KWABENA MENSAH

ASSIGNMENT_2_AAI_520_NLP_Semantic and Sentiment Analysis

Legal Sentiment Analysis

Content
The dataset consists of legal sentences for your analysis.

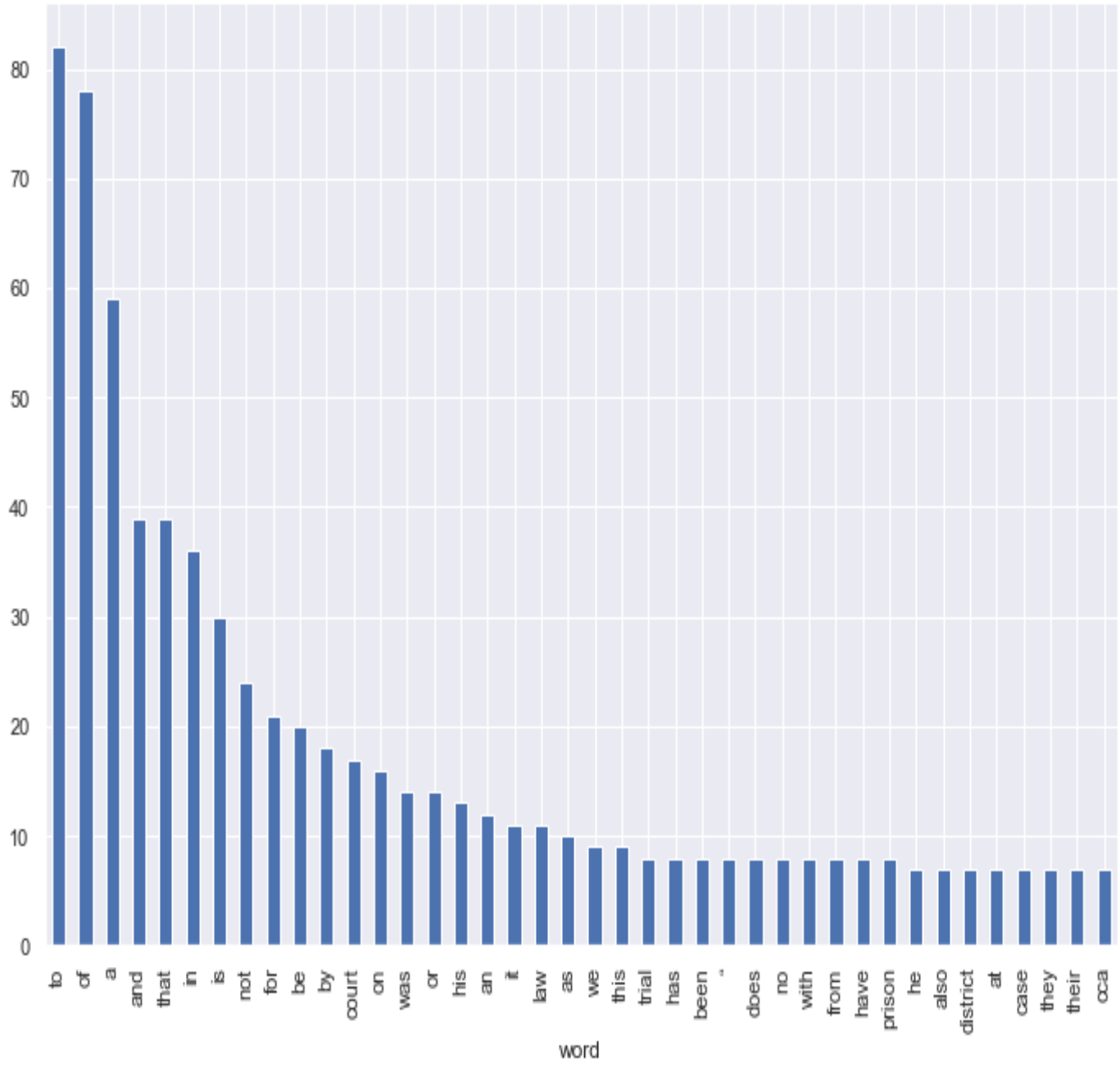
Text preprocessing
Text preprocessing is a crucial step in Natural Language Processing (NLP) and Machine Learning tasks, ensuring that the input data is cleaned and standardized to improve the efficiency and accuracy of subsequent processes. This report summarizes the preprocessing approach employed:

HTML Tag Removal:
All HTML tags present in the text are removed to ensure the extraction of pure content without any formatting instructions.
Contraction Expansion:
Contractions are expanded to their full form. For instance, "I'm" becomes "I am".
Number Removal:
All numeric values present in the text are removed to focus solely on textual content.
URL Removal:
Any URLs present in the text are removed, ensuring that external references don't influence the analysis.
Mentions Removal:
Mentions, often found in tweets, and represented by '@', are removed to keep the focus on the content and not on specific users.
Tokenization:
The text is broken down into individual words or tokens. This aids in analyzing and processing the text at a granular level.
Stopword Removal:
Common words that do not contribute significant meaning to the text, known as stopwords, are removed. Examples include "and", "the", and "is".
Punctuation Removal:
All punctuation marks are removed from the text to ensure a smooth flow and reduce noise.
Non-ASCII Character Removal:
Characters that are not part of the standard ASCII set are removed to maintain uniformity and compatibility.
Hashtag Removal:
Hashtags, which are often used in social media posts, are removed to avoid potential biases in analysis.
Lemmatization:
Words are reduced to their base or dictionary form. For example, "running" becomes "run". This helps in standardizing words and reducing the overall vocabulary.

