Report of Programing Homework1

R04522839 Xu Guanghao 1st year majored Mechanical Engineering

Describe my VSM

Method

From the given file vocab.all I can get a unigram vocabulary and the total number is not that large(less than 30000), so I simply compose my VSM by every character. So every document can become a one-dimension vector of 30000 units.
BUT that lead to my performance not good. I consider it as my major problem. I should choose to use Chinese words-split package to build my VSM. And of cause build a new vocab.all. After several trials on my VSM it's been too late for me to change it (Deadline is approaching).

Speed

I used set to form vector and export output vector to a txt document ID.txt. I think the best solution is a database and I'd like to use Mysql or Sqlserver. But this is my first time to use workstation and I am not sure it could work. I received the email sent by TAs that sqlite could be used. I am not familiar that and I think this is a place can be optimized.

Describe my TF-IDF

Method

I implemented TF and IDF, and use it in VSM as return value install of original only 1 and 0.
A problem occurred. My java heaped when I calculate whole document IDF. So installed I used one-fourth part of raw material to calculate IDF which lower my performance.

Problem

I tried to implement TF-IDF using "inverted-file "which worked out that a failure. Maybe I used it wrongly. A lot of time wasted.

Describe my Rocchio Relevance Feedback

Method

I select top 10 from 100 documents I feedback as Relevance. In fact the best relevance should be marked by human. Except that, I can't figure out a better idea.

Parameters

I put new query as 90% from original query and 10% from top 10 documents. And the round number is 1. I don't have time to try a lot of times (my code is too slow which really limited my trials) I tried 50% to original parts and the result didn't improve but down a lot which means it's really not good to give too much feedback.

Results of Experiments

First time: without any tools.no TF-IDF, no Feedback: 0.47
Second time: with TF-IDF, no Feedback: 0.60
Second time: with TF-IDF, with Feedback (50% to original query): 0.48
Fourth time: with TF-IDF, with Feedback (90% to original query): Need updated

Discussion

It's really amazing to attend this CS class and do such a midterm project as a student usually work with machines XD. I feel shame on my performance which is just around baseline. I think there are 3 points that I can improve:

1. Use words-split package install char to build VSM
2. Use sqlite install txt documents
3. Use XML parser install segmentation.

What's more, I regard to use Ruction oriented program method install of object oriented program, which brings me a lot troubles.

I will keep working on this project and try new models on this although midterm programing homework deadline havs arrived.