CS 446 Project 3 Analysis Questions

1. What is the average length of a story in this collection? What is the shortest story (and how short it is)? What is the longest story (and how long is it)? Note that for this project, "short" and "long" are measured by the number of tokens, not the number of characters.
    a. The average length of a story in this collection is 1215.1629098360656.
    b. The shortest story has StoryId: 19406-art53 and its length is 4.
    c. The longest story has StoryId: 8951-id_6 and its length is 26139.

2. What word occurs in most stories and how many stories does it occur in? What word has the largest number of occurrences and how many does it have?
    a. The word “the” occurs in most of the stories. It occurs in 996 stories.
    b. The word that has the largest number of occurrences is “the”. It has 96150 occurrences.

3. How many unique words are there in this collection? How many of them occur only once? What percent is that? Is that what you would expect? Why or why not?
    a. There are 27217 unique words in this collection.
    b. The number of terms that occur only once is 10056.
    c. The percentage is around almost around 37%.
    d. Yes, that is what I expected. The distribution of word frequencies in a collection often follows Zipf's law, which states that the frequency of a word is inversely proportional to its rank in the frequency table. This means that while a large percentage of words occur only once, there is also a small set of highly frequent words that account for a significant portion of the total word occurrences.

        Studies have shown that in large text collections, such as corpora or large-scale document collections, the percentage of hapax legomena can range from 30% to 60% or even higher. This means that a significant portion of the vocabulary consists of words that occur only once throughout the entire collection. For instance, in the example given on page 76 of the book, 70,000 words out of a 200,000-word vocabulary, only occurred once. Which is approximately 35%.

4. Your training queries have two queries that are roughly about the *scientific american supplement.* Suppose that you wanted to judge stories for relevance using a pooling strategy that takes the top 100 documents from each of those two queries. How many unique documents will you be judging? What if you only considered the top 20? Suppose you had a budget that allowed you to judge at most 25 documents. How deeply could you go into the two queries for judging to get 25 judged, no more, no less?
    a. Query 1 yielded 100+ results, whereas Query 2 yielded only 42 results. In total, there are 100 unique documents.
    b. If we only take the top 20 documents for both queries, then, we have 30 unique documents in total.
    c. We would need the top 15 queries to get 25 docs.