Kara Capps
927000068

03-03-2022

In the assignment we were given, the two tasks were to implement Vector Space Modeling and and BM25F. Starting with Vector Space modeling, I first used the getDocTermFreqs which took in Document d and Query q in order to get the number of terms in a document, and getQueryFreqs that took in Query q to get the frequency of terms in a query.

In VSMScorer.java, I normalized the term frequencies in normalizeTFs. Using two for loops, I was able to get the the value for the raw term frequencies and applied sublinear scaling through the equation 1+logf(Td) if f(td)>0, else tf = 0. I did this because it would help show the importance of each term in a document in order to improve retrieval effectiveness. After this I calculated the Document frequency which was the number of documents with the term t divided by the total number of documents.

I was then able to calculate idf: idf= ln(1+n)/(1+df(t)) where n is the total number of documents in document set, and df(t) = is the document frequency.

For the computation of netscore, I created a titleVector Hashmap containing the count of the query items in all the titles, and bodyVector Hashmap containing the count of the query items in all the titles. Lastly I created a Hashmap containing the frequency of query terms in the query. I then used the normalizeTFs(ifs, d, q) function to normalize everything, and computed the score using equation 1 in the PA#1 guideline document.

Development Data

Training Data