

Web Searching and Mining Project

Shu Hong Liu

May 20, 2019

Abstract

abstract

Keywords: TF-IDF, Okapi, Language Model

1 Introduction

I used WT2G datasets which is part of WT10G from TREC Web Corpus. Since 1992, a series of annual benchmarking evaluation exercises, called TREC (Text REtrieval Conference), have launched in the USA. TREC experiments were designed to allow large-scale laboratory testing, compare the effectiveness and performance of different information retrieval techniques. TREC has become the standard in the IR field. So, here I used WT2G that is a smaller datasets from TREC.

In WT2G, we can find that there are 247491 documents and about 1,500,000 unique words. That is a big challenge for me, because I studied in Statistic and the number of documents is too big for me to do something. Fortunately, professor Tsai recommend Indri toolkit to do information retrieval more easily than I thought before.

2 Method

Here I am going to introduce the methods that I used.

2.1 Vector Space Model

Okapi Tf and IDF

2.2 Language model with Laplace smoothing

2.3 Language model with Jelinek-Mercer smoothing

In Jelinek-Mercer smoothing, I take

3 Trials

Before using methods that I mentioned above, I construct two indexes, (1) without stemming, and (2) with stemming. In each index, I used all of methods above, so there will be 2 * 4 runs.

3.1 without stemming

3.1.1 TF-IDF

3.1.2 Okapi

3.2 with stemming

4 Idea

In many trials, We can see that the results are bad enough. However, that means there's still room for improvement

5 Conclusion

Maybe in the next time, we do this project that we can use more powerful method.

paragraph test paragraph
test indent