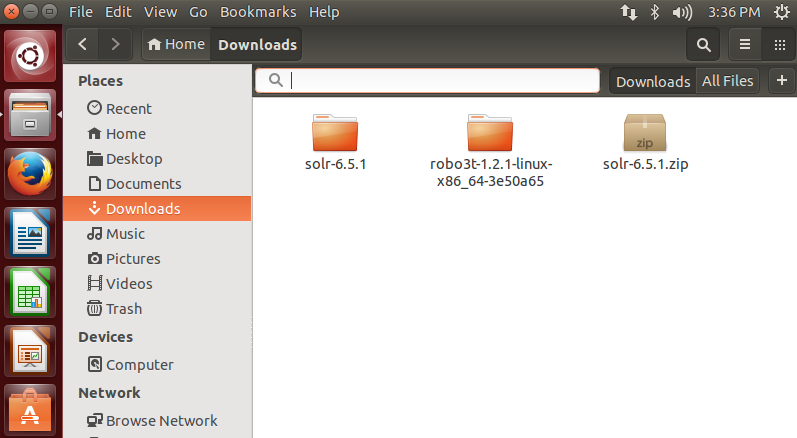
**INFS7410 Report**

The report is in the aiming of explaining how to deal with the web page crawl and use Lucene and Solr to recognize the searching engine.

1. Crawl web page with MongoDB

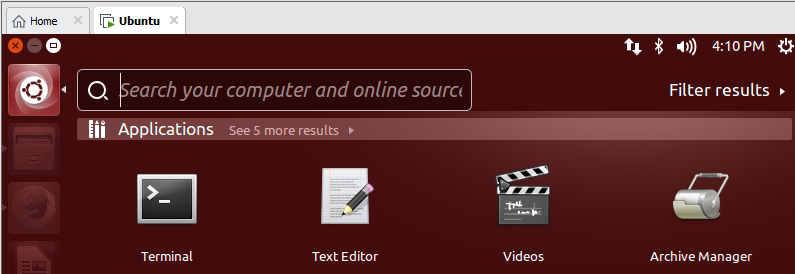
According to the introduction of the software installation, the following is the process of the crawling process, generate index, and crawl web page.

1. For the first step, install the Ubuntu 14.04 on Virtual Machine as the instruction on tutorial 1; install the software robo3t to view the download web page results.



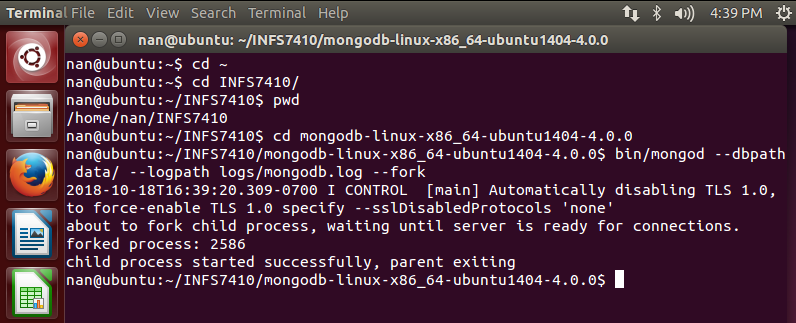
1. For the second step, install Nutch and run it to download webpage.

a. Go to the terminal interface

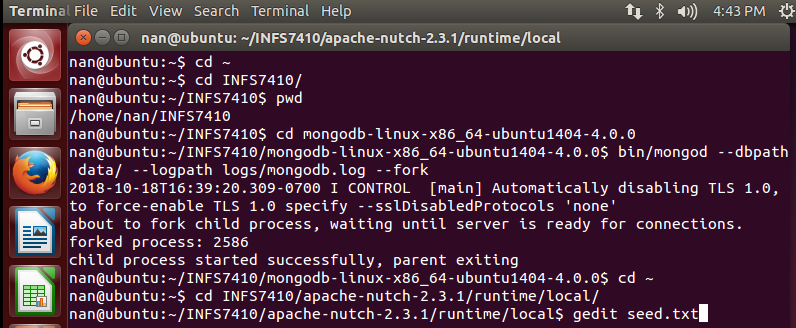


b. Start mongodb when restart Ubuntu

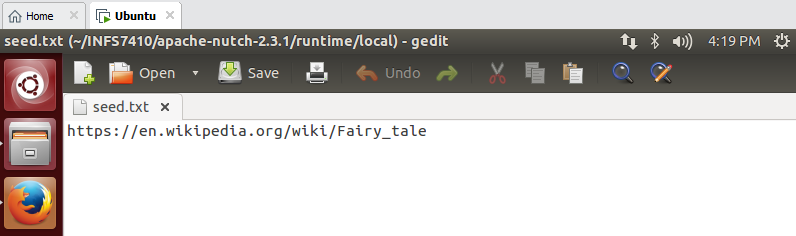
Mongodb is started successfully, by receiving the message that child process started successfully.



