

Nanako Search Engine

Linjiang Li

July 26, 2020

1 Introduction

This program is for website search using website spider,TF-IDF and inverted index to achieve.

You can crawl all the websites which can be reached from the URL you input.And run a small server on your PC. Then you can use browser to query the websites.

It also supports the search for pdf files.

2 Usage

(The usage is for Windows OS, but you can still do similar operations on Linux)

2.1

Python3($\geq 3.8.4$) is recommended.

2.2

To make sure you have the packages that will be used, run following command in your *cmd* or *bash*

```
pip install flask
pip install jieba
pip install BeautifulSoup
pip install pdfminer
```

2.3

This step is to make a database for following search

You should run *crawler.py*, *parser.py*, *index.py* one by one.

(To avoid being banned by the website server,*crawler.py* may take a lot of time.the rate of progress will be shown on the terminal)

Here is an example for Windows: run following command in your *cmd*

```
crwaler.py  
https://info.ruc.edu.cn (input the domain name you want to crawl)  
parser.py  
index.py
```

2.4

This step is about how to use the search engine.

Run *routes.py*, then use your browser access *https://127.0.0.1:5000/*

And you will see the homepage

Nanako Search

沒有人比我更懂info.ruc.e

Cai一下

已在风雨中度过 6 天 3 小时 36 分 33 秒

Figure 1: Nanako search: Homepage

Then you can input your queries in the search bar, and access the result page.



Figure 2: Nanako search: Resultpage

3 Details

3.1 Crawler

The crawler is implement with BeautifulSoup to get the URLs in label `<herf>`

Every accessible website will be downloaded in a file folder and numbered from 1 to n.

content.xxx(file type,like php or pdf), url.txt,suf.txt will be in the file folder.

Some urls will be ignored: the url with *download?*, the url with *mailto*, reduce




 content.php	2020/7/19 4:28	PHP 文件
 suf.txt	2020/7/19 4:28	文本文档
 url.txt	2020/7/19 4:28	文本文档

Figure 3: Nanako search: Crawler

index.php in urls.

3.2 Parser

The crawler is implement with BeautifulSoup, pdfminer and jieba.

The website content will be divided in to 3 fields: content,title and subtitle.

For html, content.txt includes the content with `<class=para>`, and ``. subtitle.txt includes the content with `<h1>`, `<h2>`. title.txt includes the content with `<title>`.

For html, it is difficult to divided the content into different fields. So all the content will be in content.txt

After that, each field will be divided into single words by jieba.

3.3 Index

Undoubtedly, different fields enjoy different importance. So I use Fieldweight to measure its importance. (title = 10, subtitle = 2, content = 1)

$$tf(word, document) = 1 + \log(\sum_{word \text{ in } document} appeartimes \times Fieldweight)$$

$$idf(word) = 1 + \log\left(\frac{N}{\sum_{each \text{ document}} tf(word, document)}\right) \quad N = \max(\sum_{each \text{ document}} tf(word, document))$$

$$score(query, document) = \sum_{word \text{ in } query} tf(word, document) \times idf(word)$$

`index.py` will return the highest 100 documents, and this step is implemented with multi-threads.

4 Example

Here shows some queries.



Figure 4: Example 1

本科生培养方案

100 条结果

[培养方案 - 本科生 - 中国人民大学信息学院...](#)

信息学院2017级理科试验班（信息与数学）培养方案2019-10-23 信息学院2017级图灵实验班（计算机科学与技术）培养方案2018-06-04 信息学院2016级理科试验班（信息与数学）培养方...
http://info.ruc.edu.cn/education_degree_list.php?type=1&inner=2

<http://info.ruc.edu.cn/userfiles/upload/f20170517024143922.pdf...>

信息学院SchoolofInformation理科实验班（信息与数学）专业培养方案一、培养目标本实验班培养具有扎实的数学和计算机科学与技术基础，能从事各领域的计算机与信息系统开发、应用、管理、建模与分...
<http://info.ruc.edu.cn/userfiles/upload/f20170517024143922.pdf>

<http://info.ruc.edu.cn/userfiles/upload/f20170517024337610.pdf...>

信息学院SchoolofInformation理科实验班（信息与数学）专业培养方案一、培养目标本实验班培养具有扎实的数学和计算机科学与技术基础，能从事各领域的计算机与信息系统开发、应用、管理、建模与分...

Figure 5: Example 2

信 息 学 院

School of Information

理科实验班（信息与数学）专业培养方案

一、培养目标

本实验班培养具有扎实的数学和计算机科学与技术基础，能从事各领域的计算机与信息系统开发、应用、管理、建模与分析的交叉复合型人才。本实验班含数学与应用数学专业、计算机科学与技术专业、信息管理与信息系统专业、信息安全专业和软件工程专业。学生进校时不分专业，学习期间通过选择课程形成专业，通过自主选择的培养模式和创新实践训练形成交叉、复合、个性化的知识结构和发展方向，并具备在各自感兴趣的领域进行独立分析和深入研究的能力。

Figure 6: pdf in Example 2

信息学院院长是谁

100 条结果

新闻公告 - 中国人民大学信息学院...

我校招生就业处王小虎处长来我院调研2015-06-03 靳诺书记一行到我院调研2015-06-03 第九届亚太地区信息学奥林匹克竞赛（中国赛区）在人民大学举行2015-05-13 信息学院杜小勇教授...

<http://info.ruc.edu.cn/keyword.php?name=杜小勇>

Figure 7: Example 3

新闻公告		
杜小勇(16)	· 我校招生就业处王小虎处长来我院调研	2015-06-03
	· 靳诺书记一行到我院调研	2015-06-03
	· 第九届亚太地区信息学奥林匹克竞赛（中国赛区）在人民大学举行	2015-05-13
	· 信息学院杜小勇教授当选2014年度中国计算机学会会士	2015-03-22
	· 人生，一步一步地选择——访信息学院院长杜小勇教授	2009-12-19
	· 杜小勇院长一行访问京东商城（360buy）	2011-09-05
	· 杜小勇院长和周晓方主任参加第四届中澳信息学院院长论坛	2011-12-05
	· 杜小勇院长应邀访问国立新加坡大学计算机学院并看望我院毕业生	2012-01-17
	· 杜小勇教授等参加第29届中国数据库学术会议NDBC 2012	2012-10-25
	· 杜小勇院长率团访问新加坡著名大学并看望我院校友	2013-01-25

Figure 8: Example 3 detail