Abstract

It is a simple web crawler of several search pages of ZHIHU.com. Through the project, we hope to get several kinds of data from searching-result pages of ZHIHU. Here are the keywords we have to search:

- 面试(interview)
- 实习(intern)
- 找工作(job hunting)
- 简历(CV)

You can visit ZHIHU_Search and enter the keywords to view the generated answer page.
From the answer page, we will collect several information, which forms a tuple of table.

- search_terms : the word you search
- search_rank: the rank of this tuple
- question_url: the link of the question
- question_title: the name of question
- question_follow_num: the number of followers of the question
- question_view_num: how many times the question has been viewed
- question_top_answer_username: the *name* of account whose answer ranks first among all the answers
- question_top_answer_id: the *id* of account whose whose answer ranks first among all the answers

In order to distinguish potential same tuples, we add a **"create_time"** column to record when the tuple is created.

How to start the crawler

```
python3 main.py
```

There are several command line parameters you may have to configure mannually to ensure the service can run normally.

- --db_user: the name of your account which you log in to the database
- --db_passwd: the password of your database account
- --db_name: the name of the database that you want to connect
- --db_addr: the IP address of the database that you want to connect

Outer reliance

Developing Environment

Personally, I build and test the structure on Windows 10 Professional, with MySQL 8.0.18 and Anaconda, whose Python's version is 3.7.4.

Target Environment

The project is designed to run on Linux server, relying on Python, whose version is bigger than 3.6, and PostgreSQL or MongoDB.

Docker configuration

In order to build it as a micro service, we provide **Dockerfile** in buildDocker folder. It is designed to work on CentOS 8, the latest version. Simply you merely need to add all python files into the folder and run dockerfile.

How to configure Chrome in Docker image

Though in **Dockerfile** we have configured how to install Chrome without GUI, there are still several steps which have to be done manually to ensure the project to work. Please *strictly* follow the below.

1. Run the image in docker, entering bash
2. Find the path of Chrome, create a soft link for the sake of use

```
which google-chrome-stable
ln -s [path] /bin/chrome
```

3. Solve the problem that *root* user cannot run chrome, which needs to modify file '/opt/google/chrome/google-chrome', modify the last line as:

```
exec -a "$0" "$HERE/chrome" "$@" --no-sandbox $HOME
```

4. Install chrome drive
    1. Download chromedrive built for installed version of Chrome
    2. build soft link and add 'x' mod

```
chmod +x chromedriver
sudo mv -f chromedriver /usr/local/share/chromedriver
sudo ln -s /usr/local/share/chromedriver /usr/local/bin/chromedriver
sudo ln -s /usr/local/share/chromedriver /usr/bin/chromedriver
```

Project Structure

|--main.py
|--spiders
    |--indexZhihu.py
    |--models.py
    |--multithread.py
    |--mysql_connect.py

|--to_xlsx.py

Here are the illustrations of these files.

- main.py : the entrance of whole project, accept command line paremeters, pass them to the function which connects to MySQL.
- indexZhihu.py : invoke all the modules defined in Spiders to finish the job of generating target info.
- models.py : The conglomrate of several practical functions which are used in other modules. For example, it provides the functions to capture a website, extract target urls, modify urls to standard format etc.
- multithread.py : In this file we define a class to execute multi-thread crawler. It is able to modify the number of threads you plan to use.
- mysql_connect.py : It is used to connect to the MySQL database, providing functions which respectively execute creating connection, closing connection and inserting tuples.
- to_xlsx.py: It intends to collect all tuples from table in SQL database and format the dataframe into an xlsx file.

We also have prepared createDatabase.sql for you to have a clear understanding of the design of our database. You can run it on your computer.

Demostration of running result of the project

Here is a screenshot of the table.

question_url	question_title	question_follow_num	question_view_num	question_top_answer_username	question_top_answer_id
https://www.zhihu.com/question/20602526	面试时怎样自我介绍比较好？	13891	3449848	夏姑娘	xia-gu-niang-66
https://www.zhihu.com/question/267849861	怎么判断自己面试是不是凉了？	2719	3805672	曾经的你	wxbfcc904a202fac2e
https://www.zhihu.com/question/35953016	面试应答有哪些话术和技巧？	82591	13210261	Annie姐	getready
https://www.zhihu.com/question/49088785	面试考官到底喜欢什么样的学生？	2452	628914	梁梁	liang-liang-35-5-99
https://www.zhihu.com/question/28058827	面试者如何回应面试官问的「你有哪些要问...	83345	7739130	量子位	liang-zi-wei-48
https://www.zhihu.com/question/282880854	面试想拿 10K，HR 说你只值 7K，该怎样回...	13773	19575218	陈诚	chen-cheng-58-22
https://www.zhihu.com/question/29708629	有哪些应届大学毕业生面试中能提高面试成...	33380	2430576	星海先森	deng-yu-520
https://www.zhihu.com/question/290543744	你有哪些面试失败的惨痛经验？	137656	30050123	再来一次	wei-xiao-dong-16-47
https://www.zhihu.com/question/295453435	面试当场就决定让我明天入职，靠谱吗？	1371	4124153	上官文商	shang-guan-wen-shang
https://www.zhihu.com/question/20390946	面试一定要穿正装吗？	864	922996	薇灵	shewill-48
https://www.zhihu.com/question/295572212	如何看待公司不主动给面试者倒水喝？	1129	7619526	潘神经	pan-shen-jing
https://www.zhihu.com/question/302600406	你在面试过程中哭过吗？	4185	13705711	苏桥	xing-lu-zhe-si-70
https://www.zhihu.com/question/327680271	为什么有些 95 后面试不带简历？	1217	2173811	西顾	qiu-yang-38-52
https://www.zhihu.com/question/19920401	有哪些应届生需要留心的面试技巧？	26423	2356813	江北烟雨人	Barosas
https://www.zhihu.com/question/294078513	面试中，HR一般会问什么问题？	20	21675	职越求职	careerspeedy-zhi-yue
https://www.zhihu.com/question/327988343	职场萌新，第一天去公司实习需要带什么东...	32	28560	王狗蛋	wang-zhen-hua-20
https://www.zhihu.com/question/28434997	从来没有实习经历的大学生如何找实习？	9376	1787664	Layne.Lu	layne-77-25
https://www.zhihu.com/question/34276390	实习期间你都学到了什么？	16655	1991248	空格	su-mu-39-14