This document is written by Jingfei Yang

yangjingfeiwd@gmail.com

CLPS\_Crawl\_Seek

1. How to run the program

java –jar CLPS\_Crawl\_Seek.jar <PLACE>

<PLACE> can be ALL ACT NEW NT QLD SA TAS VIC WA

FOR EXAMPLE:

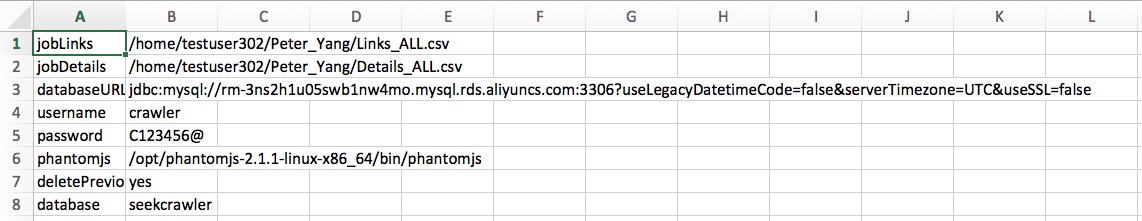
java –jar CLPS\_Crawl\_Seek.jar ALL

ALL means seek data from all of the states,

And the other words mean seeking data from specific state.

2. Configuration file

The name of the configuration file must be TheBasicDetails.csv



jobLink: The direction of the file which contains the websites’ Links

jobDetail: The direction of the file which contains the websites’ Details

databaseURL: The information and address of the database you want to input data in

username: The name of you in that database

password: Your password

phantomjs: The directory of the phantomjs

deletePrevious: yes means you will delete all of the previous data in that database

