My Search Engine Report

Wang ZiHeng

July 25, 2020

Abstract

This report will first give a brief introduction to my search engine, and then describe the purpose and implementation details of the codes used by my search engine, as well as the use of examples. At the end of the report, I will attach the way to use my search engine.

1 Introduction

My search engine is based on TF-IDF algorithm, and it supports inputting key statements to search and display the top 20 related pages in a short time. At the same time for each web page can display its title and related abstract.

2 Method/Optimization

2.1 Crawler

2.1.1 Introduction

Crawlers are the foundation of search engines.The role of crawlers is to get every URL we want.

For this big assignment, my crawler is written in python2.7 and uses libraries like 'bs4' and 'urlparse' to simplify the difficulties.The reasons for using these libraries are:

- In the 'bs4' library, 'BeautifulSoup' and other functions can easily get all the URLs under the current URL.
- In the 'urlparse' library, function 'urljoin' can easily normalize the URLs under the current URL.

2.1.2 Specific Implementation

Starting from the official website of School of Information, Renmin University of China, BFS is used to traverse each URL.

Besides, in the process of BFS, each URL is need to be marked to prevent it from being traversed again.

Then for each URL traversed, instruction 'wget -O' is used to crawl.
Finally, use file 'URL.txt' to store all the URLs.

2.1.3 Details to Note

- Need to judge all run out of the information institute web page.
- Remove duplicate sites(like index.php and the URL with '#').
- Delete illegal links (such as download links(you can check if the url have 'download' and the common file suffixes, such as .doc, .xls, etc.), email(urls which have '@') and inaccessible urls).

2.2 Word Processing

2.2.1 Introduction

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.

- In the 'jieba' library, function 'cut' can cut text into terms.
- In the 'HTMLParser' library, we can recognize the elements of HTML to find the text we want.

2.2.2 Specific Implementation

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

Besides, we can create an inverted index of each term and record the number of times term appear in the document. Also we can generate abstract of all web pages.

In addition, in this link, we also need to preprocess the following documents:

- wordlist.txt: Store all the words we get from the web pages.
- title.txt: Store the title of all web pages.
- doclen.txt: Store the vector length of all web pages(the method of vector modulus length involves the following part of the algorithm).
- index.txt: Stores the inverted index of each term and the number of times it appears in the document.

2.2.3 Details to Note

- Stop words need to be removed after word cutting (stopwords.txt).
- It is necessary to record whether each piece of text is necessary to obtain (for example, endnotes and other irrelevant contents, we do not need to obtain).

  Method: Use the 'Htmlparser' to get the tag, and then record the type of text after the tag according to the tag, such as:

  - Title / H1 - H6: title.
  - Class names in div tag include 'essay', 'news_Content', etc. : text.

  Stack can be used to know which text belongs to which tags.

2.3 TF-IDF Algorithm

2.3.1 Introduction

TF-IDF algorithm is the core of search engine. It is basically the same as what is taught in class. Use the following formula to calculate the weight of the term t in document d:

$$w_{t,d} = (1 + \log_{10} tf_{t,d}) \times \log_{10}(N/df_t) \tag{1}$$

In this formula, $tf_{t,d}$ is the number of times the term t appears in document d, df_t is the number of documents in which the term t appears, and N is the number of documents.

Then construct the vector of each document with these weights, and calculate the cosine value of the angle between the vector of each document and the search vector. According to the value, which document is similar to the search can be determined.

2.3.2 Optimizations

- The elements with large IDF are weighted.
- When there are too many query terms, delete terms which have small IDF.
- Add weight when the current term is time.
- It is assumed that the words in the title are more important, so it can be assumed that the words appearing in the title are repeated several times in the article, so as to increase the importance of the title.

These optimizations can ensure the search speed and improve the accuracy.

2.4 Web

2.4.1 Introduction

Complete by using python2.7's 'flask' library. Besides, 'Bootstrap' was also used to make the HTML more beautiful.

2.4.2 Specific Implementation

It is mainly about the method of abstract.

Because of the particularity of UTF-8 encoding, finding the abstract with keywords can't be done like ordinary string matching. So I cut the abstract, then match each term and keyword, and then splice the term with the term before and after, and finally form the abstract.

