Ricardo Diaz Individual Project
NLP Fall 2022
December 11, 2022

Introduction

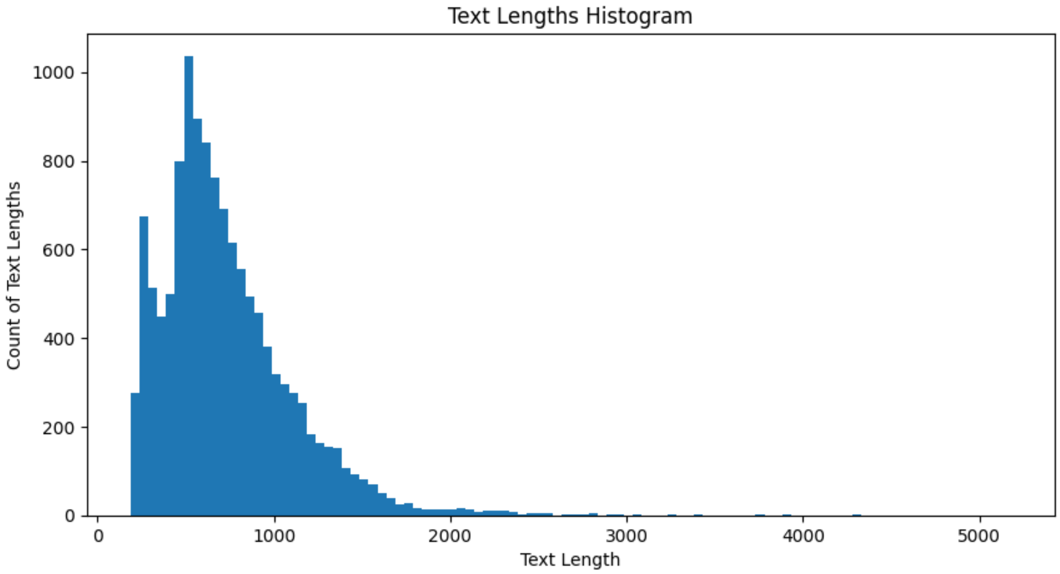
The purpose of the project was to identify the relationship between summarized and text data against the model, furthermore evaluating the model performance with the word length of the summarized and text data. To achieve this evaluation we used a variety of rogue metrics and created a text ratio metric that measures the comparison between summary and text length. The division of the work was:

- EDA: Ricardo Diaz
- Models: Alexis Kaldany
- Results: Sagar Tripathi

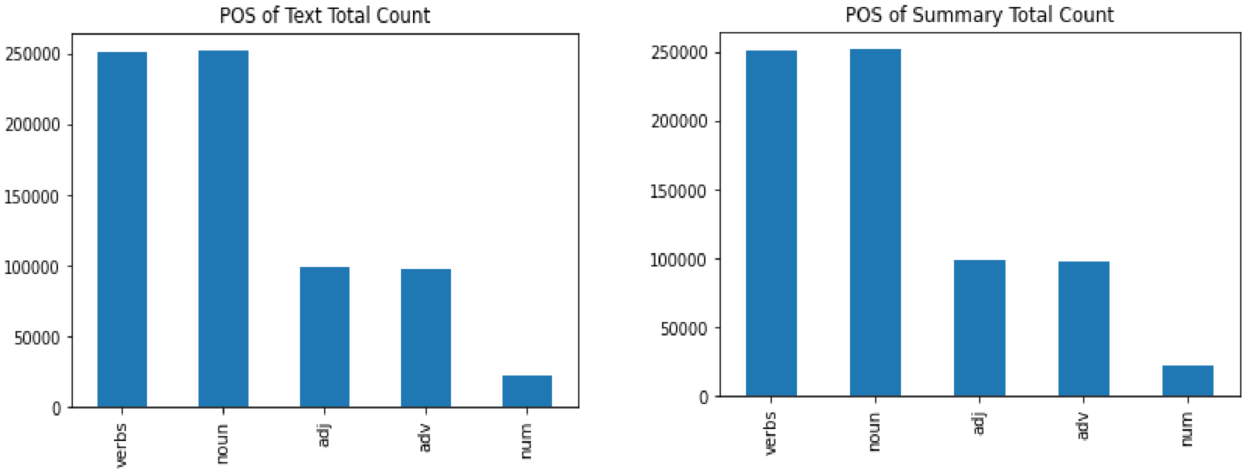
Description of your individual work.

My individual work consisted is grabbing the initial insights of the dataset, using text statistics learned from class, part of speech provided by the spacy library and showing a variety of graphs using matplotlib and seaborn.