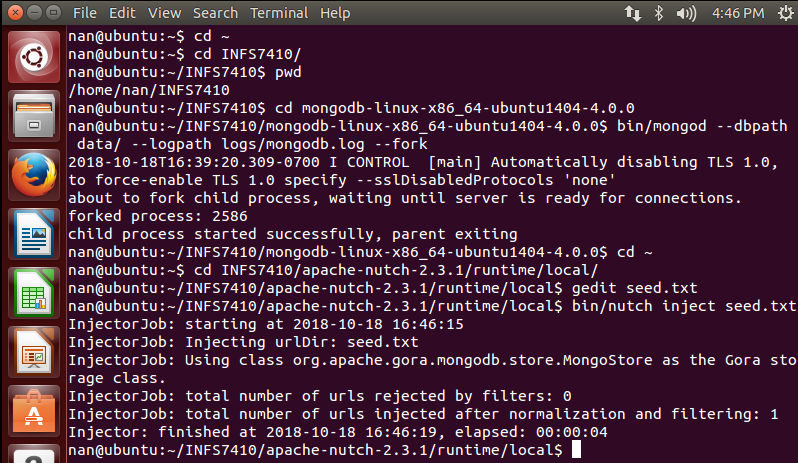
c. Input the url need to crawl into the file ‘seed.txt’



The website try to download is ‘https://en.wikipedia.org/wiki/Fairy\_tale’.

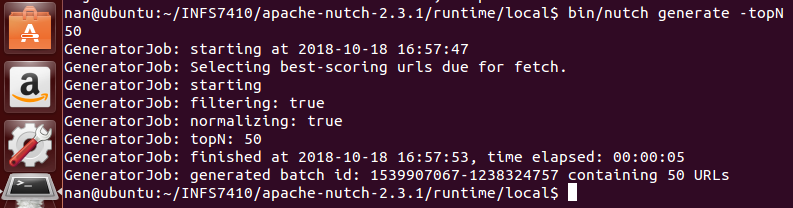


d. Inject the seed

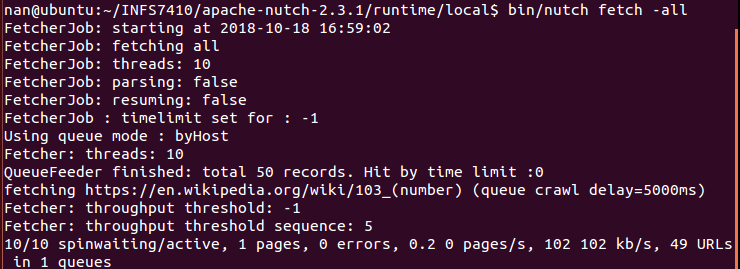


e. In order to analytic the web page and the sub-webpage of the link input, it needs repeat the following four steps many times. Usually, the first time is used to generate the main web page, without analytic most of the sub-webpage, and the sub-webpage is generate out since the second round.

