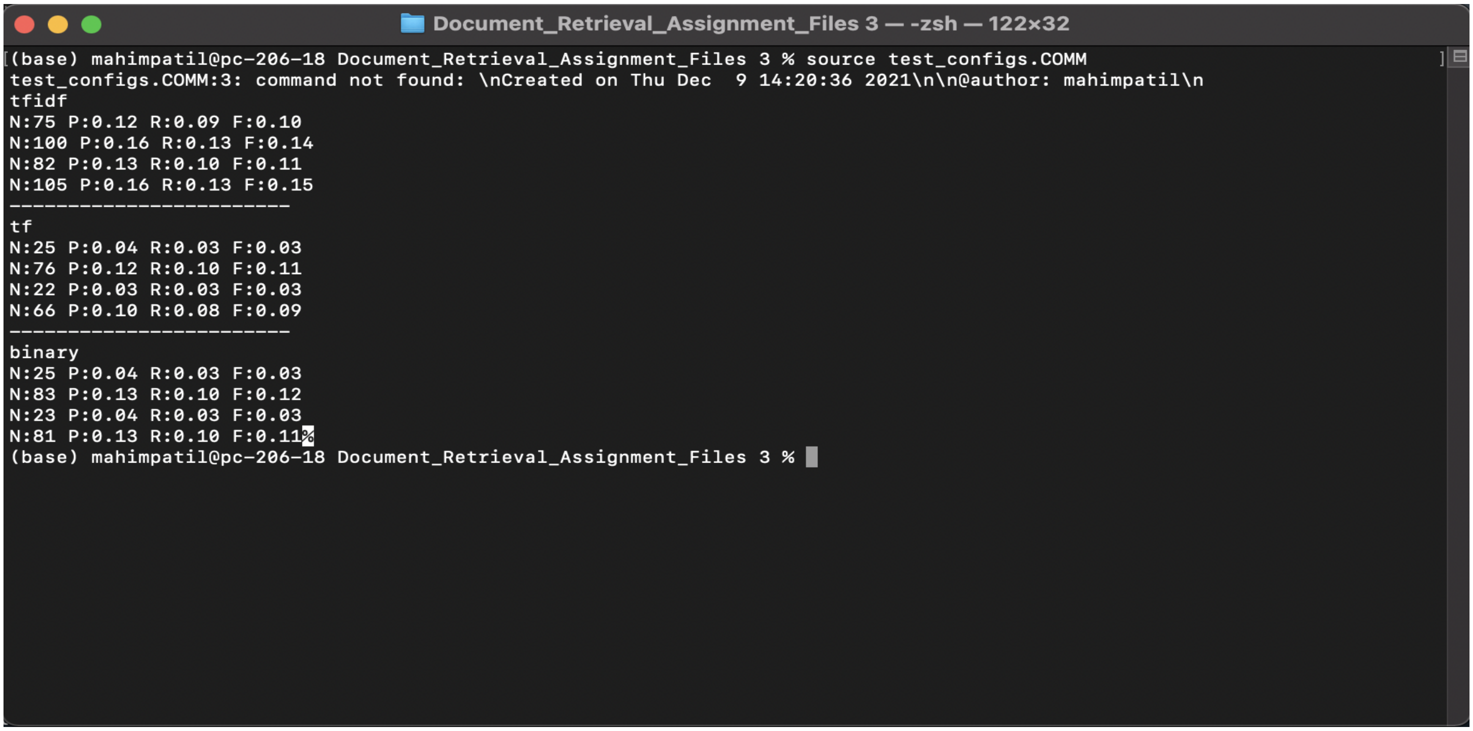
1. Implementation Details

Weighting Scheme	Details
Binary	In Boolean model implementation, the weights are either 1 if term present in document else 0. Based on that, The Document's score depends on number of query terms present in the document. The model was implemented successfully.
TF	For Term Frequency scheme, Document vector size has been calculated at the beginning to reduce compute time. The Query term frequency is multiplied with Document term frequency. Then the cosine similarity is calculated to get the top 10 relevant documents.
TFIDF	For TFIDF scheme, TF of inverted index is precalculated at the beginning to reduce compute time. Here we find the IDF for each document and then Σ *qidi* is calculated by multiplying each query term IDF with IDF of document term.

