Report

1.Description of implementation

This IR system is very simple, the main method is the vector space model. According to the term weighting schemes I construct a vector, or embedding, for doc and query.

For binary :

$$score = 1$$

For tf : score =

$$score = \frac{word_{freq}}{word_{count}}$$

It indicates the ratio of the number of times a word appears in the document to the total number of words in the document.

For tfidf : score =

$$score = \frac{word_{freq}}{word_{count}} * log(\frac{nums_{docs}}{nums_{worddocs} + 1})$$

"idf" represents the ratio of the total number of documents in the corpus to the total number of documents containing the word.tfidf is the product of "tf" and "idf"

Then I calculate the cosine similarity between the doc vector and the query vector, and sort according to the cosine similarity to get the final recall document.

2.Result

The following are the results of my experiment. In order to do the ablation study, I used the controlled variable method.

Experiment	Weight Schemes	Stoplist	Stemming	Precision	Recall	F-measure
1	binary	no	no	0.07	0.06	0.06
2	binary	no	yes	0.10	0.08	0.08
3	binary	yes	no	0.13	0.10	0.12
4	binary	yes	yes	0.16	0.13	0.15
5	tf	no	no	0.08	0.06	0.07
6	tf	no	yes	0.11	0.09	0.10
7	tf	yes	no	0.17	0.13	0.15
8	tf	yes	yes	0.19	0.15	0.17
9	tfidf	no	no	0.20	0.16	0.18
10	tfidf	no	yes	0.26	0.21	0.24
11	tfidf	yes	no	0.22	0.17	0.19
12	tfidf	yes	yes	**0.27**	**0.22**	**0.24**

I drawed the following conclusions from the above results.

- **1.**The best results appear when using stoplist, stemming and tfidf.
- **2.**Use stoplist and stemming both have a positive effect on the improvement of results, and the improvement of stoplist is greater（except in “tfidf”）. I don’t know which stemming method professor used (maybe by nltk), it doesn’t seem to be very effective, maybe I can try the tokenizer method in BERT[1].
- **3.**Why stoplist is effective? because it filters some words that often appear but have little meaning (such as "the", "of", "by", etc.), it make feature engineering more effective.
- **4.**From the results, the metrics of "tf "and "binary" are not much different.. This is because they only consider the frequency of a word in the document without considering the importance of the word itself. At this time, "idf" is needed, and "idf" will give common words smaller weights, so features are more robust, but "tfidf" also has problems. This calculation can not reflect the location information and the importance of the word in the context. Therefore, it is better to use some large natural language models. such as BERT.

Reference

[1] Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.