# 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

# 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.