ReadME

Description of used software

We use the crawler we used in project 1 which is crawler4j. It is an open source web crawler for Java which provides a simple interface for crawling the web pages.

Installation

NOTE:

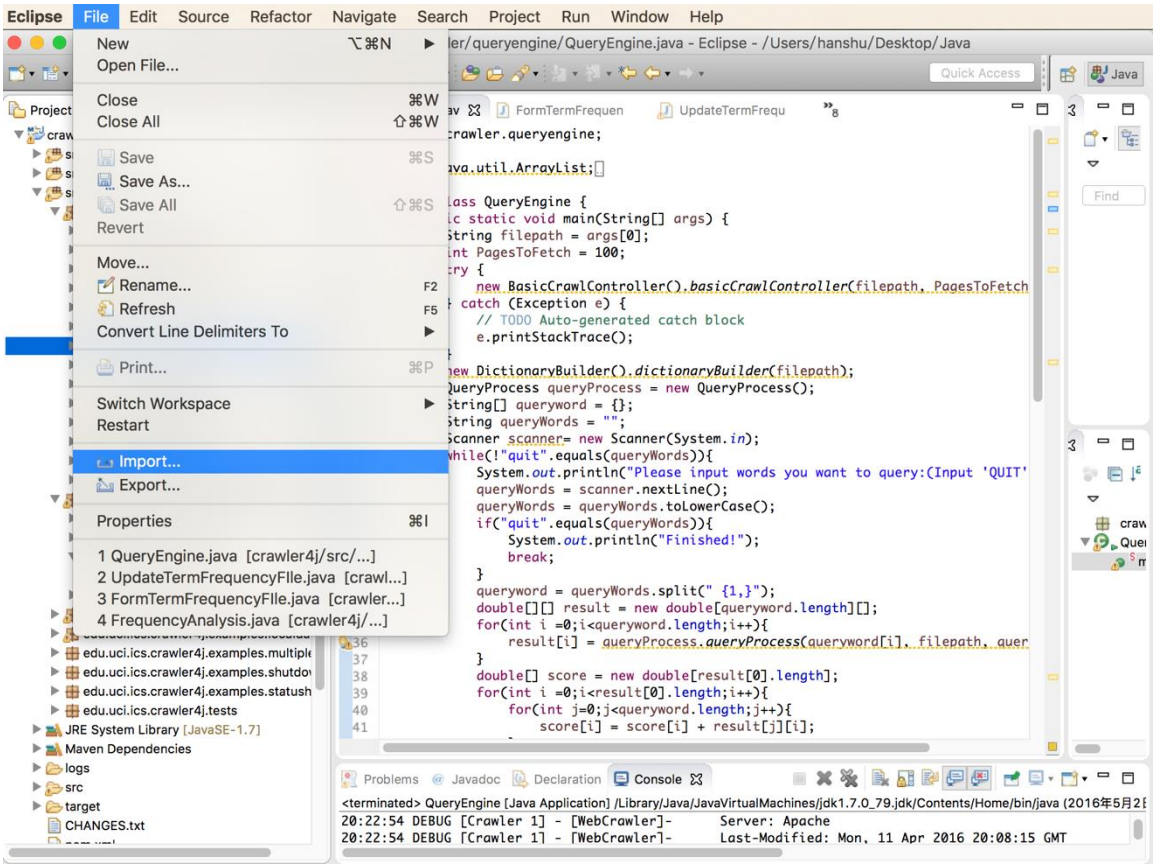
Using Maven

To use the latest release of crawler4j, please add the following snippet in the "pom.xml' file which is contained in "crawler4j-crawler4j-4.1" project

