1 Implementation of Weighting Scheme

1.1 Binary

The binary scheme is fully implemented. The term weight is either 1 or 0. For every candidate document, $q_i d_i$ is the number of common terms that appear in both query and the document, and d_i^2 is the number of terms in the document. The similarity score of a document is higher if it contains more terms from the query.

1.2 Term Frequency

The TF scheme is fully implemented. To reduce the retrieval time, the size of each document vector ($\sqrt{\sum_{i=1}^{n} d_i^2}$) is calculated before the loop. For every candidate document, $q_i d_i$ is the multiplication of the frequency of the query term with the frequency of the query term in the document. The similarity score of a document is higher if it contains more terms that have higher frequency in the document and the query.

1.3 Term Frequency - Inverse Document Frequency (TFIDF)

The TFIDF scheme is fully implemented. To prevent repeating the same calculation in the loop, the TFIDF value of each term in the query is calculated beforehand. For every candidate document, the $q_i d_i$ value is the multiplication of the TFIDF of the query term with the TFIDF of the query term in the document.

2 Results

No stoplist, **No** stemming			
	Binary	TF	TFIDF
Rel_Retr	44	50	132
Precision	0.07	0.08	0.21
Recall	0.06	0.06	0.17
F-measure	0.06	0.07	0.18
Time (sec.)	0.26	0.32	0.76

No stoplist, **With** stemming			
	Binary	TF	TFIDF
Rel_Retr	59	73	166
Precision	0.09	0.11	0.26
Recall	0.07	0.09	0.21
F-measure	0.08	0.10	0.23
Time (sec.)	0.29	0.36	0.80

With stoplist, **No** stemming			
	Binary	TF	TFIDF
Rel_Retr	81	103	140
Precision	0.13	0.16	0.22
Recall	0.10	0.13	0.18
F-measure	0.11	0.14	0.19
Time (sec.)	0.09	0.07	0.18

With stoplist, **With** stemming			
	Binary	TF	TFIDF
Rel_Retr	105	126	172
Precision	0.16	0.20	0.27
Recall	0.13	0.16	0.22
F-measure	0.15	0.18	0.24
Time (sec.)	0.11	0.13	0.28

Table 1: Results of three weighting schemes under four different configurations

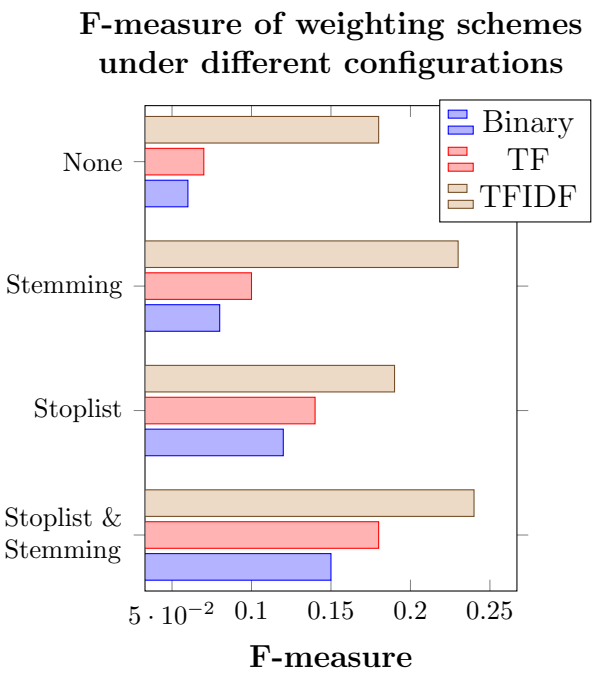

Figure 1: F-measures of three term weighting schemes

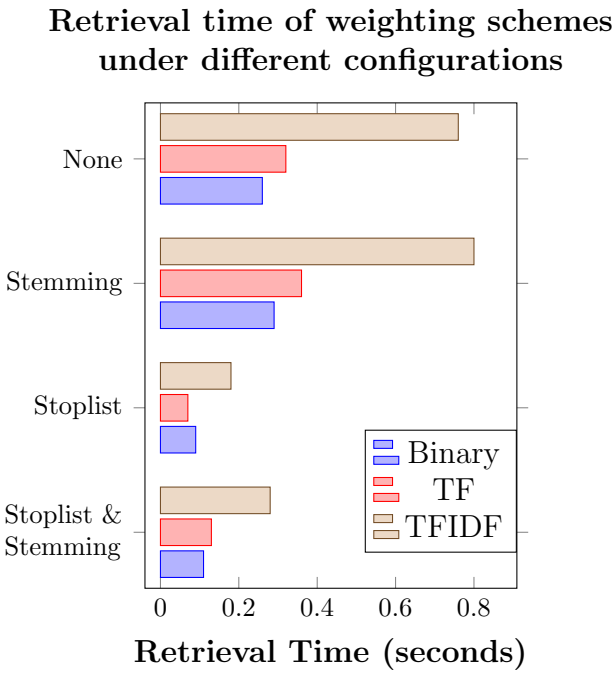

Figure 2: Retrieval time of three term weighting schemes

3 Discussion and Conclusion

According to **Figure 1**, it has been shown that TFIDF weighting scheme has the best performance, and binary weighting scheme has the poorest performance. In **Figure 2**, the retrival time is completely opposite, where TFIDF weighting scheme required the longest retrieval time, and binary weighting scheme is the overall fastest.

Based on the results in **Table 1** and **Figure 1**, it is proved that the results will improve when words are preprocessed by stemming and trivial words are removed using the stoplist. It is noticed that the TFIDF weighting scheme has very small improvements when stoplist is used, but huge performance gain when stemming is used. In contrast, binary and TF weighting scheme have larger improvements with stoplist but smaller improvements with stemming. Reason is that the words that are infrequent in the document collection are regarded as important in TFIDF weighting scheme, and it is likely that an important keyword in both query and document will reduce to become a same 'word' after stemming.

According to **Figure 2**, it is proved that the retrival time will be greatly reduced when using stoplist, but slightly increase with stemming. This is because stemming will increase the number of common words between the query and the document, thus the number of loops and calculations are increased which will result in a longer retrieval time. In addition, the retrieval time is affected by code refactoring as well. Some values can be calculated in the initialisation stage to prevent repeated calculations in the loop.

As a conclusion, the performance of the IR system could be improved by the usage of term manipulation techniques such as stoplist and stemming. The efficiency of the IR system is mainly affected by the implementaion approach, where sensible code refactoring could prevent repeated code executions. Lastly, given the results shown above, it would be reasonable to choose TF over binary weighting scheme for situations where both time and quality of results are the main priority.