2. Results and Discussion



```
[(base) mahimpatil@pc-206-18 Document_Retrieval_Assignment_Files 3 % source test_configs.COMM
test_configs.COMM:3: command not found: \nCreated on Thu Dec  9 14:20:36 2021\n\n@author: mahimpatil\n
tfidf
N:75 P:0.12 R:0.09 F:0.10
N:100 P:0.16 R:0.13 F:0.14
N:82 P:0.13 R:0.10 F:0.11
N:105 P:0.16 R:0.13 F:0.15
------------------------
tf
N:25 P:0.04 R:0.03 F:0.03
N:76 P:0.12 R:0.10 F:0.11
N:22 P:0.03 R:0.03 F:0.03
N:66 P:0.10 R:0.08 F:0.09
------------------------
binary
N:25 P:0.04 R:0.03 F:0.03
N:83 P:0.13 R:0.10 F:0.12
N:23 P:0.04 R:0.03 F:0.03
N:81 P:0.13 R:0.10 F:0.11%
(base) mahimpatil@pc-206-18 Document_Retrieval_Assignment_Files 3 %
```

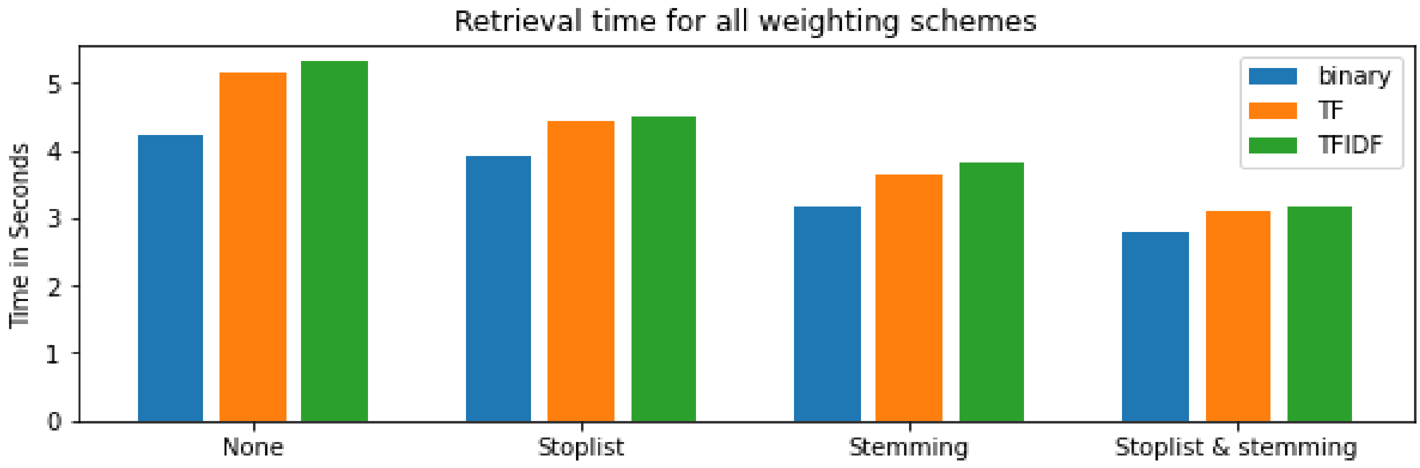
Observations

From the above terminal screenshot, it is evident that TFIDF weighting scheme has overall better performance. The best score is when stop list and porter stemming is used. It can observe that in TFIDF weighting scheme, The score is lowest among TFIDF results, when no term manipulation technique (I.e stoplist and porter stemming) is applied. The score is highest among every result, when stoplist and porter stemming and TFIDF is used.

For TF and binary weighting scheme, the scores are similar when no term manipulation technique is applied. But binary has better score than TF when both stoplist and porter stemming is used.

Insights

Based on observed details, it is proved that when words are preprocessed by stemming and common words are removed using the stoplist. There is huge boost in performance in all 3-weighting scheme, when stoplist is used. On contrary, there is very low or minute change in performance when only porter stemmer is used. In conclusion, TFIDF with stemming and stoplist has the best precision, recall, F1-scores. That is because, the TFIDF score is inversely proportional to informativeness.



In the above bar graph, The X-axis contains all 4 different configurations (1. without using stoplist & stemming, 2. using Stoplist, 3. using Stemming, 4. using Stoplist and Stemming) and Y-axis represents time in seconds. It is clearly evident that when stoplist and stemming is used, the time taken is lowest for all the 3 weighting schemes because stoplist removes all trivial words and stemming preprocesses terms to their roots. When term manipulation is not done, the time taken is highest among the results, which is obvious as many common words are present.