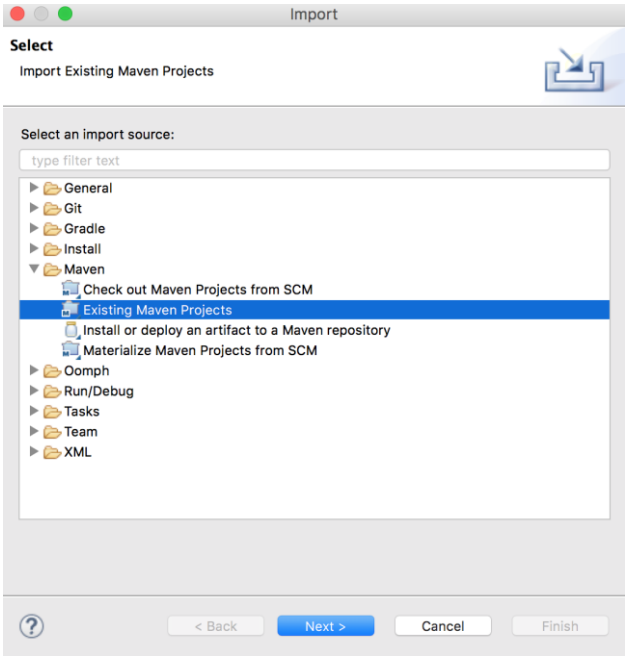
```
<dependency>
    <groupId>edu.uci.ics</groupId>
    <artifactId>crawler4j</artifactId>
    <version>4.2</version>
</dependency>
```

Step:

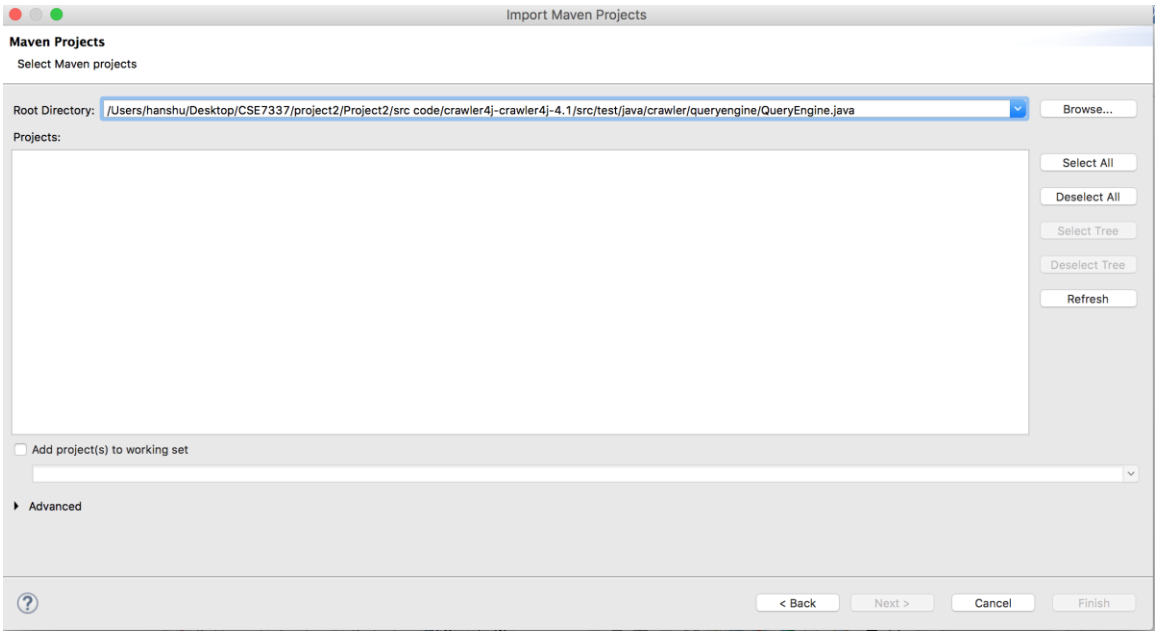
1. open Java IDE(Eclipse)--->click on "file" button--->select "Import" tab as show below:



2. select the “Existing Maven Projects” option under the Maven folder in the Import interface as show below:



3. select the “crawler4j-crawler4j-4.1” file as your “Existing Maven Projects”--->click on “next” which equals to “OK” as show below:



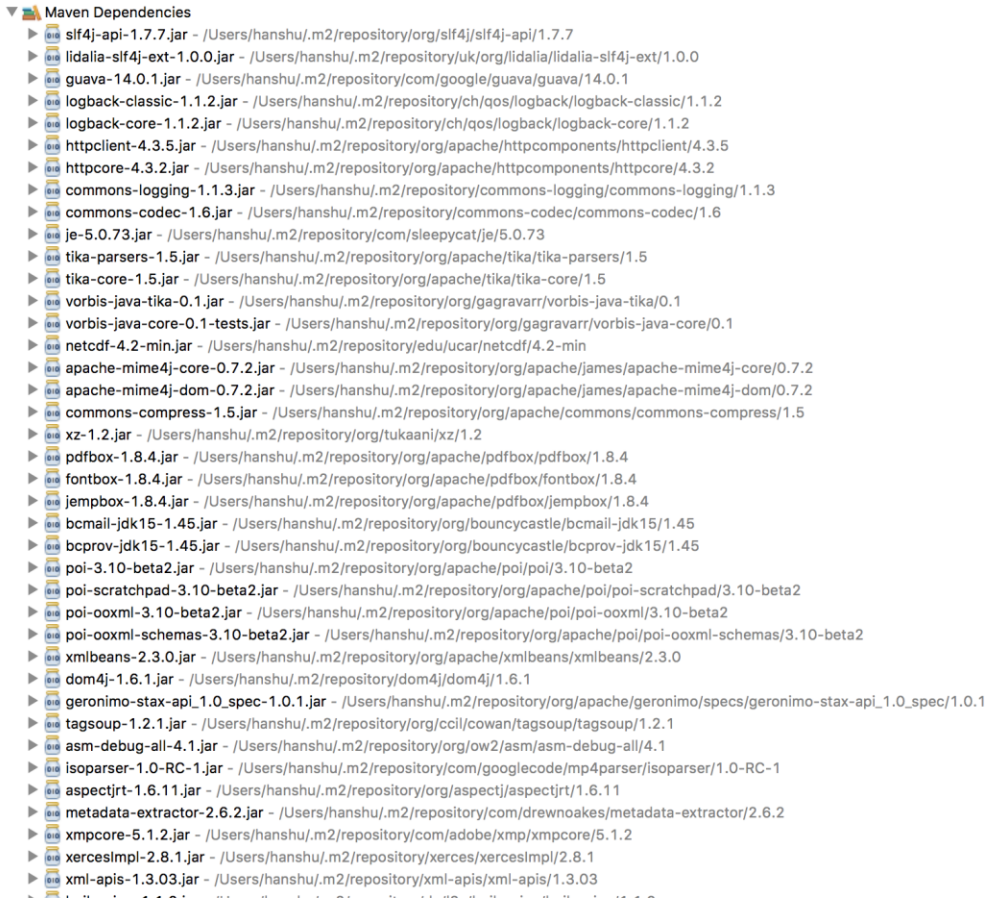
Then, set the right path.

4. as the notation content before, add the snippet

```
<dependency>
            <groupId>edu.uci.ics</groupId>
            <artifactId>crawler4j</artifactId>
            <version>4.2</version>
    </dependency>
```

into the pom.xml file, then the import process started.

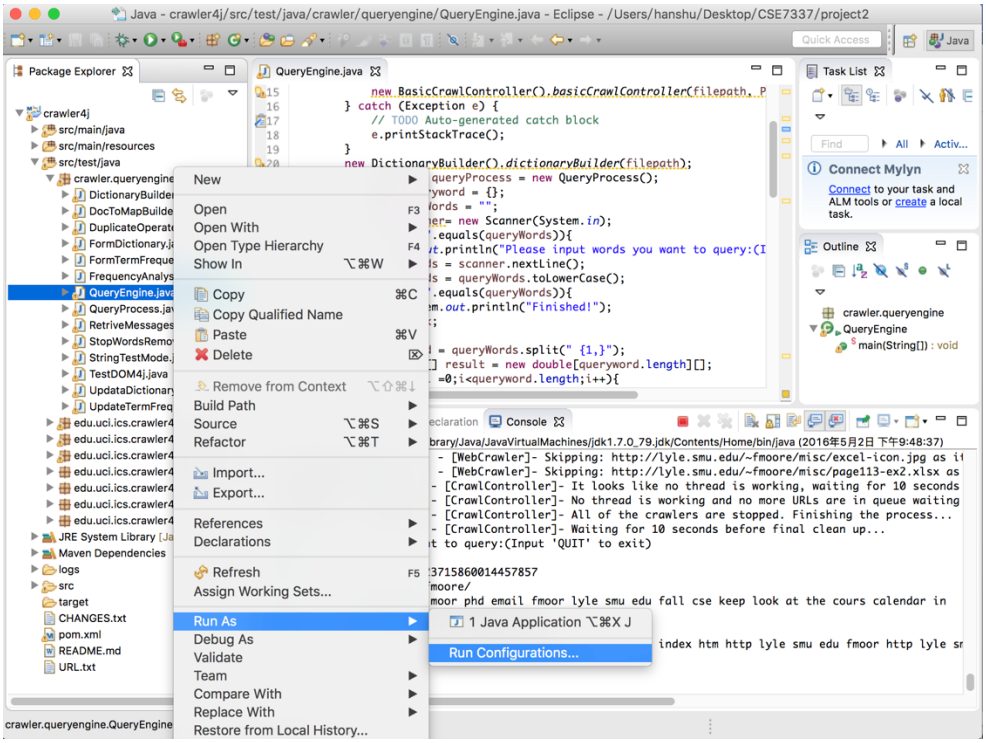
As we have added the dependency snippet and dependency package has already existed, therefore we cannot take a screen shot of the process, but the screen shot of the dependency files are listed below:



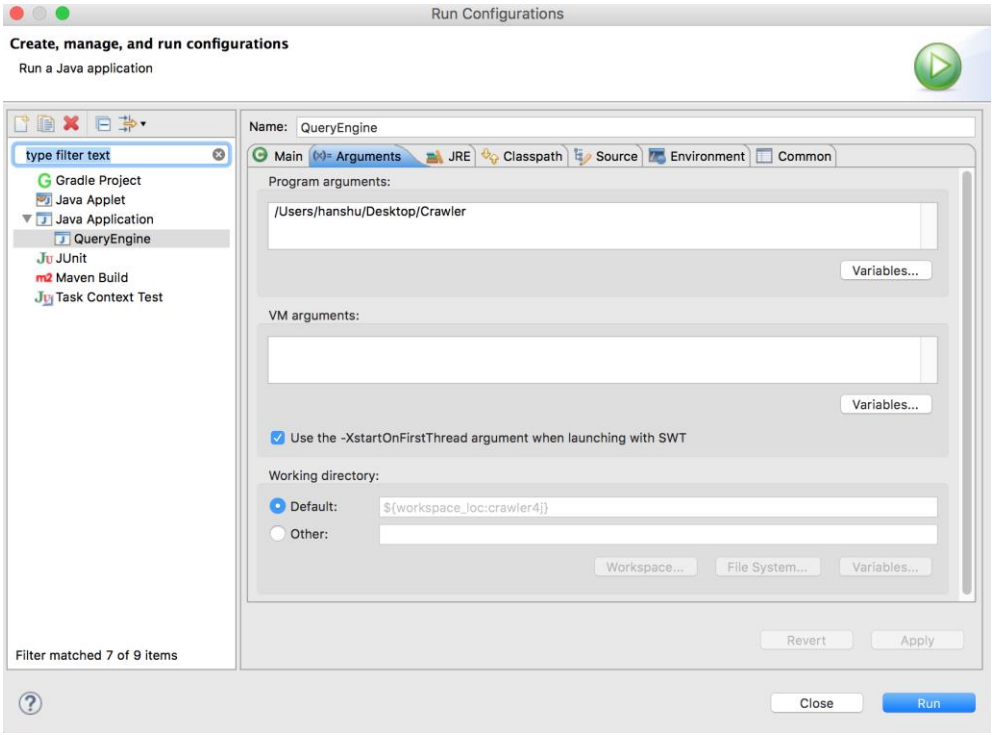
5. Completed!

Execution instructions

1. create a file to store the crawler result and create a stopwords.txt file under the file newly created to store the crawler result--->add the stop words as provided--->you have to run the java file “QueryEngine.java” as configuration



