CS 446 Project 3 Report

Kanav Talwar

Dec 11, 2023

First Analysis Question

The collection's stories have an average length of 1214.2 tokens. The shortest story, which is "19406-art53," comprises only 4 tokens and the longest story, which is "8951-id_6," contains 26139 tokens.

Second Analysis Question

The word "the" appears in the highest number of stories, being present in 966 stories. It is also the word which occurs the most frequently, having 96151 occurrences

Third Analysis Question

There are 27217 unique words in the given collection, out of which 10056 occur only once. Given the numbers, approximately 37% of the words occur only once. This is a pretty reasonable outcome as given a collection of all possible tokens which occur, some of them can occur only once in the collection.