Introduction to elgooG Search Engine

Zhipeng Chen

July 20, 2020

Abstract

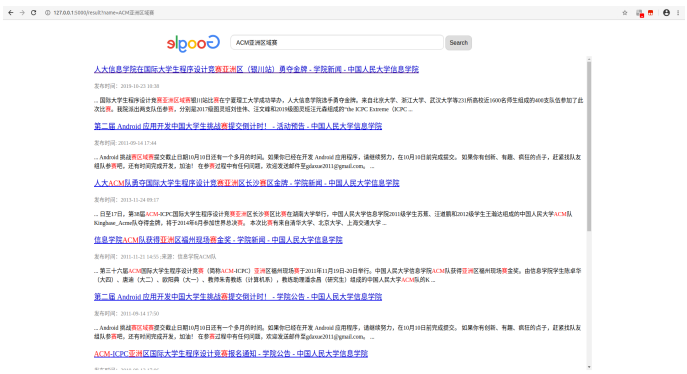
elgooG Search Engine can search the websites of the School of Infomation in Renmin University of China, and present the article which user wants to read.The search engine uses inverted index so that it can present the pretty relevant passage from all websites.

1 The function of engine

1.1 UI

The initial interface contains the logo and a text frame.

When user search something on the engine, it will jump to the interface showing the results.



1.2 Function

In the initial interface, users can type something into the text frame to search for the relevant passage. The engine will to show the resulte about the question that users asked.

Each result contains the title, time, source and a brief instruction for the article(if they exist). The relevant element in the detial of the result will be showed in red.

In the result interface, users can search something different in the text frame. Engine will return the result about the new question.

2 The way to make engine

2.1 Crawler

In order to get all the websites of the School of Infomation, I use a crawler to download the websites. The crawler is written by C++ programming language. When it reach a website, crawler will search the html document about this website for the urls which haven't been reached. Then the crawler will download the unreaching websites by using the system command 'wget' and make a flag for it. Through the BFS(Breadth First Search), the crawler will download all the websites we need.

2.2 Normalized and divided

The code to translate the html document to text is written by python programming language. First, the program will read each document that have been download before. Then, it will find the title, source, time and main body of the article by using BeautifulSoup4(a package in python). Finally, it will save the detial in other documents.

The code to divided the normalized document is written by python programming language. The program will cut the document into some words by using jieba(a package in python), and save the results in other documents. After cutting document, it's convenient to search for the most relevant website.

2.3 Search for result

The elgooG uses the inverted index which is the classical model to build a search engine. The process is worked before the search engine start.

After that, the search engine can return the question that users asked. When the engine get a query, it will cut the query into some words by using jieba. Then it will calculate the sorce of each website and return the top 10 website as result. The fomula to calculate the sorce is the one below:

$$Sorce_i = \sum Value_{i,id}$$
$$Value_{i,id} = (1 + \lg(t_{i,id})) * (\lg(\frac{N}{n_{id}}))^{1.75} * (\lg(len_{id} + 1))$$

The $Sorce_i$ is the sorce of ith website; the $Value_{i,id}$ is the value of idth word in ith website; the t_{id} is the times of the idth word appearing in ith website; the N is the sum of the words in all website; the n_{id} is the times of the idth word appearing in all websites; the len_{id} is the lenght of the ith word.