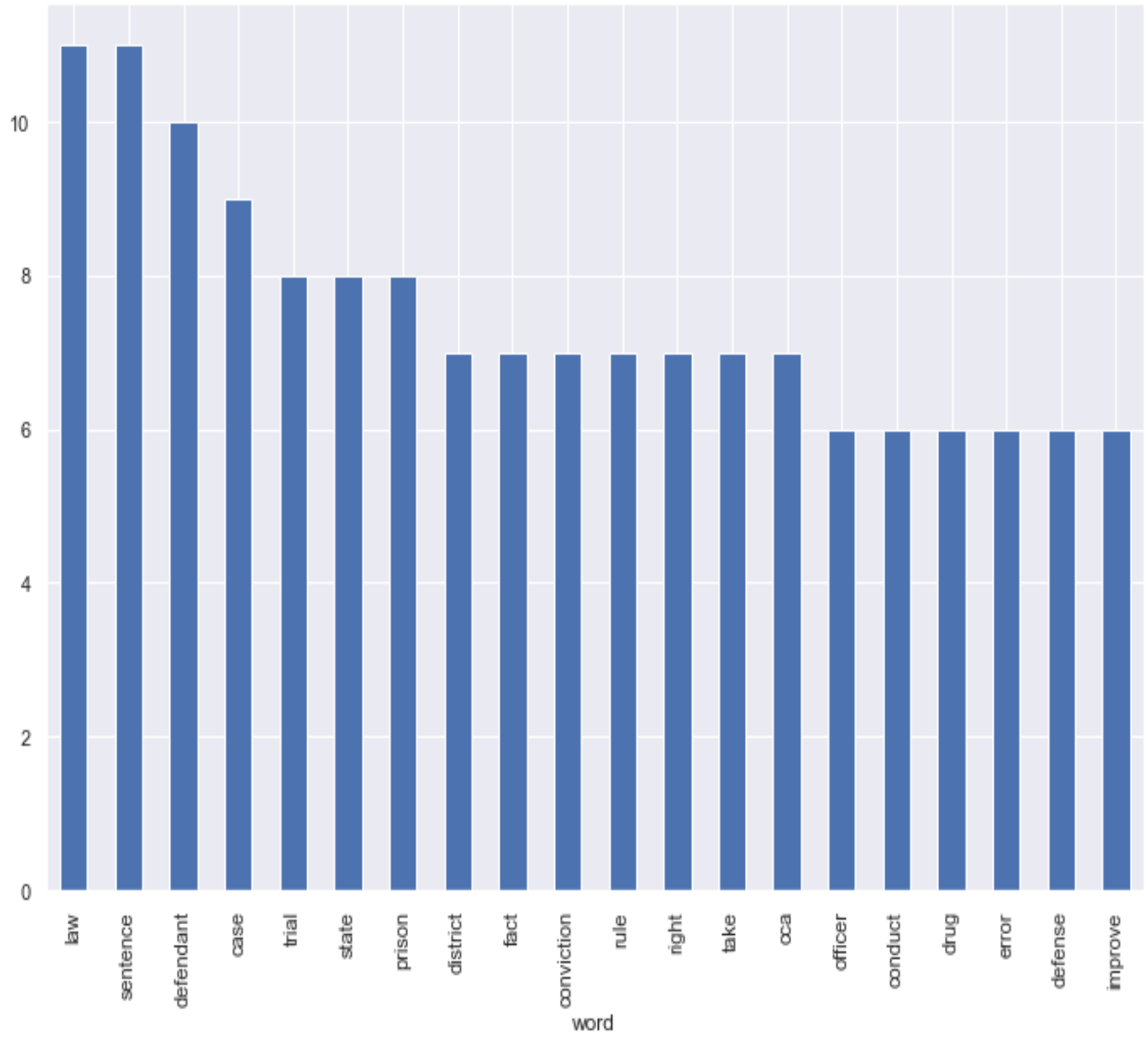
Challenges:
The main challenge was that the data set was small, just 576 rows.

Using the Positive Labels, to illustrate the importance of text preprocessing, we can see how different words stand out after text preprocessing.

Visualizing the DataFrame Before Text Prepocessing – Positive



Visualizing the DataFrame Before Text Prepocessing - Positive



Observations/Findings:

We explored two models, Random Forest and LSTM

Random Forest

An F1-score of 0.3333 suggests that the balance between precision and recall for the model's predictions is not optimal. It indicates that the model may be making a significant number of false positives and/or false negatives.

LSTM

An F1-score of 0.3430 suggests that the balance between precision and recall for the model's predictions is not optimal. It indicates that the model may be making a significant number of false positives and/or false negatives. The score is lower than the accuracy you reported, which means that while the model might be getting 50% of the predictions right, the predictions it gets wrong have a significant impact on precision and recall.

Please See Notebooks for expanded walkthrough of code and steps