database: The scheme that you will use in that database

3. The program

3.1 All of the classes the program includes:

Main

HandleCSV

Crawl\_Link

EachState

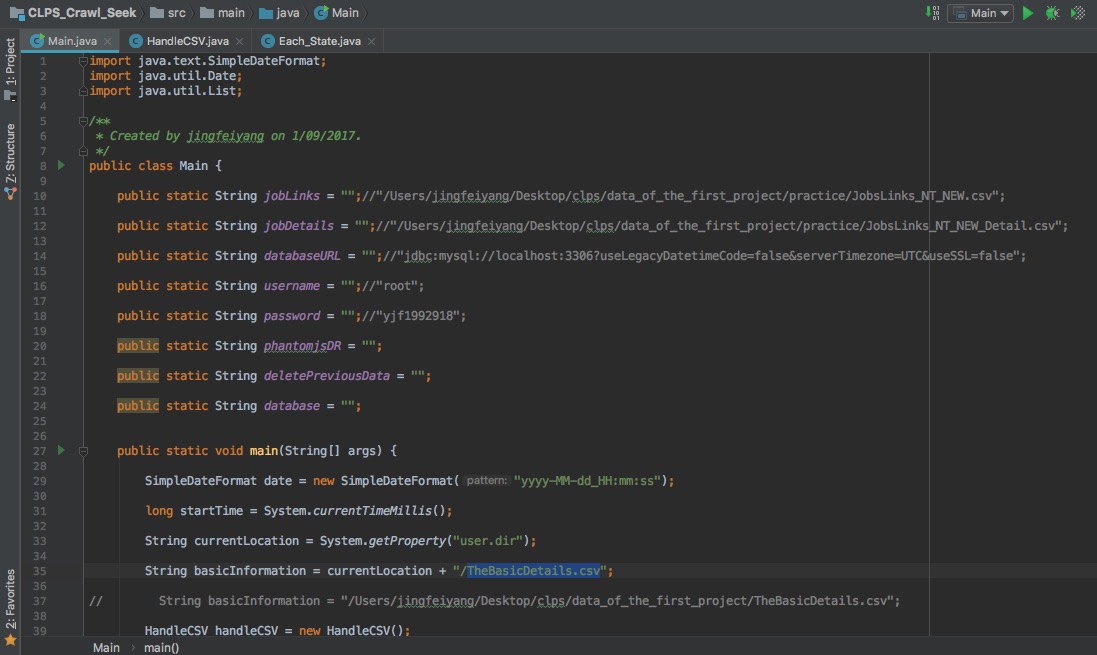
Crawl\_Details

HandleDatabase

3.2 Each class

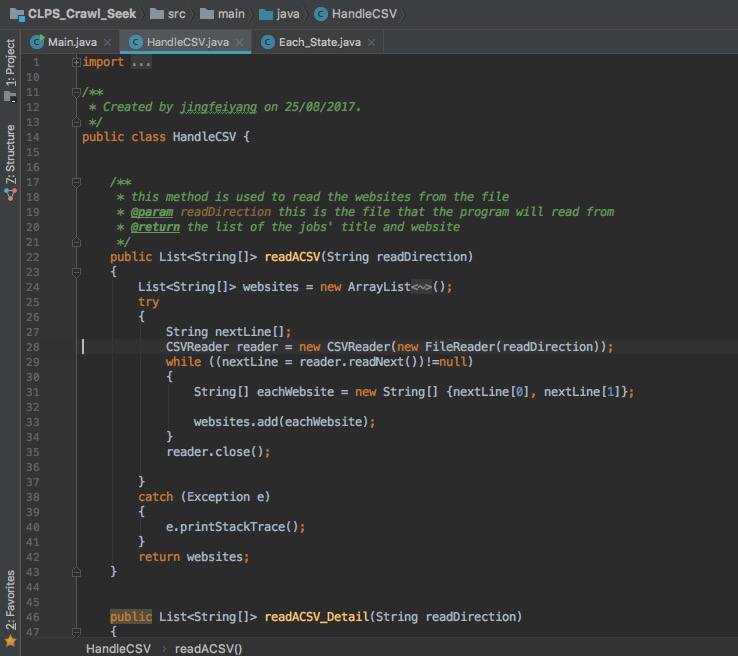
3.2.1 Main

This is the main class.



3.2.2 HandleCSV

This class is used to handle the CSV file



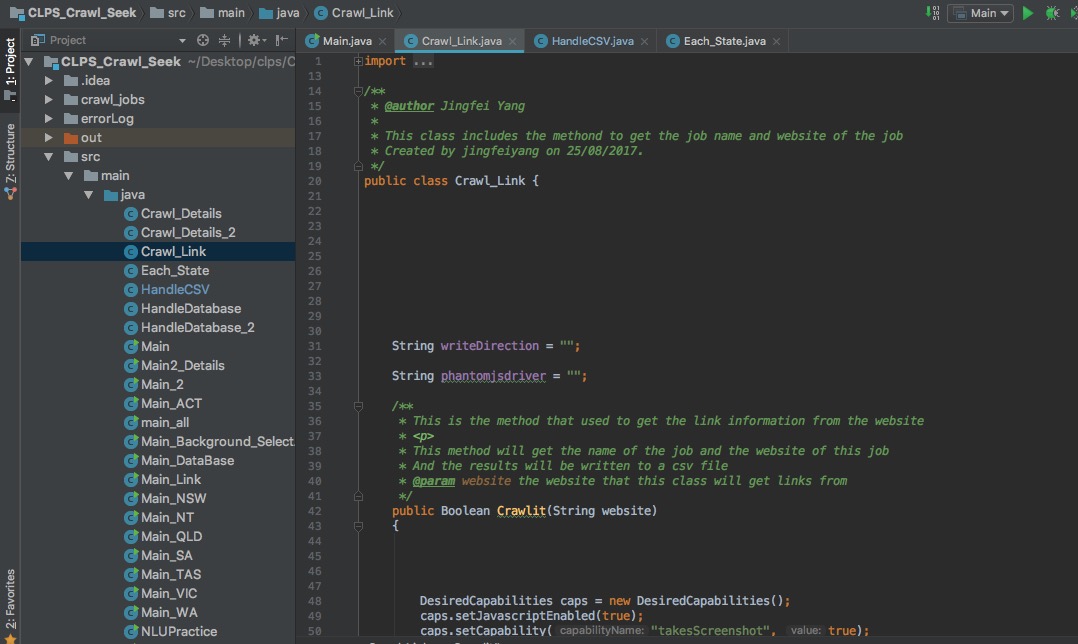
readACSV is used to read the link file and the basic file