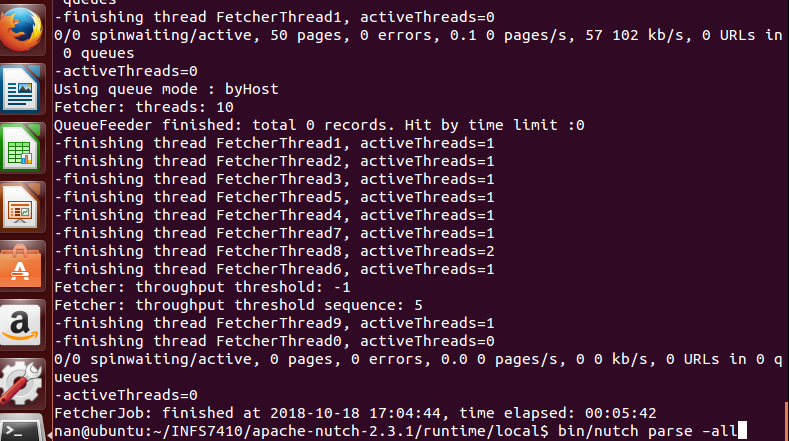
* Generate: bin/nutch generate -topN 50



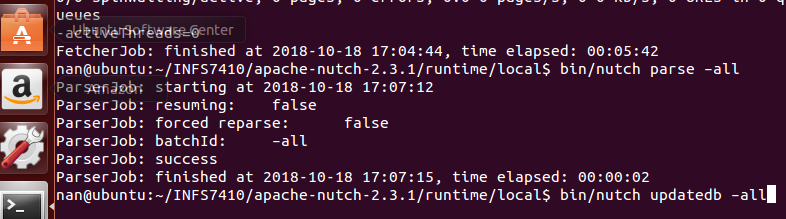
* Fetch : bin/nutch fetch –all



* Parse : bin/nutch parse –all



* Updatedb : bin/nutch updatedb –all

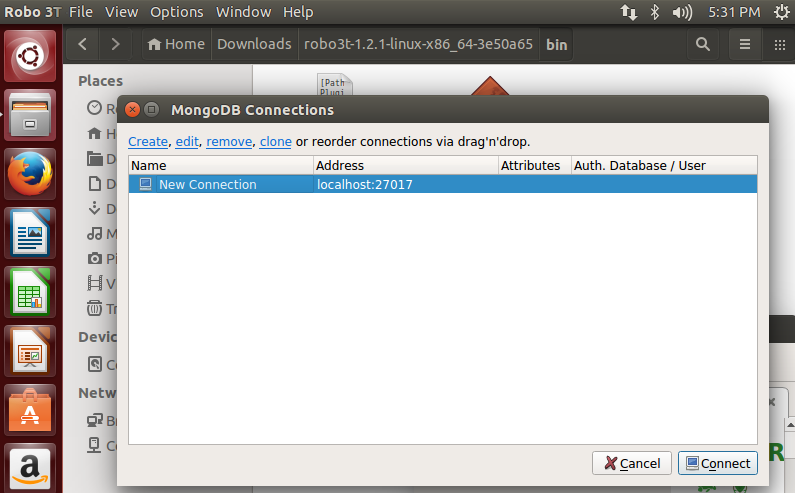


1. View the download result

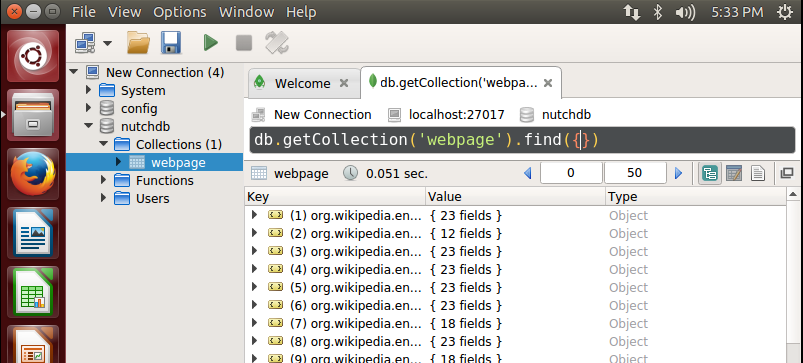
a. Go to the location of robo3t and start the program.



b. Connect the robo3t with MongoDB.

