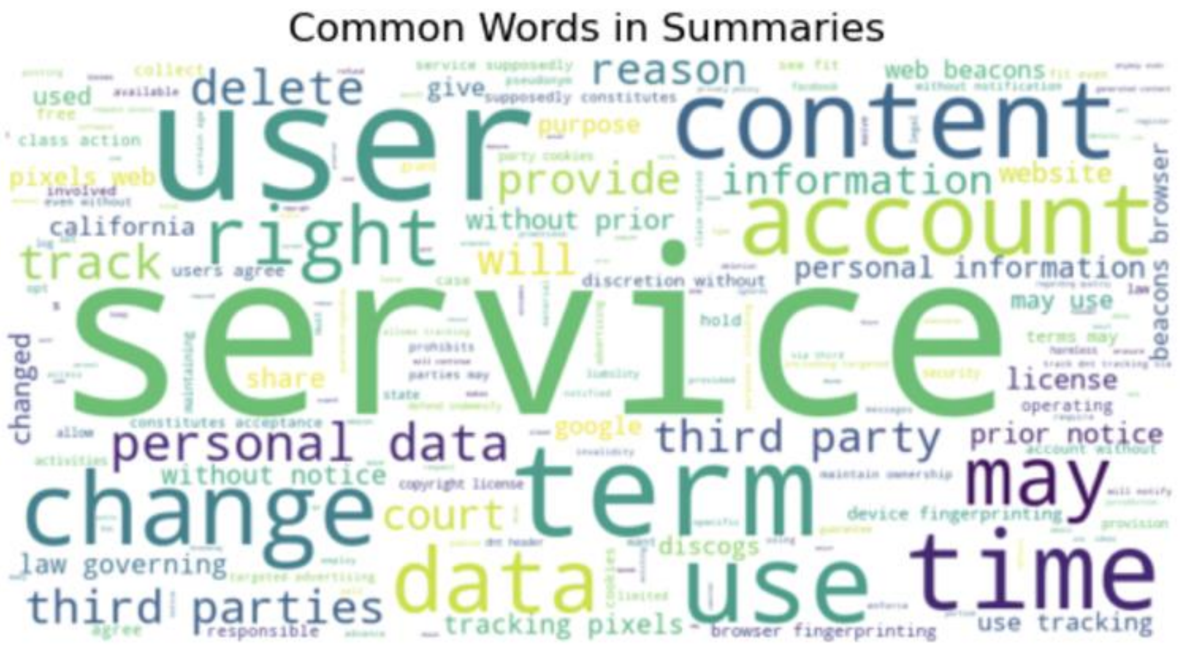
```
       legal_text_length  summary_length
count         361.000000      361.000000
mean           71.548476       15.019391
std            82.973011       10.226717
min             7.000000        3.000000
25%            29.000000        9.000000
50%            50.000000       13.000000
75%            83.000000       17.000000
max           783.000000       80.000000
```