In the words from query, if a word is made by number, it will have higher value. It's value can be calculate by the fomula below:

$$Value_{i,id} = (1 + \lg(t_{i,id})) * (\lg(\frac{N}{n_{id}}))^{1.75} * (\lg(len_{id} + 1)) * (\lg(len_{id} + 1) + 1.6)$$

3 The skills of programming

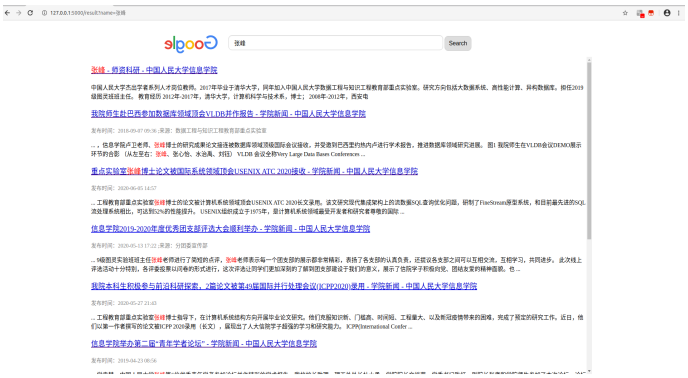
First, a huge project can be divided into some small part. Finishing the small part is easier than finishing the whole project at once. At the smae time, it's also easy to code and debug in the small part of project.

Second, coding with a clear mind. If you don't know what you are doing, you will make many bugs instead of writing correct code.

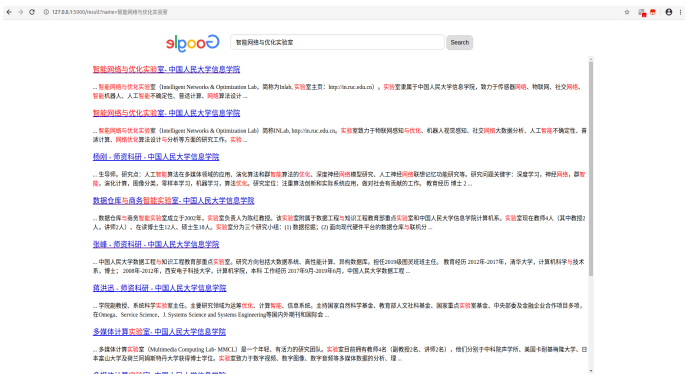
Third, learning the grammer of new programming language first, if you want to write code through it.

4 Some examples

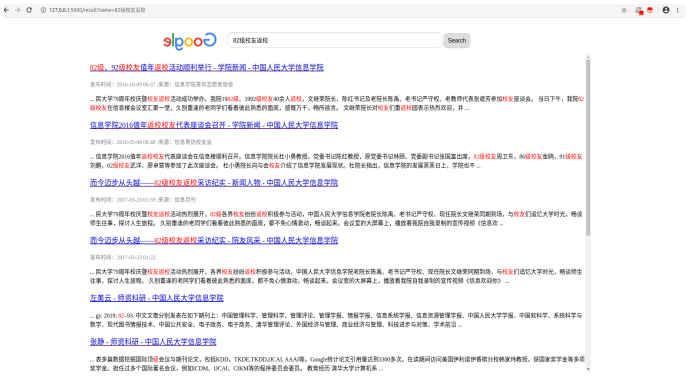
Search for teacher:



Search for laboratory:



Search for news:



5 Summary

The elgooG Search Engine can present the relevant website for each query. I use C++ to make a crawler to download, use python to cut document into words and use inverted index to solve the query.

6 How to build it in your computer

The search engine is based on Flask. You can download Flask by entering the command:

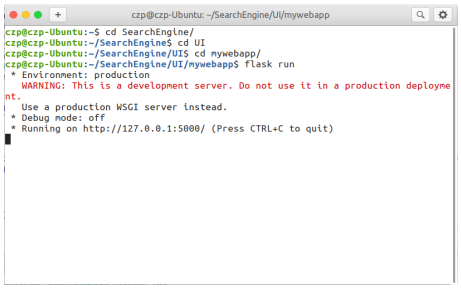
```
1 ~$ pip3 install flask
```

in terminal.

If you want to make the search engine work, you should enter the command below in the path $YourPathToEngine/SearchEngine/UI/mywebapp/$:

```
1 ~$ flask run
```

Then terminal will show the local url in your computer, like the picture below:



You can start the engine by entering the url into your browser.