Information Retrieval Report (Assignment 2)

Part 3 (Worked on 3B)

Name: Abhinandan De
Roll no: 19CS10069

Key Features:

We reuse the following implementations from part 2A

- *Transpose the inverted index to decrease space complexity*
- *Tf-idf implementation*

The difference here lies in the use of NumPy arrays instead of dictionaries. Since we do not need to optimize for time to such an extent, we use NumPy since the code becomes legible with the easy use of vectorized operations and it doesn't necessitate the use of custom_add functions for adding two sets of tokens.

General Procedure

All in all, we just compute the transpose of the document vectors and obtain a similar representation for the query vectors. The tf-idf implementation is varied as per the methods which were required.

Now, we just fetch the queries along with cord ids from the fetched documents and then fetch the corresponding document vectors from the transposed index constructed earlier.

After that, we take the mean for the top 10 documents corresponding to each query which gives us a vector of size |V|.

Now we simply take the 5 greatest values for the tokens and use an inverse mapper to find the exact words which have the highest importance.