Search Engine Presentation

ZiHeng Wang

Directory

Title

Method/Optimization

Crawler

Crawlers are the foundation of search engines.

For this big assignment, my crawler is written in python2.7 and uses libraries like urlparse and bs4 to simplify the difficulties.

Crawler

Specific implementation:

From the very beginning of the school of Information website, each URL is traversed using BFS.

For each url traversed, use wget for crawling.

Crawler

Details to note:

① Need to judge all run out of the information institute web page.

② Remove duplicate sites(like index.php and the URL with '#').

③ Delete illegal links (such as download links, email and inaccessible urls).

Word Processing

Word processing is also a very important part of search engines.

It is also written in python2.7 and uses libraries such as jieba and HTMLParser to simplify the difficulties.

Word Processing

Specific implementation:

After you normalize the web page (which you can do using BeautifulSoup), get all the text through handle_data(HTMLParser), and then cut the words with jieba.cut.

Word Processing

Details to note:

① Stop words need to be removed after word cutting.

② It is necessary to record whether each piece of text is necessary to obtain (for example, endnotes and other contents, we do not need to obtain).

TF - IDF Algorithm

Tf-idf algorithm is the core of search engine. It is basically the same as what is taught in class.

$$w_{t,d} = (1 + \log \mathrm{tf}_{t,d}) \times \log_{10}(N / \mathrm{df}_t)$$

TF - IDF Algorithm

I made some simple optimizations:

① The elements with large IDF are weighted.

② When there are too many query words, delete words which have small IDF.

③ Add weight when the current word is time.

④ It is assumed that the words in the title are more important.

Optimization can ensure the search speed and improve the accuracy

Web

Complete by using python's flask library.

Bootstrap was also used to make the HTML more beautiful.

It is a pity that I can't do better because of the time and my poor skill.

Web

Method/Optimization

Web

Title

Some of the Methods I Explored

WAND

A Safe algorithm.

However, I have tried and found that the length of crawling web articles is usually short, and there are not many documents, so using the WAND algorithm is still likely to make all documents need to accurately calculate their vectors, or even slower than the plain one, so it is abandoned.

(This also took some time to study, so sad

Title

Actual Display

Title

Summary

Summay

In this is not a very short week, from the beginning of nothing, to now although the search engine has done, but the results are still so-so.

It is also a pity that I failed to finish many things I wanted to do because Of my weak basic skills and lack of knowledge in this field. Python was also the first time I came into contact with, so I spent a lot of time in searching materials and programming.

Summay

However, this course is at least a start. I learned a lot of knowledge and tried to put it into practice. I encountered many setbacks and had the help of teachers, teaching assistants and classmates.

In a word, I still have to work hard in the future.

THANK YOU FOR YOUR LISTENING!