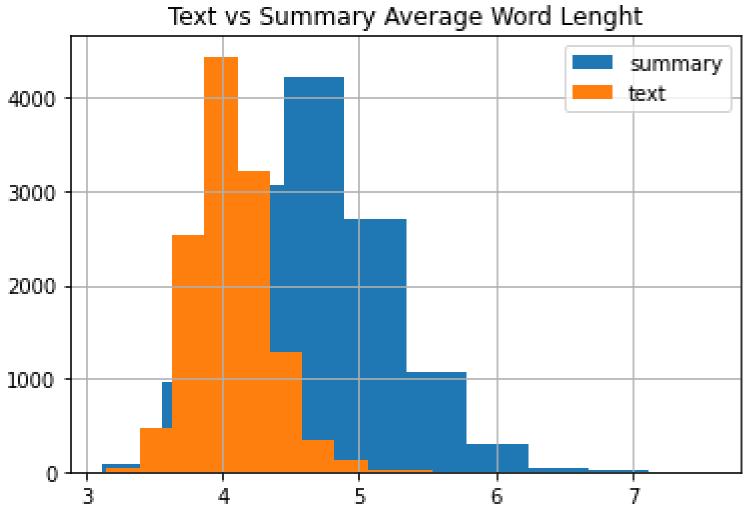
Results



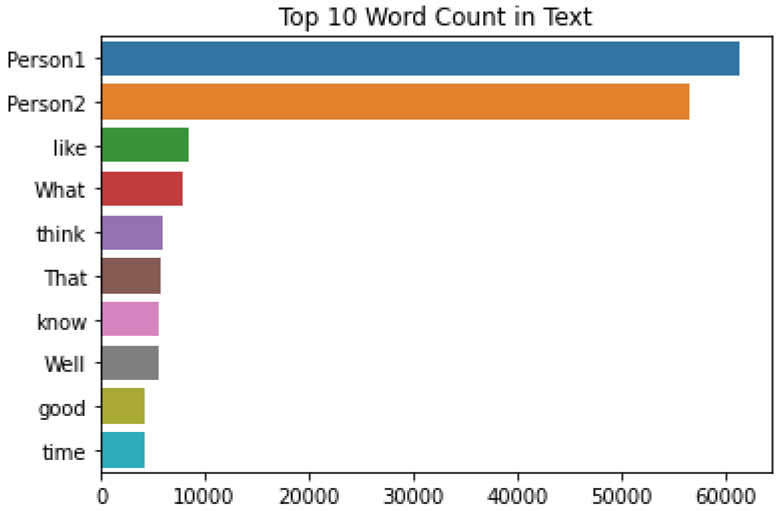
Showing the histogram of the length showed us the distribution of the text length, the text is not large and most of the text has less that 1000 which relatively small.



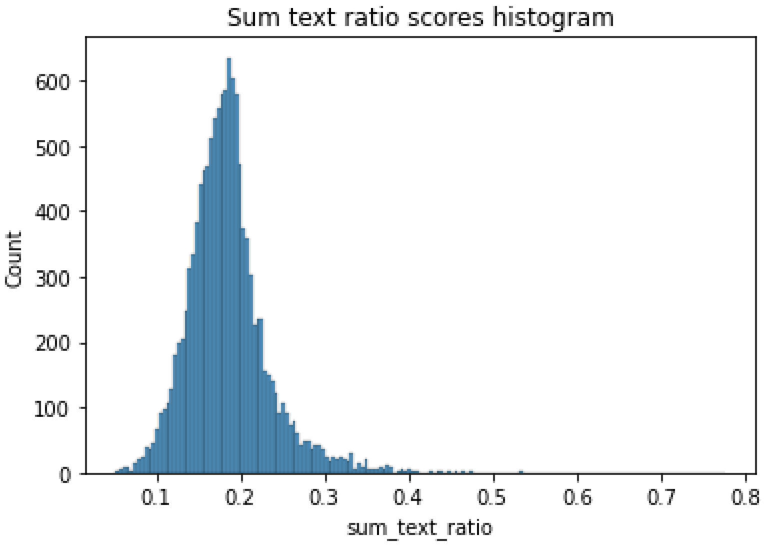
I wanted to compare the part of speech tagging between the summarize and the raw text and I noticed the distribution of POS was similar.



Checking the average word length between summarize and text, one could notice the summary version has longer average word length.



To obtain the most common words in our text, I used ntlk and I found that Person 1 – Person 2 were the most common words, this was sense since is a dialogue, but we realized is only a dialogue between two persons a insights the dataset didn't describe.



We created a feature call sum_text_ratio which compares the relationship between length from the summary to the raw text. This dependent variable created has a normal distribution. You can see here that the difference between text word length and summarize word length is not large since all the values are below 0.4.
There was more exploratory data analysis, but these graphs are enough to give us a good view on what we are dealing with in the data perspective.

Conclusion.

The results of the EDA were that the dataset is a dialogue between two persons, the text contains a lot of nouns & verbs, the raw text is not a large dialogue because it only includes mostly less than 1000-words length. In the future we could change the dataset we are using since our dialogues maybe bias since our raw text is relatively small and is a dialogue dataset and create a different metric which measure the text ratio between summarize text/raw text.

Code
20– 3 / 20 + 150 * 100 = 10%

References:

[Exploratory Data Analysis for Natural Language Processing: A Complete Guide to Python Tools

- neptune.ai]