1 Implementation of Weighting Scheme

Before implementing the weighting schemes, it is necessary to create a list of candidate documents for each query to improve the retrieval efficiency. Candidate documents are those that contains at least one 'term' from the query. Hence the list of candidate documents can be generated using list comprehension with sets intersection function. In addition, when calculating the size of each document vector, it is only required to loop through all the 'term' that is in the document, thus saving a lot of computational cost.

All term weighting schemes are implemented using vector space model, which is represented as the modified equation with the component $\sqrt{\sum_{i=1}^{n} q_i^2}$ dropped:

$$SIM(Q, D) = \frac{\sum_{i=1}^{n} q_i d_i}{\sqrt{\sum_{i=1}^{n} d_i^2}}$$

1.1 Binary

The term weight is either 1 or 0. $q_i d_i$ increases by 1 if the term is in the query, and d_i^2 increases by 1 if the term is in the document. The similarity score of a document is higher if it contains more terms from the query.

1.2 Term Frequency

The term weight is equal to the term frequency in a specific document and the query. The similarity score of a document is higher if it contains more terms that have higher frequency the document and the query.

1.3 Term Frequency - Inverse Document Frequency (TFIDF)

The TFIDF scheme is fully implemented. The term weight depends on its frequency in the relevant documents, query, and in the collection.

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

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

With stoplist, **No** stemming

	Binary	TF	TFIDF
Rel_Retr	84	104	140
Precision	0.13	0.16	0.22
Recall	0.11	0.13	0.18
F-measure	0.12	0.14	0.19

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

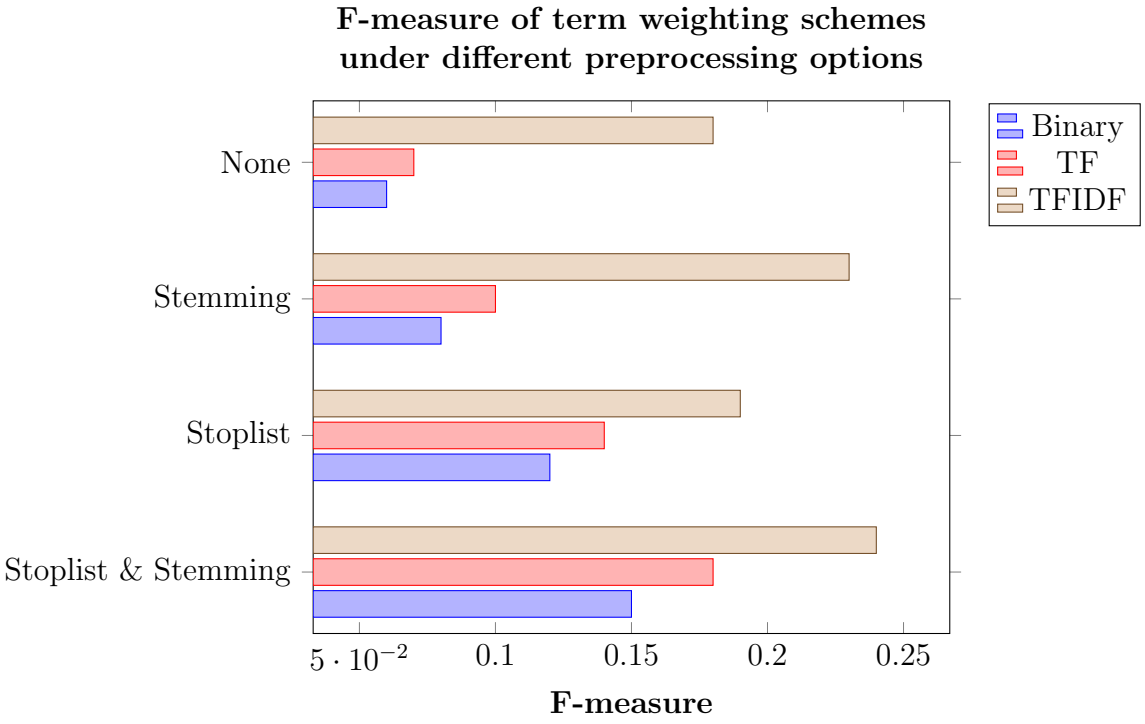

Figure 1: F-measure bar chart of three term weighting schemes

3 Discussion and Conclusion