CMPS 470 PA1 Project Report

Team: ChatBLT

Team Member: Zachary Gros, Rhett Hill

Given that our data is comprised of strings, we decided to first remove any unnecessary characters such as numbers, stopwords, and any special characters in order to better work with the data. In order to show the changes that our preprocessing has done to our data, we chose to show the data in a plot comparing the string length to the total word count in the data for both the raw data and preprocessed data. These plots can be seen below. Looking at the plots, we can notice by the axis labels that the preprocessed data has been significantly simplified with the maximum string length being reduced to only 25,000 from 30,000 and the word count being reduced to 3500 to 8000. We also created a separate plot using min-max scaling on the preprocessed data to attempt to see any further changes to the data, but none were found.

A graph with blue and orange dots

AI-generated content may be incorrect.A graph with blue and orange dots

AI-generated content may be incorrect.