SWE 225 Information Retrieval
Assignment 3 Milestone 3 Optimized Inverted Index and Ranking

Pooja Bhatia, 14567073
Rutuja Pansare, 55443146
Ella Dodor, 31129450
Samyak Jhaveri, 13043185

Test queries:

1 – cristina lopes

2 - machine learning

3 - ACM

4 - master of software engineering

5 - pooja lopes

6 - cristina rutuja lopes

7 - master of computer science at University of California Irvine

8 - Information retrieval

9 - 2022

10 - Master of Science 2022

11 - UNIVERSITY OF CALIFORNIA IRVINE

12 - collaboration

13 - distributed computing research

14 - global object tracking

15 - sourcererCC

16 - genetic duplication

17 - neurochemical transmitter

18 - donald bren

19 - deep learning

20 - accessibility research

1. In boolean retrieval as of Milestone 2, we took intersection of common documents of the words in the query. As a result we only got results for documents where all of the words in the query were present. If even one word was not present in the corpus, the output of the boolean query was empty due to no common documents. We did not implement simhash so we got repetitive pages and content.
    - Ranking performance was not very good as compared to cosine similarity we implemented in M3.
    - Runtime performance took longer than 1 second in M2 because inverted index was not modified.

2. In cosine similarity, we optimized our index by creating 26 different files (all alphabets in english) for faster retrieval of documents. Using simhash we avoided documents with similar content.
    - Ranking performance was good as compared to M2.
    - Runtime performance took longer than 500 milliseconds because the inverted index was not sorted.

3. Later on we sorted our index, for faster retrieval of docs. We then used binary search to find word of query in the index
    - Ranking performance was good as compared to M2.
    - Runtime performance took longer than 500 milliseconds because the inverted index was not sorted.

Consider the queries ‘pooja lopes’, ‘cristina rutuja lopes’

In boolean retrieval we got 0 results whereas in cosine we got relevant docs

Rest all queries, more relevant docs were displayed on top by using cosine similarity

EXTRA CREDITS:

1. Simhash similarity for near duplicates (>90% similarity)
2. Implemented Web Interface