COM6115: Text Processing (2019/20)
Assignment: Document Retrieval

Student Registration Number: 19018654

Results and Discussion

To speed up the code, I used Dictionary Comprehensions, preprocessing, and other methods frequently, while doing as little as possible operation in the for loop. In the end, the maximum time for the code to run is within 0.3 seconds (the minimum time is 0.146s).

Time analysis

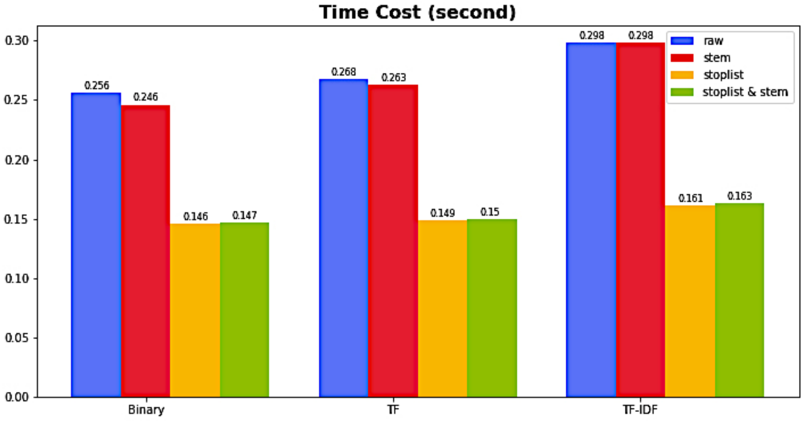

Figure 1: Time Cost

In figure 1, whether it is binary, tf or tfidf, the raw data (blue bar) takes the longest time; after using stemming (red bar), the time is slightly shortened; when stoplist is used (orange bar), the time is almost reduced by half; when using both stemming and stoplist (green bar), the time consumption is also reduced by half.

Using stemming: a word and its different forms are considered the same word, so the total number of words is reduced. Thus, the speed will be slightly faster.

Using stoplist: According to Zipf's law, frequently occurring words account for nearly 50% of the total number of words. These words are not the most useful for retrieval. So, when using stoplist, these frequently occurring words are excluded, and the total number of words is reduced by about half. As a result, code execution is also nearly half faster.

Evaluation score analysis

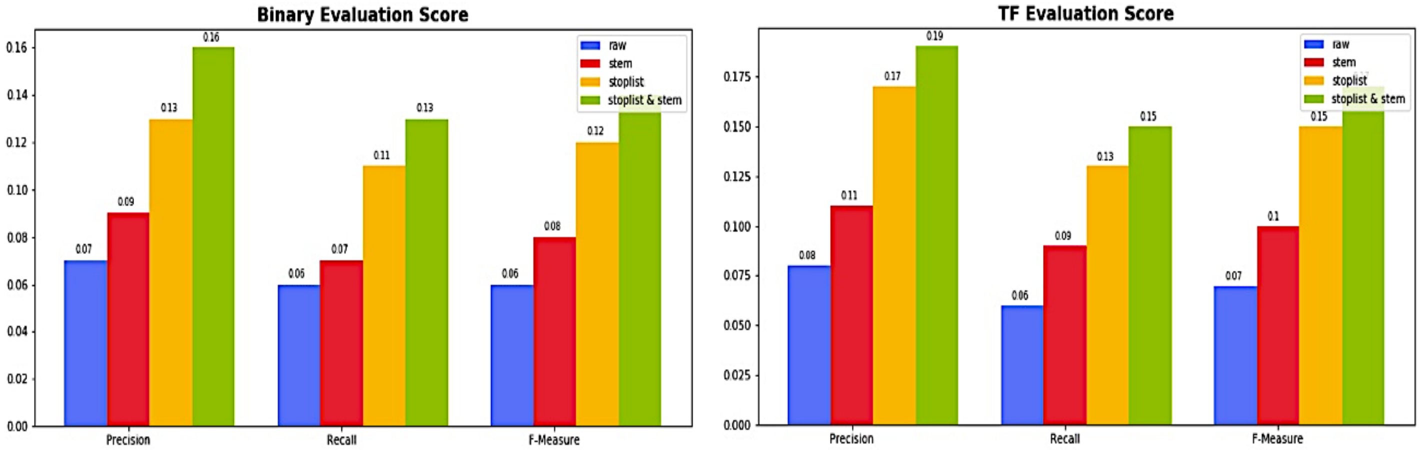

Figure 2: Binary and TF Evaluation Score

The figures above show the evaluation scores, including precision, recall and F-measure from left to right in Binary and TF modes, respectively. When using stemming (red bar), due to the mixture of morphological variations, words with the same meaning and different forms are treated as the same word, and the score improves. Even better is the use of stoplist (orange bar), which cuts out a large number of unimportant high-frequency words. Works best when used with both methods (green bar).

However, in the case of TF-IDF, the results will be a little bit different, just as shown in figure 3.

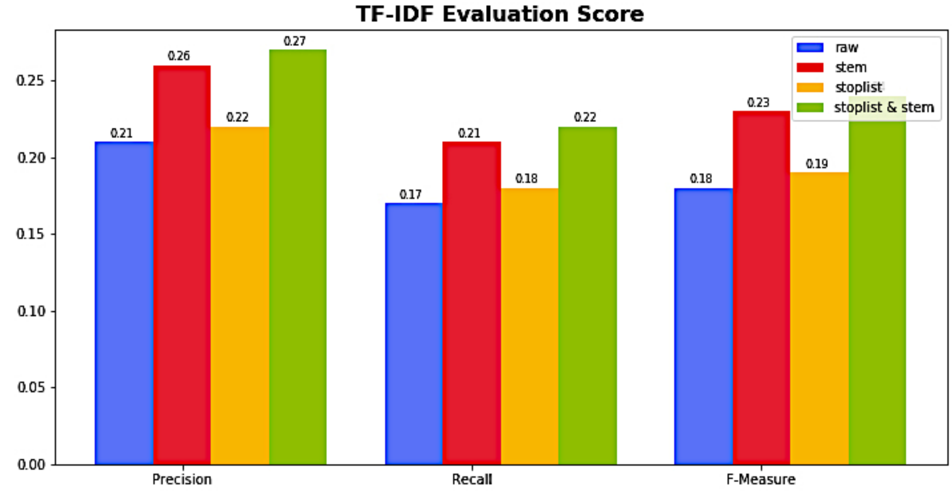

Figure 3: TF-IDF Evaluation Score

In the case of TF-IDF, all words with their weights are taken into account. There may be a small number of common words that are useful for information retrieval. However, after using stoplist (orange bar), these useful words have been removed, so the score dropped compared to stemming. The use of stoplist has little help for the TF-IDF mode, while in the current mode, stemming (red bar) has improved the score considerably.

Experimental Data Table

Mode	Data Sources	Time(s)	Precision	Recall	F-measure
Binary	Raw	0.256	0.07	0.06	0.06
	Stoplist	0.146	0.13	0.11	0.12
	Stemming	0.246	0.09	0.07	0.08
	Stoplist & Stemming	0.147	0.16	0.13	0.14
TF	Raw	0.268	0.08	0.06	0.07
	Stoplist	0.149	0.17	0.13	0.15
	Stemming	0.263	0.11	0.09	0.1
	Stoplist & Stemming	0.15	0.19	0.15	0.17
TF-IDF	Raw	0.298	0.21	0.17	0.18
	Stoplist	0.161	0.22	0.18	0.19
	Stemming	0.298	0.26	0.21	0.23
	Stoplist & Stemming	0.163	0.27	0.22	0.24