CS 446 Project 3 Report

Ilmin Cho

May 16, 2023

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 longis it)? Note that for this project, "short" and "long" are measured by the number of tokens, not the number of characters.

   Average length = 1215.1629098360656 / 19406-art53 = 4 / 8951-id_6 = 26139

2. What word occurs in the most stories and how many stories does it occur in? What word has the largest number of occurrences and how many does it have?

   ‘the’: 966 docs / ‘the’: 96151 times

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?

   27217 unique terms, 10056 only once. It is around 37%. It is difficult for me to predict the exact value, but it can be calculated to some extent using zipf'law and the larger the size of the corpus and the more diverse the types, the higher the rate of unique words that appear only once, so our sample has a moderate amount of sample and diversity, so 37% is a reasonable number.

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?

   58 unique documents on top 100 docs. / 0 unique documents on top 20.

   To get 25 unique documents, go in to 64 ranking

5. Run the query *amherst college* -- where that means the two words separately and *not* the phrase -- using either BM25 or QL (your choice). For any 10 of the top 50 top ranked documents, look at the text of the document and mark whether it is relevant. Put your judgments in a file called *amherst-YOURUSERNAME.qrels* Your should include the 10 storyIDs and a judgement of relevant that is 0 = has nothing to do with Amherst College, 1 = Amherst College is mentioned, or 2 = substantially relates to someone from or

something that happened at Amherst College. Use the qrels file format from P2, with the queryname being "amherst", then the skip value of 0, then the storyID (NOT your internal docid), and then the 0/1/2 judgment, one of those per line.