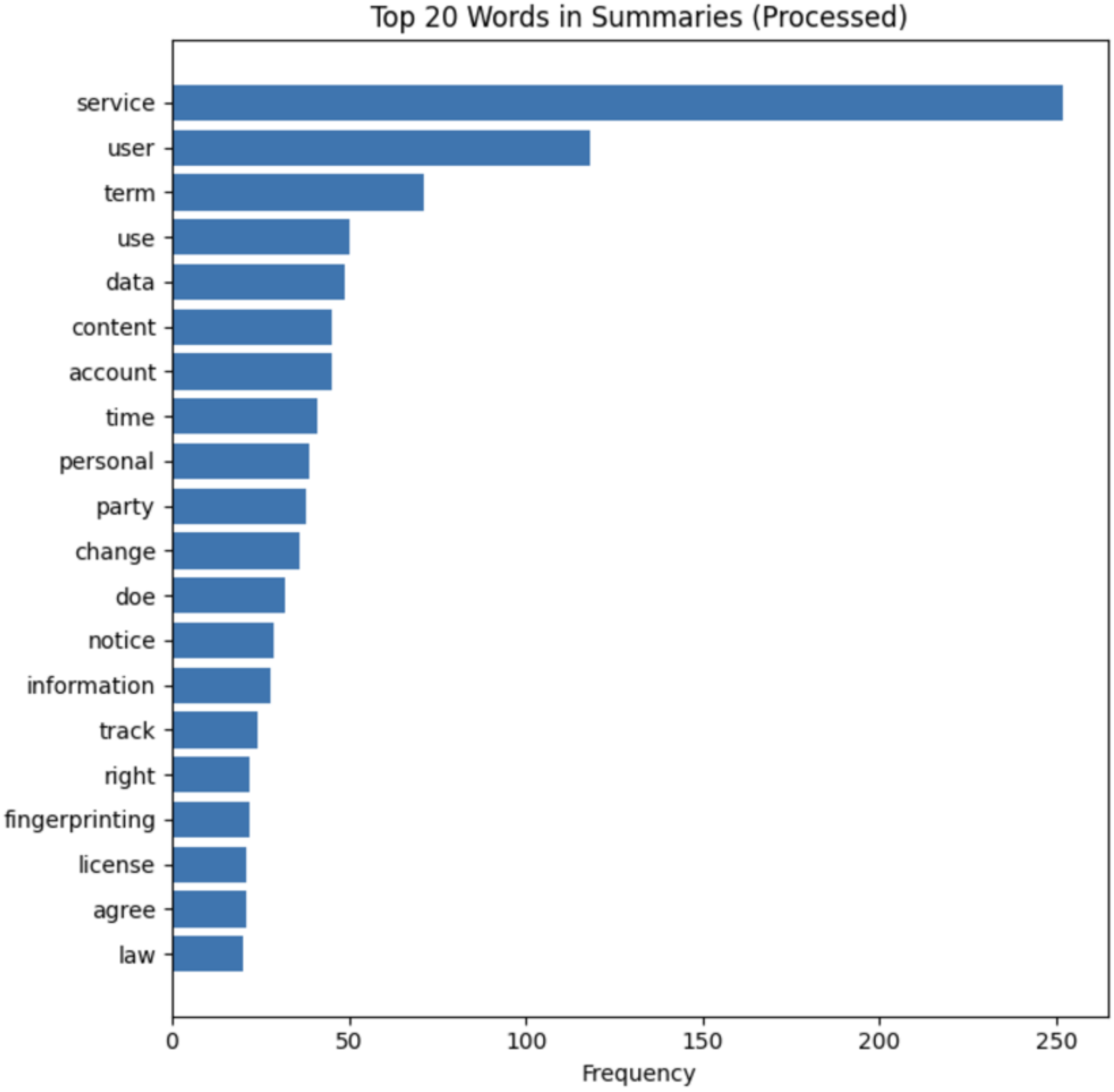
## Common Words


Common Words in Legal Texts

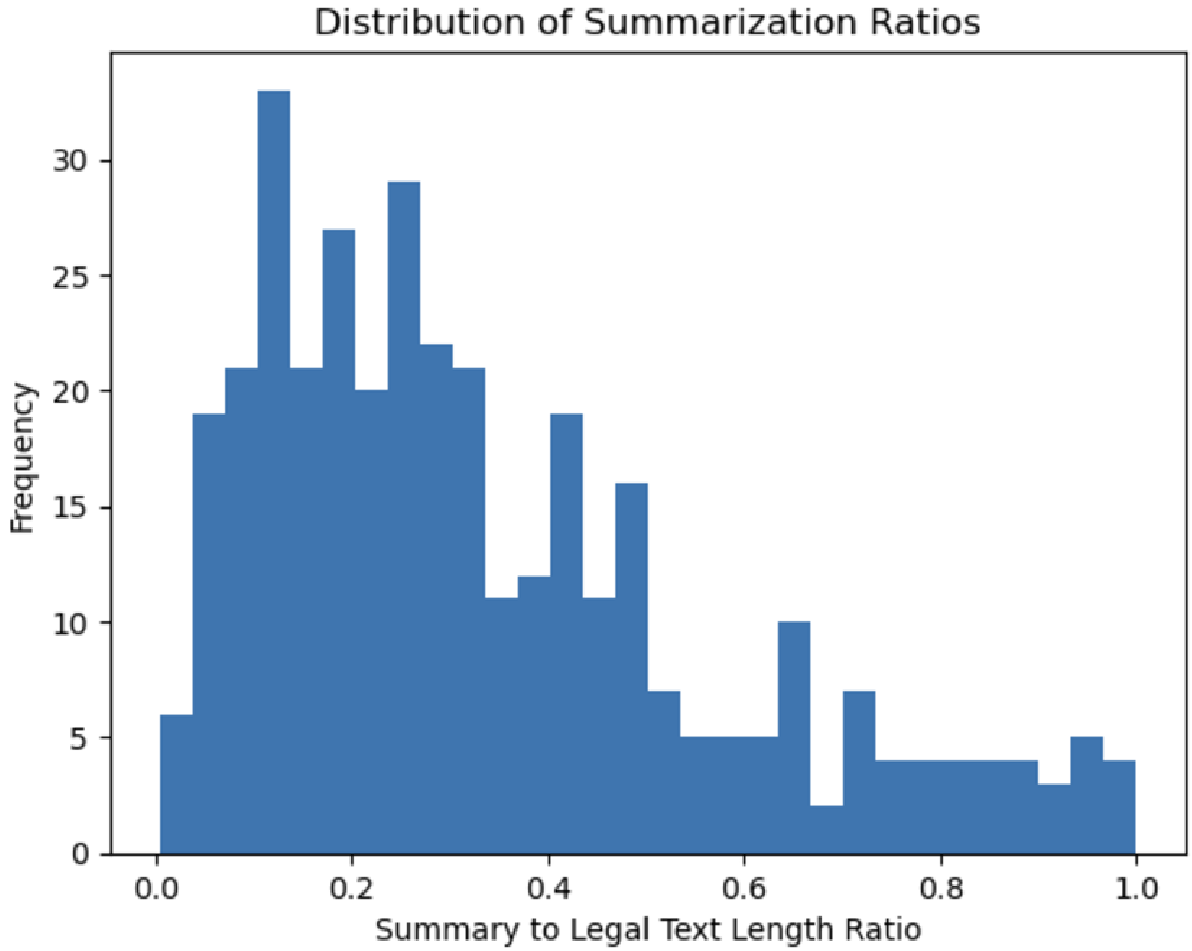

Top 20 Words in Legal Texts (Processed)
service
use
information
content
term
data
account
party
right
user
site
time
device
cooky
law
personal
change
provide
agree
website
0
50
100
150
200
250
300
350
Frequency



Common Words in Summaries



Top 20 Words in Summaries (Processed)



### Feature Engineering

To enhance model performance, additional features or identifiers may be derived from the dataset, depending on insights gained during EDA.

### Data Splitting

The dataset is split into training, validation, and testing subsets to ensure a robust evaluation of model performance.

## Modeling Approach

The modeling process leverages Hugging Face's **facebook/bart-large-cnn** transformer, a pre-trained sequence-to-sequence architecture well-suited for abstractive summarization. This model is fine-tuned on the privacy policy dataset, with preprocessing tailored to the needs of legal text summarization.

### Evaluation Metrics

To assess the model's performance, a suite of standard and advanced metrics is employed:

1. **ROUGE:** Measures n-gram overlap between generated summaries and reference summaries. While effective for extractive tasks, ROUGE may penalize summaries with valid paraphrasing or lexical variation.
2. **BLEU:** Focuses on precision by evaluating how many words in the generated summary match the reference summary, often favoring exact matches.
3. **BERTScore:** Leverages contextual embeddings from transformer models to evaluate semantic similarity, offering robustness to paraphrasing, synonyms, and syntactic variations. By comparing cosine similarity between token embeddings in generated and reference summaries, BERTScore provides a nuanced understanding of meaning beyond lexical overlap.

## Observations and Insights

During EDA, a consistent observation was that the summaries were significantly shorter than their corresponding legal texts, often reducing the word count by a substantial margin. The goal is for the summarization model to replicate this level of reduction while maintaining semantic fidelity and clarity.

Preliminary results suggest that ROUGE and BLEU are limited in their ability to assess abstraction, as they primarily rely on surface-level word matching. In contrast, BERTScore captures the deeper semantic relationships between words, making it a preferred metric for evaluating abstractive models.

## Conclusion

This project demonstrates the potential of transformer-based architectures for summarizing complex legal texts. The Hugging Face BART model performed effectively, particularly when evaluated with metrics like BERTScore, which account for semantic similarity and contextual awareness. This underscores the importance of advanced metrics in capturing the nuances of abstractive summarization tasks, where meaning and readability often outweigh exact lexical matches.

Future work will explore additional model architectures, such as T5 and GPT-based summarizers, and extend the evaluation framework to include user studies for assessing readability and perceived usefulness of the generated summaries.

# References

The project was based on these articles and datasets.

https://github.com/maastrichtlawtech/awesome-legal-nlp

https://aclanthology.org/W19-2201/

https://github.com/lauramanor/legal_summarization/tree/master