A brief report on the implementation of BM_25 algorithm

BM_25 is an effective ranking algorithm based on binary independence model which also includes document and query term weights.

$$\sum_{i \in Q} \log \frac{(r_i+0.5)/(R-r_i+0.5)}{(n_i-r_i+0.5)/(N-n_i-R+r_i+0.5)} \cdot \frac{(k_1+1)f_i}{K+f_i} \cdot \frac{(k_2+1)qf_i}{k_2+qf_i}$$

-K1,K2 are parameters whose values are set empirically. In my implementation k1=1.2, k2=100.
-Based on this value of K is calculated by the following formula

$$K = k_1((1-b) + b \cdot \frac{dl}{avdl})$$

-dl is the document length and b=0.75.

-avdl is the average document length in the corpus.

-dl and avdl is calculated based on the unigram_tokens file produced in the previous assignment which contains word and its number of occurrences in the corpus.

-qf is as explained in the below example. It's the frequency of query words

For a query with two terms 'hurricane isabel' – qf =1

-No relevance information is given so ri and R as taken as 0

-N = size of the corpus. In my case its 1000

-"hurricane " occurs in 226 documents. So n1=226.This is being from the dictionary created based on the unigram_Df table produced from the previous assignment. This unigram_Df file contains terms and the list of documents the term occurs in.

-"hurricane" occurs 40 times in a particular document. So f1 = 40. This information is fetched from a dictionary created based on the unigram inverted index produced from the previous assignment. This file contains terms and the list of documents it occurs in along with the number of times the term occurs in each document.

- In this way value for each and every variables present in the formula is calculated. These values are substituted in the formula and score for each document pertaining to a particular query is calculated.

-A dictionary containing the documents and its score is created and sorted based on the score in the descending order. It is then written into an output file in the given format.