COL106
Assignment 7 (CORPUS Q&A TOOL)
Report

In the second segment of the assignment, our primary focus was on optimizing the existing code to yield enhanced results when executing the same query. To achieve this, we implemented several methodologies, detailed below:

1. **Elimination of Stop Words:**
In an effort to refine the accuracy of the query, our initial step involved the removal of stop words. These words, commonplace in the English language, are often excluded by search engines during the crawling or indexing process to conserve space and processing resources. Examples of such stop words include "the," "is," "at," "which," and "on."

2. **Word Score Adjustment:**
Our next approach involved modifying the score of words using a tailored method, aiming to derive superior scores for individual words. The goal was to identify the top k paragraphs within the corpus.

For the modifying the score of words in query we have used formula that :

$$\boldsymbol{Score(w) = log[\frac{Total\ word\ in\ corpus}{Count\ of\ w\ in\ corpus\ +1}]}$$

After that we have simply calculated the score of a paragraph as formula for a query q:

$$\boldsymbol{Score(para)} = \sum_{\boldsymbol{s\ in\ q}} \textbf{Score}(\boldsymbol{s}) * (\textbf{count of word in para})$$

These methods are better than the ones previously used part 1 because they assign higher scores to less frequent words in the corpus. By doing so, we increase the importance of rare words, leading to more accurate search results for the given query.

3. **Fixed Value of K:**
Our final adjustment involved setting the value of K to 4. This decision was made strategically, as opting for a smaller value might lead to data loss, while a larger K could potentially retrieve irrelevant data.