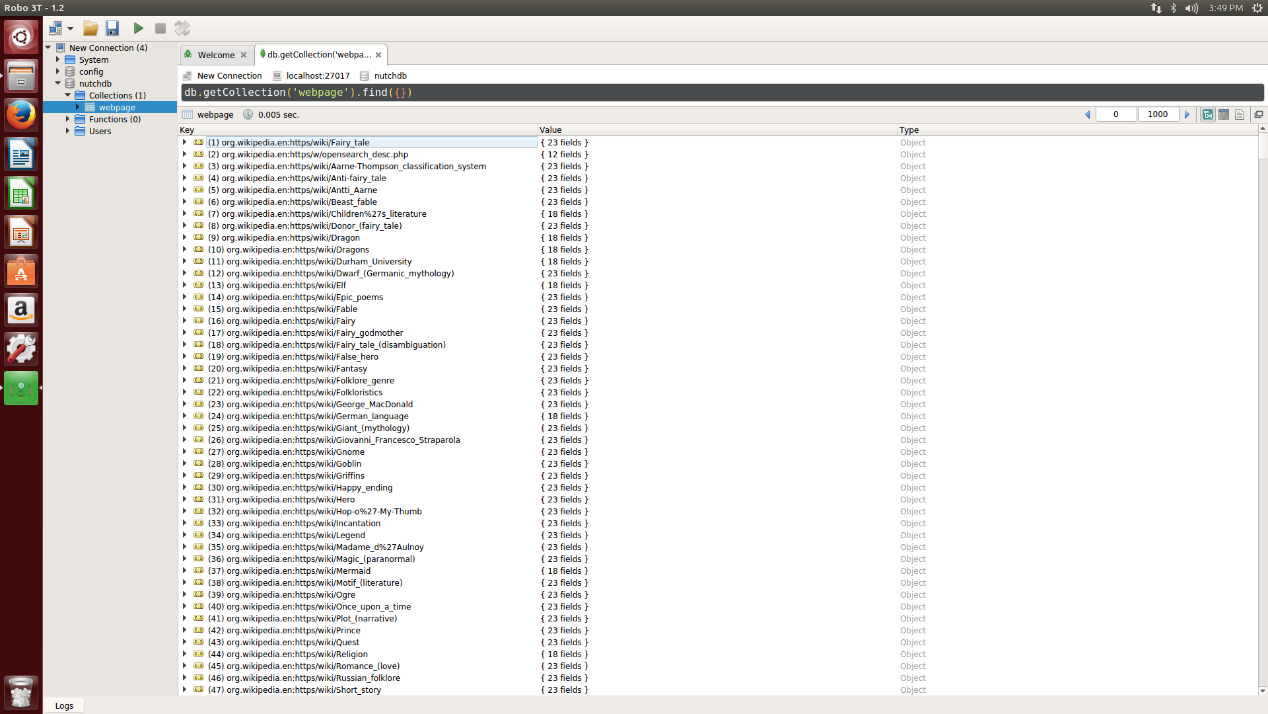


c. Connect the robo3t with MongoDB, and the result could be view on the page by path nutchdb -> collections -> webpage.

