COMPSCI 446
Search Engine
P4 Report

Professor: David Fisher
Student: Yunfei Luo

March 25, 2020

1.
The source code of all things about indexing is defined in the class called *inverted_index* in *class_lib.py*. Both BM25 and Query are defined in their corresponding class, which is also in *class_lib.py*.

2.
This project use the same codes in the previous project for building inverted index. For both Query Likelihood and BM25 ranking method, they share the same structure:

```
for each document do
    for each query terms do
        compute cumulative score
    end for
    update score for the document
end for
return result
end of algorithm =0
```

Except they have different criteria for calculating the score.

3.
a) The package *gzip* was used for reading the input file.
b) The package *time* was used for visualizing the running time for each steps.
c) The package *json* was used for extract the content in the input file as python acceptable structure.
d) The package *math* was used for calculate the $\log(\cdot)$.

4.
I didn't expect the results for $Q6$ to be good. Because "to be or not to be" didn't contain much meaningful information regarding the content of Shakespeare's works. Moreover, the majority terms appear in this queries seems more like stop words that are much likely to be removed from the original raw documents. As a result, the scores related to probability and frequencies will be very small,

which could directly lead to the result that is not considerable.

5.
The result could not be as good as expected. Since every scene need for setting things at some certain positions, the frequencies of these query terms could be high in most documents in collection, which will lead to the final scores with little differences.

6.
The result from BM25 on $Q5$ appears to be better which contains comedies.

As shown in the file *judgements.txt*, the QL method turns out to be better. Some of the top results are relevant to the query terms, and the rest are not that relevant. However, BM25 do show some relevant result, but they are not that ”top”.