The
University
Of
Sheffield.

Text Processing

Assignment 2: Document Retrieval

Name: Daotan Liu

Student ID: 200206596

Implementation

The code implements text retrieval in different configurations (weighting strategies and preprocessing strategies). The results of retrieval have been compared with the gold standard results. The running efficiency of the program is recorded as well.

The code retrieves the document in three different weighting strategy.

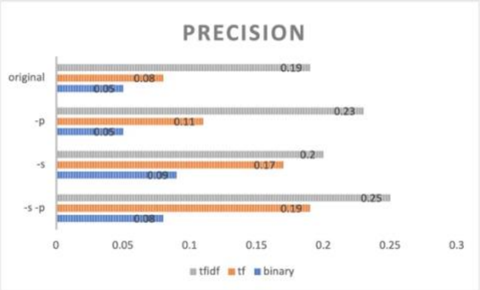
1. In the first strategy (*binary*), the value is set to 1 when the document contains the word in the query, otherwise it is set to 0.
2. In the second strategy (*tf*), the weight is determined by calculating the *document freq*.
3. In the last strategy (*tfidf*), using (1) to produce the result.:

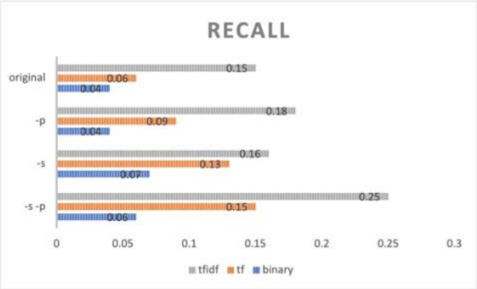
$$tf.idf = term\ freq * \log_{10} \frac{size\ of\ collection}{document\ freq} \qquad (1)$$

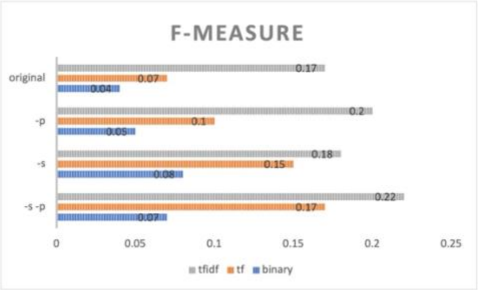
Finally, use the cosine of angle is used to calculate the similarity.

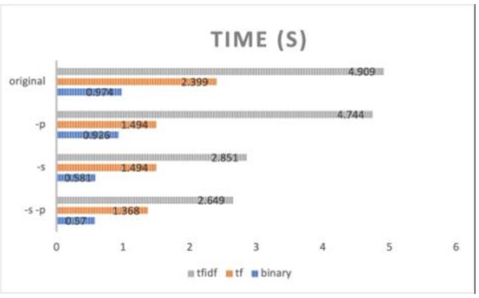
Performance in different configurations:

label	Options	Precision	Recall	F-measure	Time(second)
tfidf	-s -p	0.25	0.2	0.22	2.649
	-s	0.2	0.16	0.18	2.851
	-p	0.23	0.18	0.2	4.744
	original	0.19	0.15	0.17	4.909
tf	-s -p	0.19	0.15	0.17	1.368
	-s	0.17	0.13	0.15	1.494
	-p	0.11	0.09	0.1	2.313
	original	0.08	0.06	0.07	2.399
binary	-s -p	0.08	0.06	0.07	0.57
	-s	0.09	0.07	0.08	0.581
	-p	0.05	0.04	0.05	0.926
	original	0.05	0.04	0.04	0.974









Inference

Comparison between the weighting strategies.

In terms of precision, the tfidf weighting strategy performs the best, and tf is the second while binary is the worst. In terms of Recall, tfidf is also the best, and is significantly ahead of the other weighting strategies, with binary being the worst. Again, tfidf achieves the highest F-score, followed by tf, with binary coming in at the bottom. However, tfidf is the slowest and binary is the fastest, mainly because the tfidf model is the most complex one and involves the most computation while binary only needs to determine whether the term is present in the document and does not require complex computation.

The compare of different pre-processing strategies in a single weighting strategy.

It can be seen that for F-score, recall and precision, the combination of -s and -p is usually the best one and the original is the worst one. Single -s or single -p have their own advantages and disadvantages. This is mainly due to the fact that the introduction of the stoplist reduces the number and interference of common words, thus giving more weight to the core words. The stemming of words makes it easier for the algorithm to integrate related words, making the discrimination more accurate. In terms of time, the introduction of stoplist and stem reduces the running time significantly, mainly because the dimensionality of the vectors is reduced, and the operations are simpler.

In summary, the tfidf is most effective, for it bring the importance of the word into calculation. The introduction of stoplist and stem not only improves the performance of the model, but also reduces the running time.