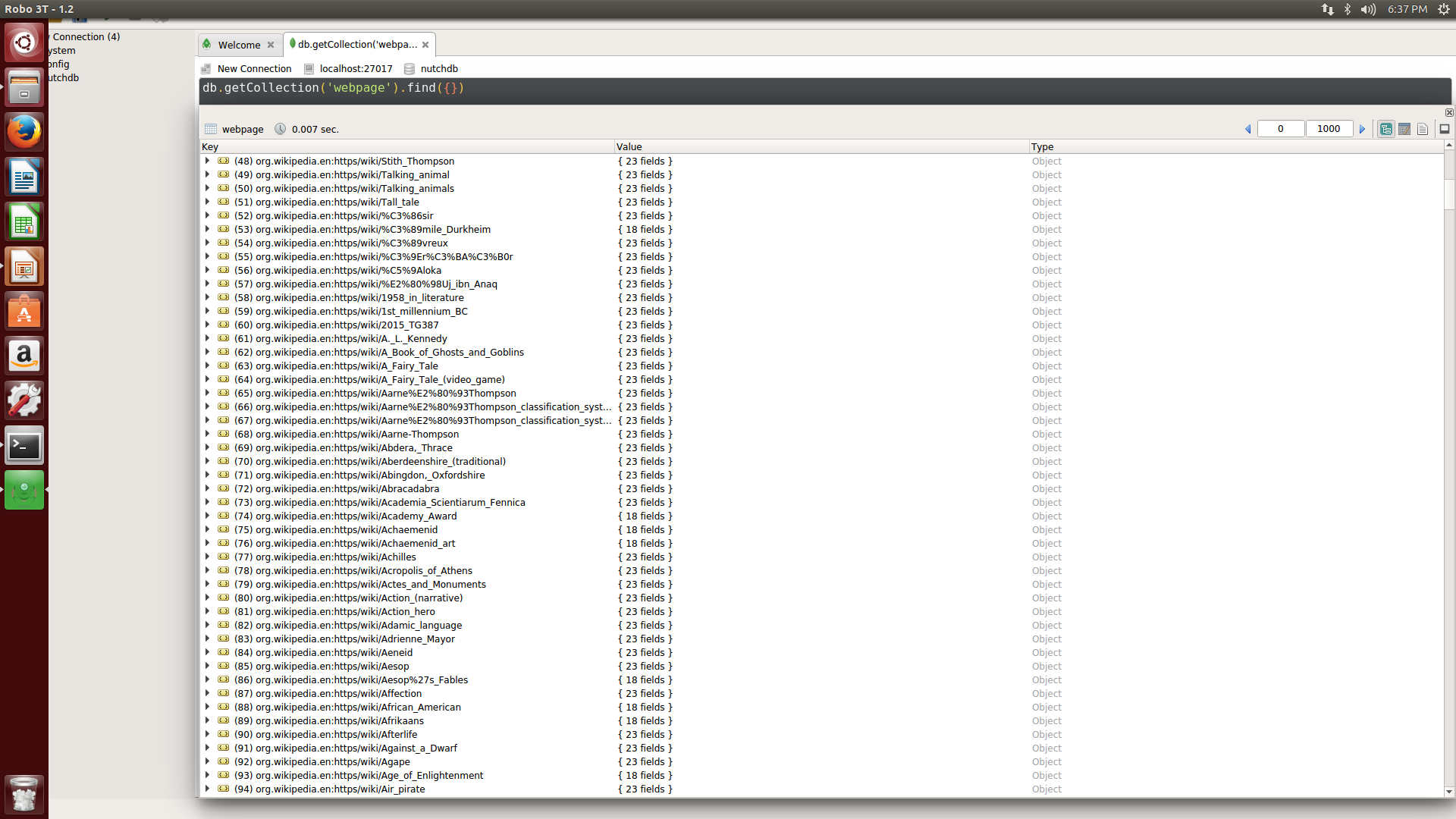


1. Result screenshot

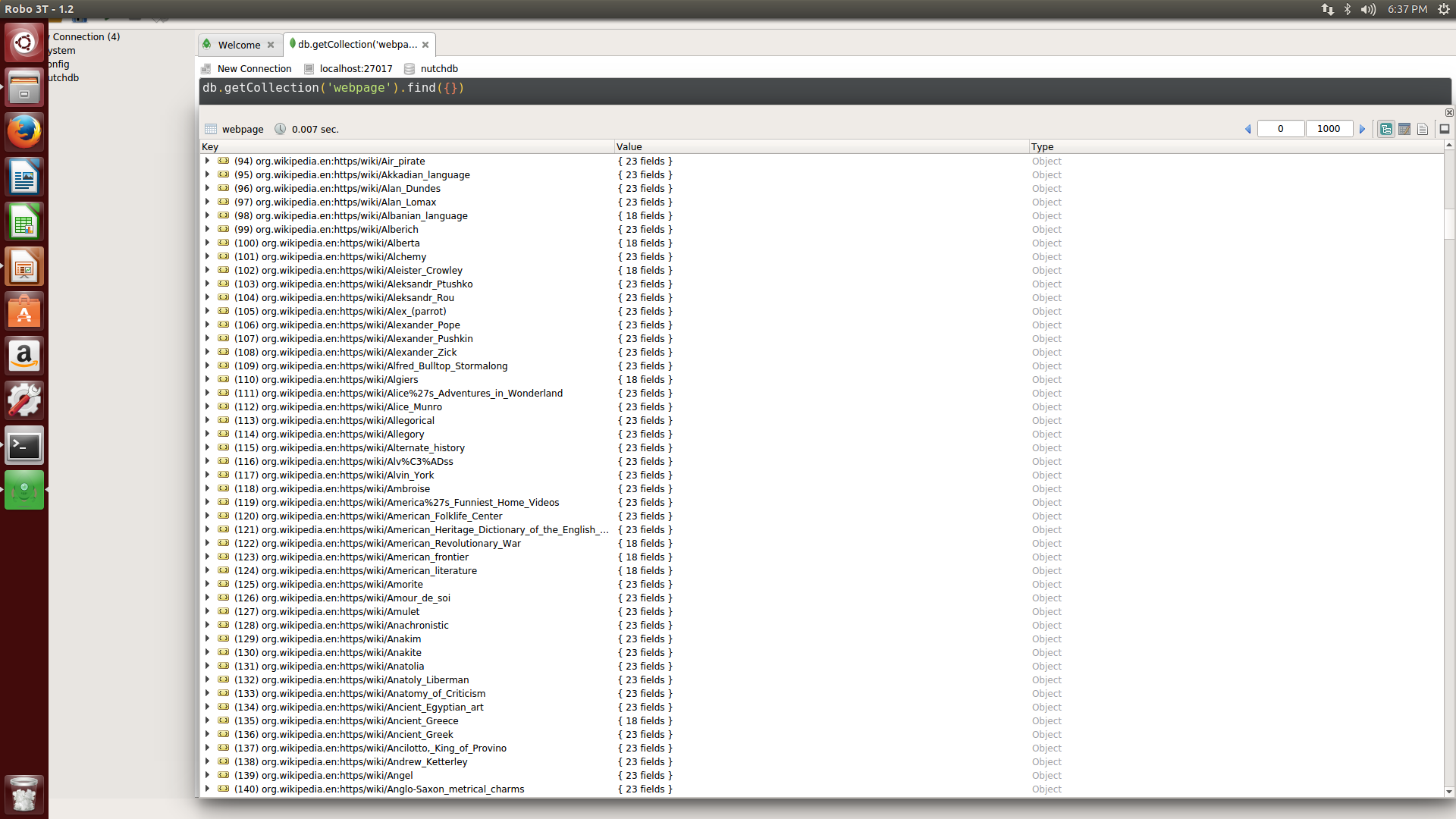
Graph 1



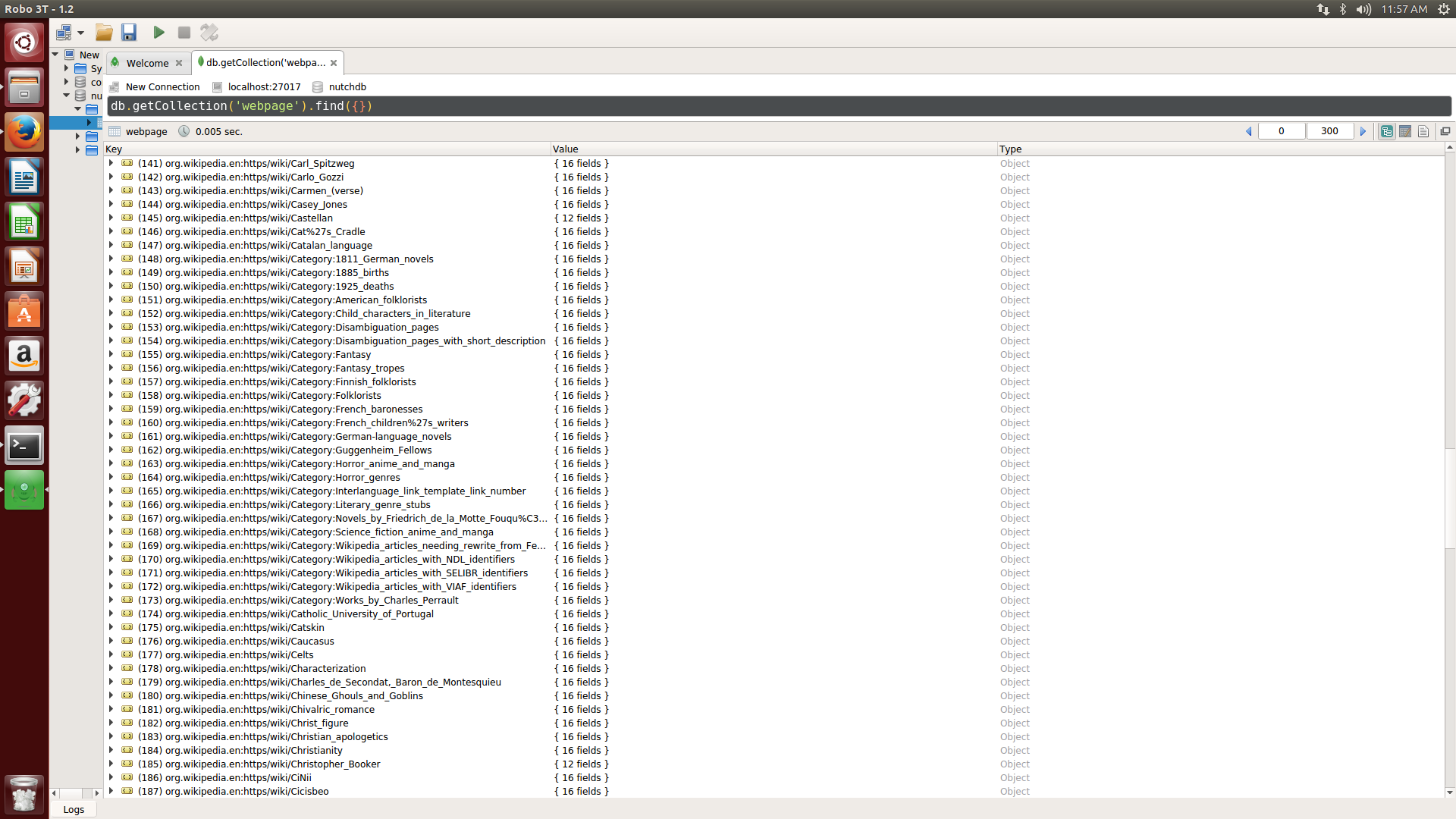
Graph 2



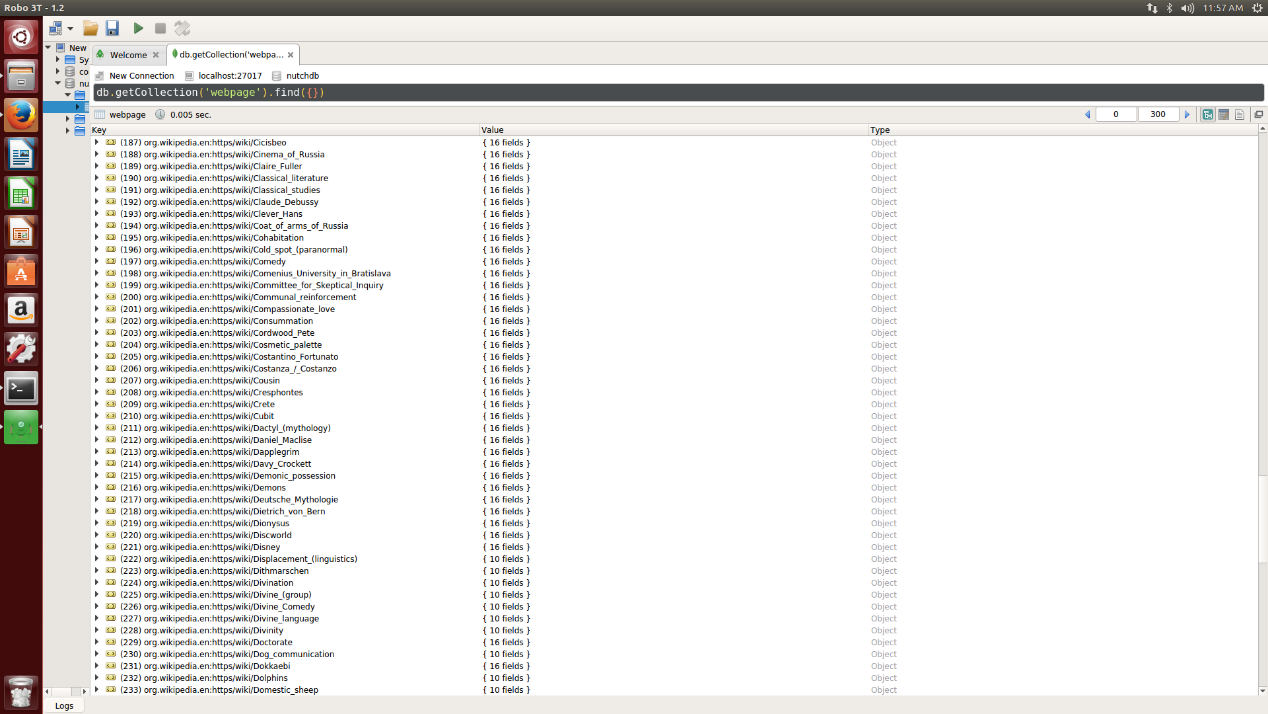