readACSV\_Detail is used to read the detail file

writeACSV is used to write the data to csv file

3.2.3 Crawl\_Link

This class is used to seek each job and its websites

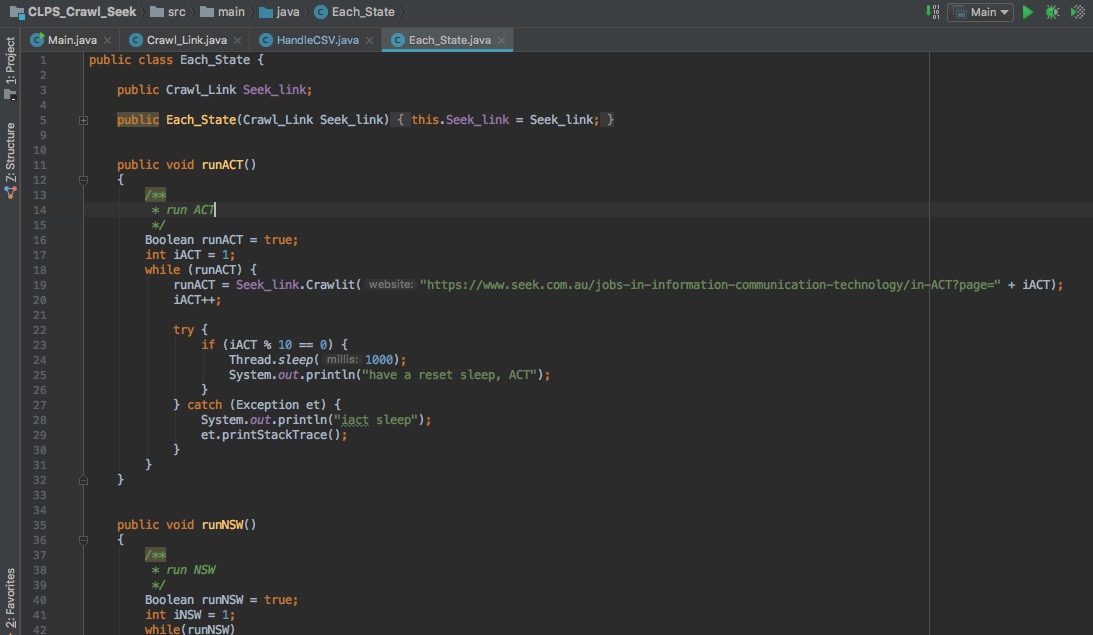


The crawl driver this class uses is phantomjs

The data this class has got will be written to link file

3.2.4 Each\_State

This class is about getting data from different states



The constructor method is

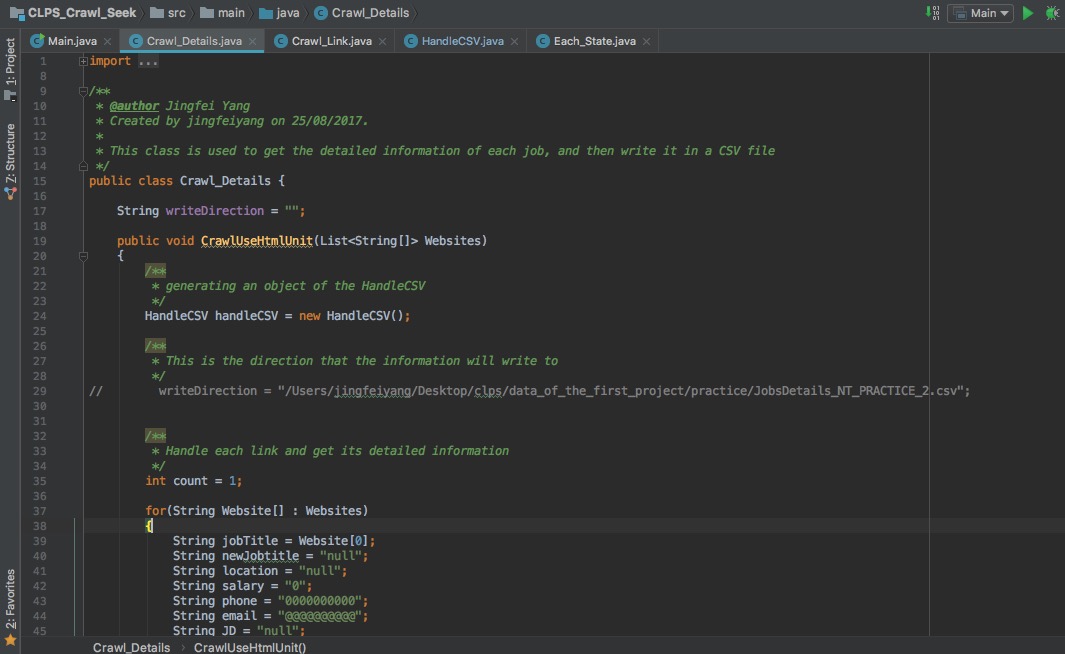
public Each\_State(Crawl\_Link Seek\_link)  
{  
 this.Seek\_link = Seek\_link;  
}

Seek\_Link must be an object of Crawl\_Link

And to avoid making the cpu handleing too much data, the program will sleep 1000 millis per 10 website pages