3 Examples

数据工程与知识工程教育部重点实验室 搜索

共展示20条结果

数据工程与知识工程教育部重点实验室 - 机构设置 - 中国人民大学信息学院

...数据工程与知识工程教育部重点实验室成立于2006年，是"985工程"二期"数据工程与知识工程"科技创新平台承担单位，中国人民大学建设的第一个省部级重点实验室。现任主任为教育部科技进步一等奖获得者杜小勇教授。 实验室以中国人民大学的学科优势为依托，...

来源：http://info.ruc.edu.cn/overview_structure_dept.php?id=13

数据工程与知识工程研究所 - 机构设置 - 中国人民大学信息学院

...中国人民大学数据工程与知识工程研究所是一个集教学、科研与应用开发为一体的高科技学术机构。担负着计算机科学与技术专业博士点的建设任务。 自1987年建所以来，始终站在学科前沿，跟踪国际先进技术，承担了大量国家重点研究项目，积极开展对外...

来源：http://info.ruc.edu.cn/overview_structure_dept.php?id=17

数据库与商务智能教育部工程研究中心 - 机构设置 - 中国人民大学信息学院

...商务智能教育部工程研究中心成立于2001年，致力于数据库与商务智能新技术研究、产品研发以及应用推广。 目前，该中心成立数据库与商务智能工程研究中心研究实验室、研发中心、测试中心以及办公室，并以人大金仓公司为市场运作实体成立了数据库...

来源：http://info.ruc.edu.cn/overview_structure_dept.php?id=15

明德图灵 搜索

共展示20条结果

UIUC张振杰博士为"明德图灵"学员做首场报告 - 学院新闻 - 中国人民大学信息学院

...为"明德图灵"学员做首场报告 11月12日，"明德图灵"厚重人才成长支持计划邀请美国加州大学香槟分校（uiuc）张振杰博士作首场报告。报告的的题目是"浅谈科研工作的影响力和流行度"，项目导师陈跃国副教授主持讲座，来自...

来源：http://info.ruc.edu.cn/news_convert_detail.php?id=1322

"明德图灵"主题培训之美国大学生数学建模竞赛培训会召开 - 学院新闻 - 中国人民大学信息学院

..."明德图灵"主题培训之美国大学生数学建模竞赛培训会召开 2016年11月5日，"明德图灵"主题培训之美国大学生数学建模竞赛培训会在信息楼417会议室召开。辅导员刘玲初与竞赛获奖学生姜亚宁、岳天泽介绍竞赛基本情况，并...

来源：http://info.ruc.edu.cn/news_convert_detail.php?id=1329

我院师生及"明德图灵"成员参观Pivotal公司 - 学院新闻 - 中国人民大学信息学院

...及"明德图灵"成员参观pivotal公司 2017年11月8日下午，"明德图灵"厚重人才成长支持计划成员与信息学院师生共40余人参观了pivotal公司，深入了解pivotal公司的研究成果和技术理念，并与pivotal中国地区副总裁 elisabeth hendrickson...

来源：http://info.ruc.edu.cn/news_detail.php?id=1464

4 Summary

In this is not a very short week, from the beginning of nothing, to now although the search engine has done, but it is a pity that 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. That's why I haven't been able to improve the accuracy of my search.

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.

5 How to Use This Search Engine

Before you start, you need to download the following libraries for your python2.7:

- bs4
- lxml
- jieba
- flask
- urlparse
- HTMLParser

Then you need to download 'SearchEngine.zip' from the github, which is a search engine that does not contain abstracts.

Run it after unzipping:

```
$ python Main.py
```

For the version which has abstracts, you need to download 'Spider.py', 'Check.py', 'Get_Abtract.py' to generate a text folder and then place the text folder in the location of Main.py.

```
$ ./Spider.py
$ ./Check.py
$ ./Get_Abtract.py
```

If you need to update, you need to download 'Spider.py', 'Check.py', 'Get_Text.py' and do the following:

```
$ ./Spider.py
$ ./Check.py
$ ./Get_Text.py
```

Then copy all the generated txt files('doclen.txt', 'index.txt', 'title.txt', 'URL.txt', 'wordlist.txt') to the location of Main.py.