Graph 3



Graph 4



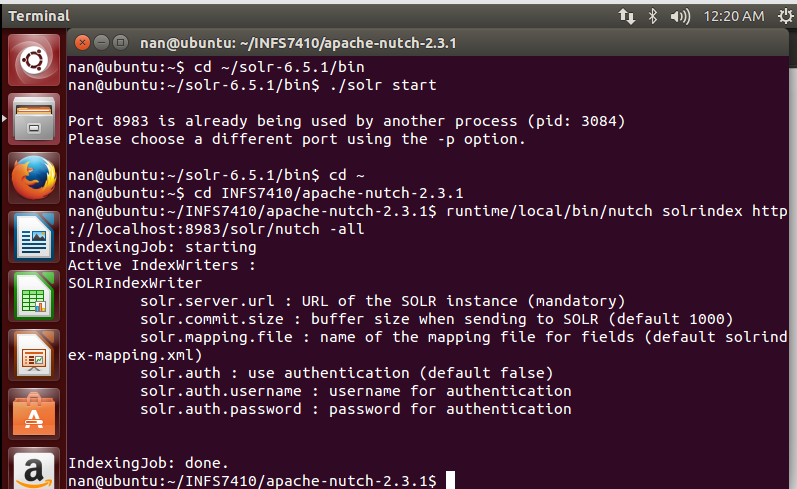
Graph 5



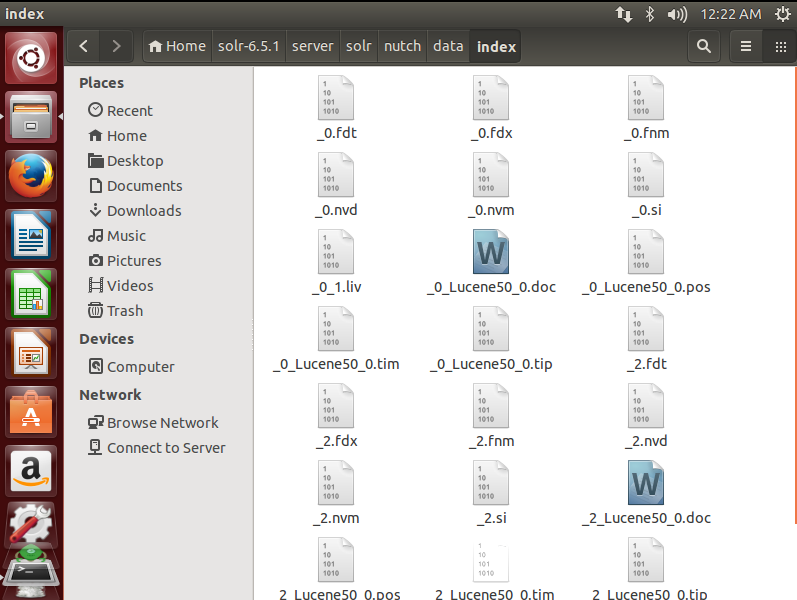
2. Use Lucene and Solr to implement the search engine.