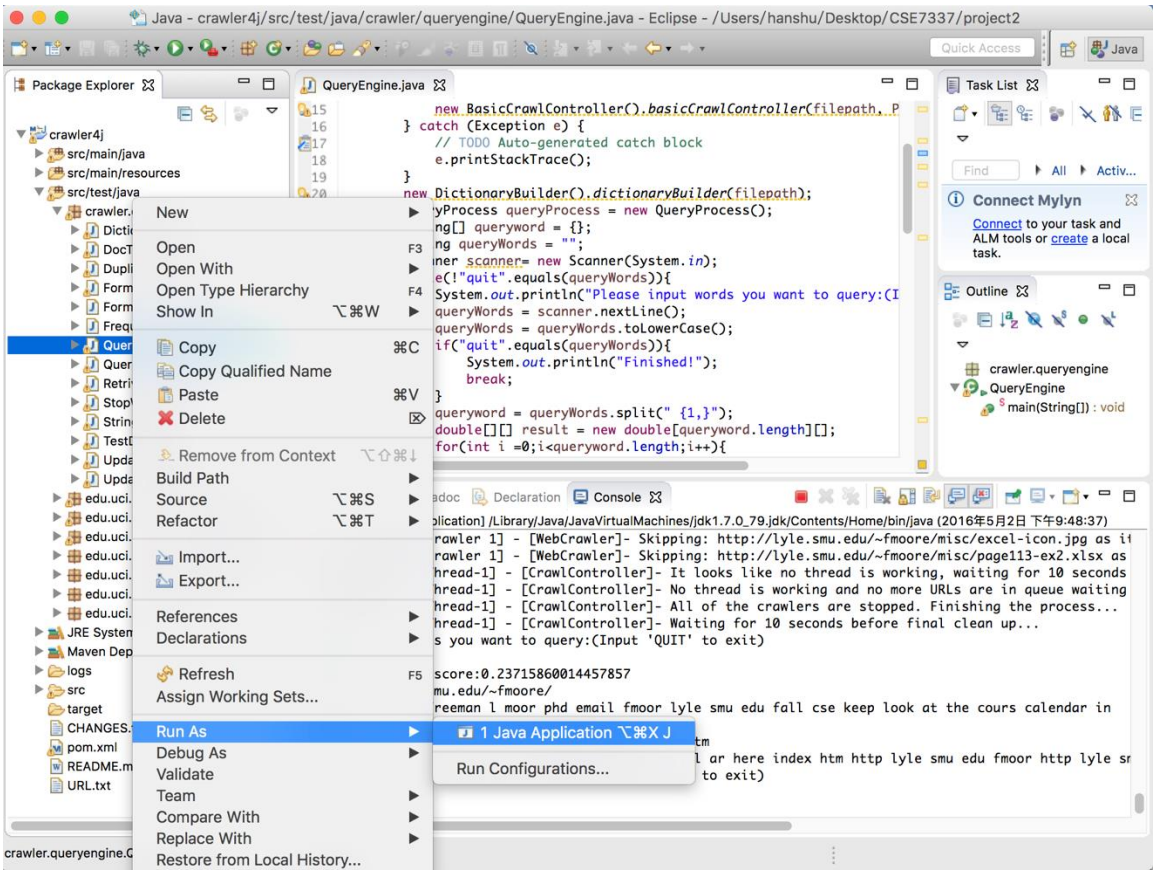
2. type in the Arguments---> press “run” button



set the right path to the folder where contains a stopwords.txt file, and it will be

used to save other result files.

3. run the QueryEngine.java as Java Application



4. type in the query words in the reminded district and the engine will start work

```
21:49:59 INFO  [Thread-1] - [CrawlController]- It looks like no thread is working, waiting for 10 seconds to make sure...
21:50:09 INFO  [Thread-1] - [CrawlController]- No thread is working and no more URLs are in queue waiting for another 10 seconds to make sure...
21:50:19 INFO  [Thread-1] - [CrawlController]- All of the crawlers are stopped. Finishing the process...
21:50:19 INFO  [Thread-1] - [CrawlController]- Waiting for 10 seconds before final clean up...
Please input words you want to query:(Input 'QUIT' to exit)
moore smu
docID:1-----rank score:0.23715860014457857
URL:http://lyle.smu.edu/~fmoore/
Content: spring freeman l moor phd email fmoor lyle smu edu fall cse keep look at the cours calendar in
docID:27-----rank score:0.15811388300841894
URL:http://lyle.smu.edu/~fmoore/misc/urlexample1.htm
Content: exampl link februari how mani distinct url ar here index htm http lyle smu edu fmoor http lyle smu
Please input words you want to query:(Input 'QUIT' to exit)
```