3.2.5 Crawl\_Deatils

This is the class to catch the details of each job



The crawl driver this class uses is HtmlUnitDriver

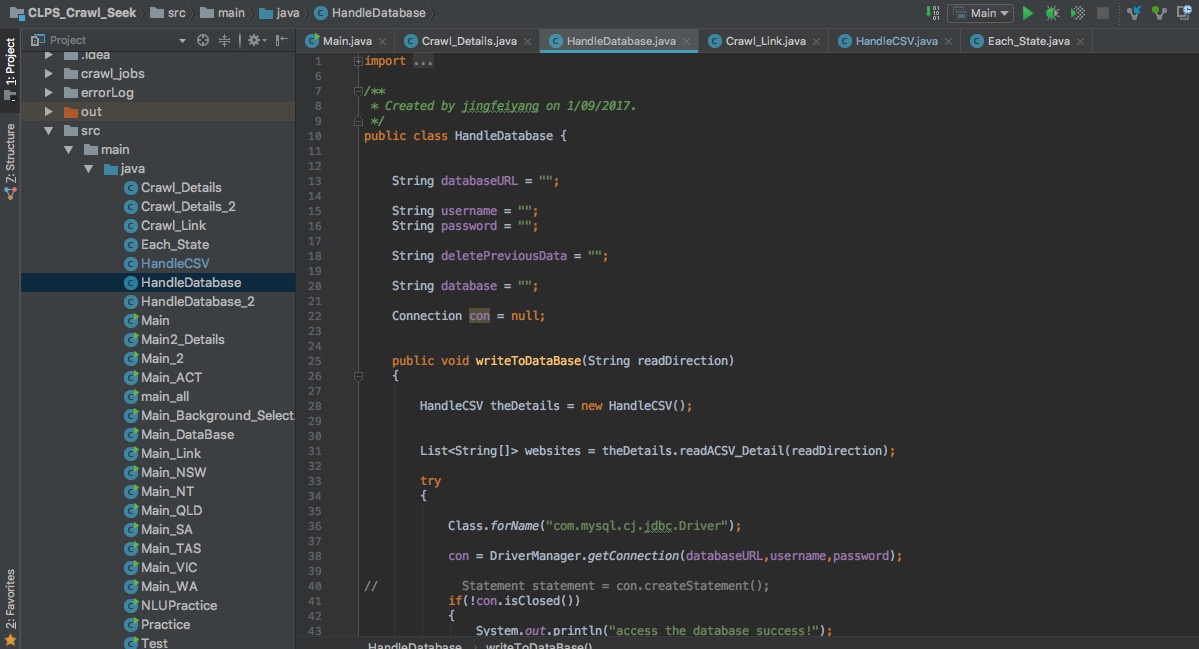
(Using which driver depends on the situation)

The data will be written to Detail file

And after seeking data from every 50 websites, the program will sleep 2000 millis

3.2.6 HandleDatabase

This class will handle the database



The databse driver I use is

com.mysql.cj.jdbc.Driver

This class will create a new table named seekDataDetail in the specific scheme

You can choose to create the scheme and table first or not, this class will check them, if they are not created, this class will create them automatically

3.3 the process of the program

a. read the TheBasicDetails.csv, get the configuration data

b. get job titles and websites from specific states based on the command line argument

c. write the link information to link file

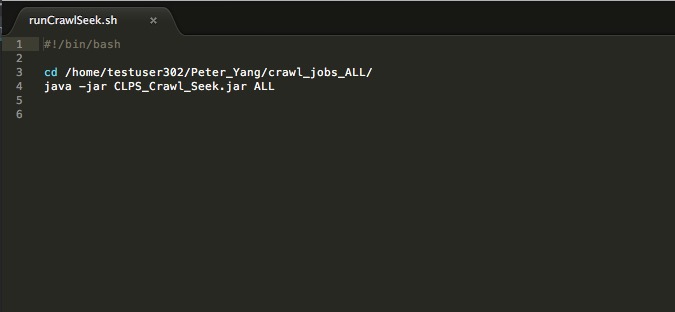
d. read link information from link file and get job details from the specific websites and write the details to Detail file

f. read the details information from the Detail file and write it to databse

4. The server script and command

4.1 runCrawlSeek.sh

This script is used to run the java file in ubantu server



4.2 crontab command

This command is used to let the script run at 3 pm every day and no output file.

0 15 \* \* \* /home/testuser302/Peter\_Yang/runCrawlSeek.sh