Document retrieval report

(Yuhang Hu)

This report is organized as two parts: performance results and running time analysis. The first part reports the evaluation results under different configurations, the second part looks into the execution time of the code in different configurations and presents some findings about retrieval efficiency.

1.Performance results under different configurations

The following table summarizes my performance results under different configurations.I'll discuss my results and conclusions in different term weighting schemes.

Term weighting scheme		Binary	TF	TF*IDF
No stemming or stopwords	precision	0.07	0.07	0.19
	recall	0.06	0.05	0.15
	F-score	0.07	0.06	0.17
Stopwords	precision	0.13	0.16	0.2
	recall	0.11	0.13	0.16
	F-score	0.12	0.14	0.18
Stemming	precision	0.1	0.1	0.23
	recall	0.08	0.08	0.18
	F-score	0.09	0.09	0.2
Stemming and stopwords	precision	0.16	0.19	0.25
	recall	0.13	0.15	0.2
	F-score	0.14	0.17	0.22

Figure 1. Performance results

1.1 Binary term weighting

Under the binary weighting scheme, a term in the query is given value 1 if it's in a document or 0 otherwise, regardless of the number of its occurrences in the document. One can expect that the information retrieval accuracy can be low under this simple weighting scheme as two terms would have the same weight even if one term appears more times in a document than another. From the table above, we can see that this aligns with the results obtained without any term processing method(i.e. skipping stop word or stemming), which are 0.07,0.06,0.07 in precision,recall and F-score respectively.

1.2 Term frequency(TF) weighting

TF weighting weights term as number of times it appears in a certain document. So different terms have different significance when retrieving documents according to a given query, leading to better results

comparing with binary setting. In my results, this may not seem evident without any term processing method at first, but it achieved better results when both stop words and word stemming are used (0.19, 0.15 and 0.17). Given the importance of skipping stop words in text processing, only ignoring stop words without doing any word stemming also achieve a slightly better result(0.16, 0.13 and 0.14).

1.3 Term frequency and inverse document frequency(TFIDF) weighting

In addition to term frequency, inverse document frequency (IDF) also consider a term's frequency in a document collection.The IDF can be expressed as follow:

$$\text{IDF} = \log\left(\frac{\text{number of documents in a collection}}{\text{number of documents containing a certain term}}\right)$$

A term's TF*IDF value would be large if it appears in a small number of documents but multiple times in a particular document, so TFIDF weighting takes a term's informativeness within a document collection into account and achieves better performance when retrieving relevant document. Not surprisingly, the performance measures under TFIDF scheme in my results are highest among all the weighting schemes, no matter word processing methods are used or not(the last column in the table above).

2.Execution time analysis

The efficiency of information retrieval is also important. The following graph shows retrieval time in seconds between different weighting configurations. It's clear that when both stop word list and term stemming are used, the execution time is reduced significantly in all the weighting scheme. These preprocessing methods have indeed improved the retrieval efficiency.

Another interesting finding is that the retrieval efficiency improves a little under TF and TFIDF weighting scheme when only the term stemming method is used, and the execution time even increased a bit under binary weighting, which is possible since term stemming needs extra time to process, whereas the efficiency improved remarkably when only the stop word method is used. This finding suggests stop words processing technique carries more weight in improving retrieval efficiency and is an important part of performing information retrieval.

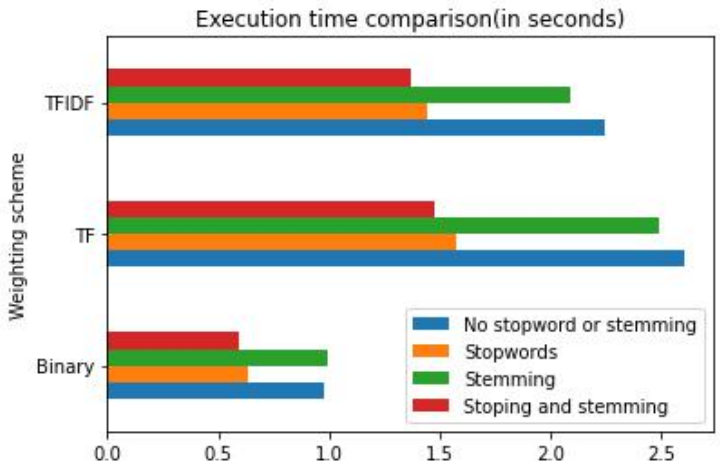

Figure 2.Retrieval time comparison(in seconds)