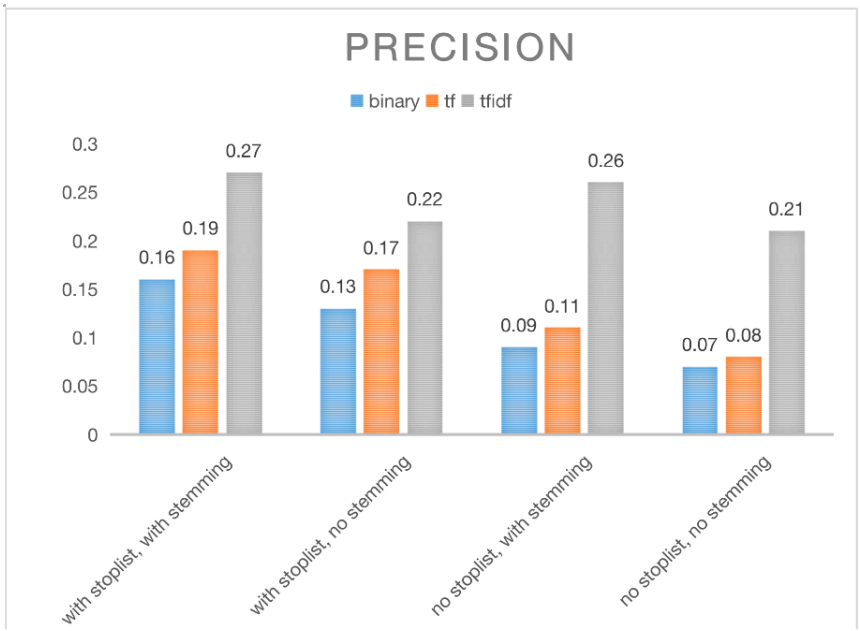
Information Retrieval Assignment Report

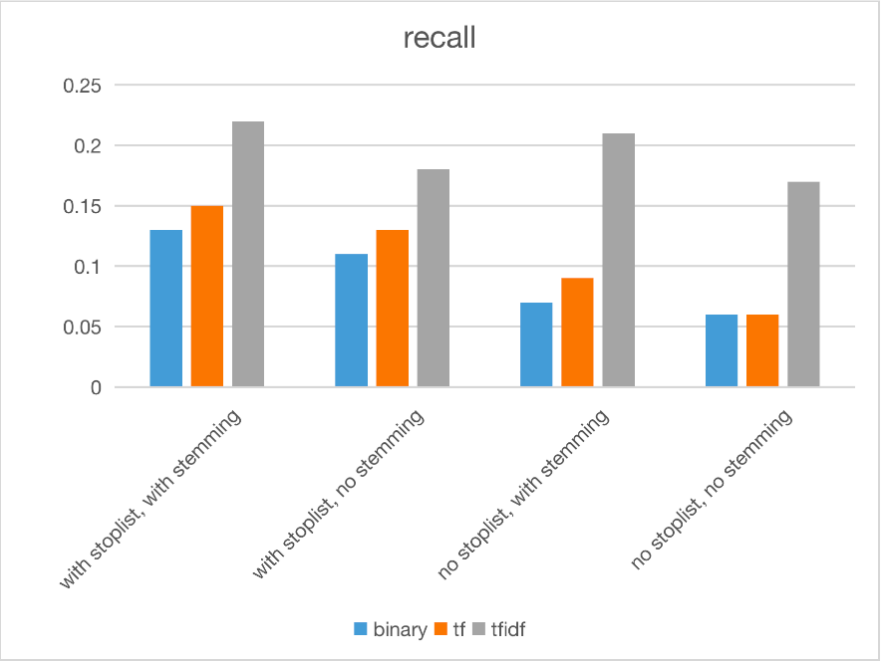
In my code file, I have implemented the extraction and comparison of document index and query index. At the start of my code, I create a new dictionary to store dicid, term and counts of terms by using a two-level dictionary. Then I calculated the required values, including total number of documents D, inverse document frequency of each term, the size of each document .etc. According three different term-weighting schemes, I computed similarities between query and each document based on the vector space model, and then sorting them in the sequence of their values. After testing all 12 conditions, the total time of code running is less than 30 seconds.

Here are the evaluations of results by comparing output files with gold standard file.

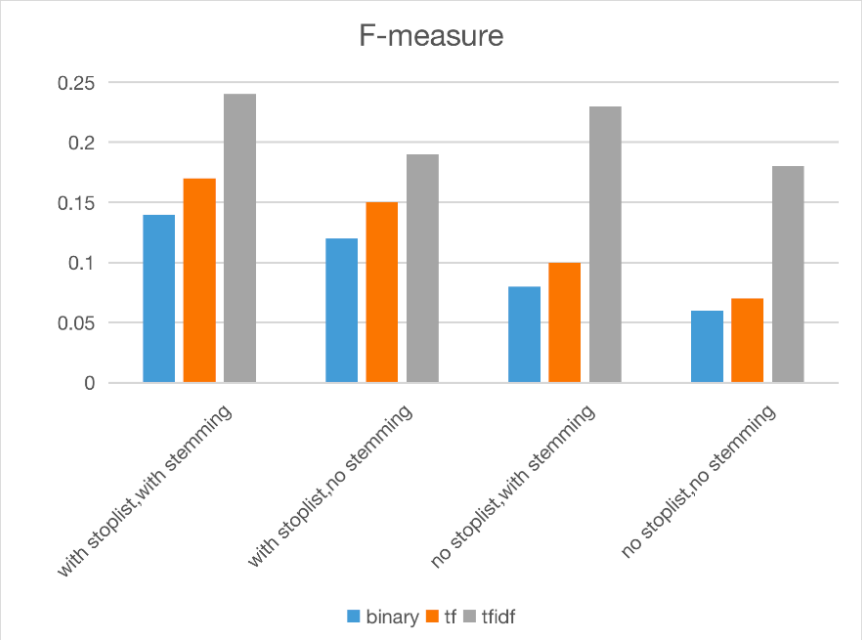


Graph 1

For the graph1, it shows the precisions of 12 different conditions. As we can see, the precision of TFIDF scheme is always the highest comparing with binary and TF schemes. When the index file and query file are with stop list and stemming, the precision is higher than other index files and query files.



Graph 2



Graph 3

For the graph 2 and graph 3, they show the recall and F-measure values across all queries, and we can draw the same conclusion: the value of TFIDF scheme is the highest, the value of files which are with stop list and stemming is higher than any other index and query files.