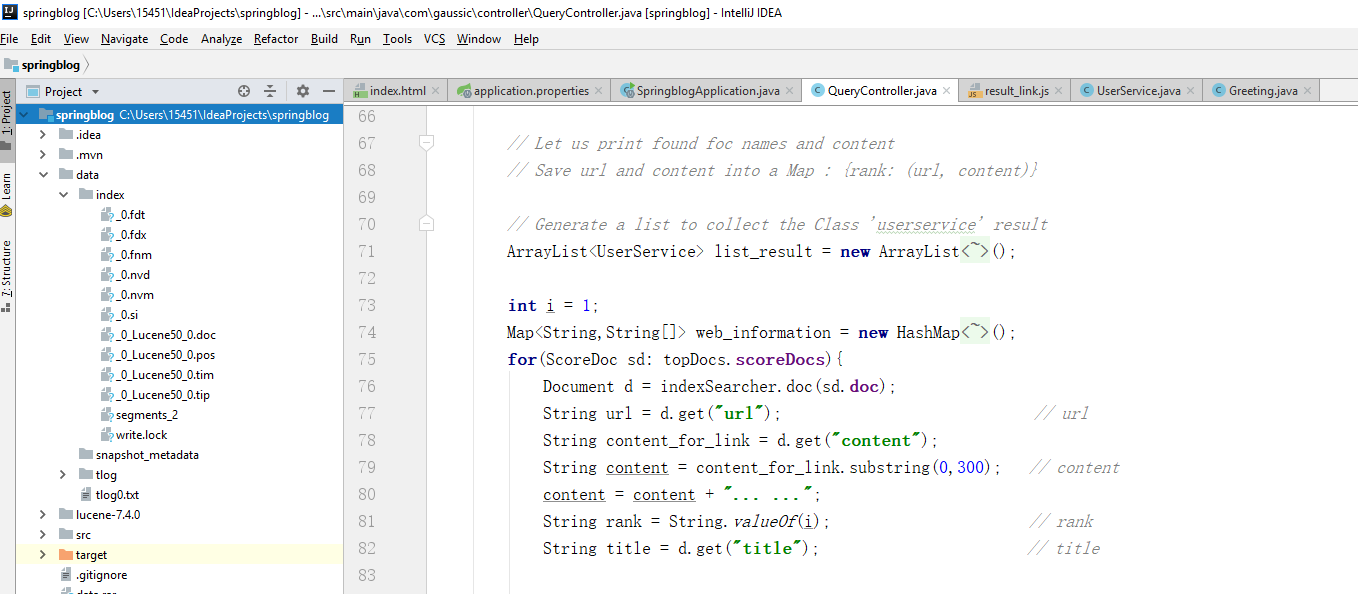
The Solr software is applied to generate out the index, which has the effect to search out some document objects, which contain various documents information, such as Url, document content, and title. Lucene package is used to generate out, obtain, search and deal with document content.

1) Restart the Mongodb same as the part one. Install Solr software



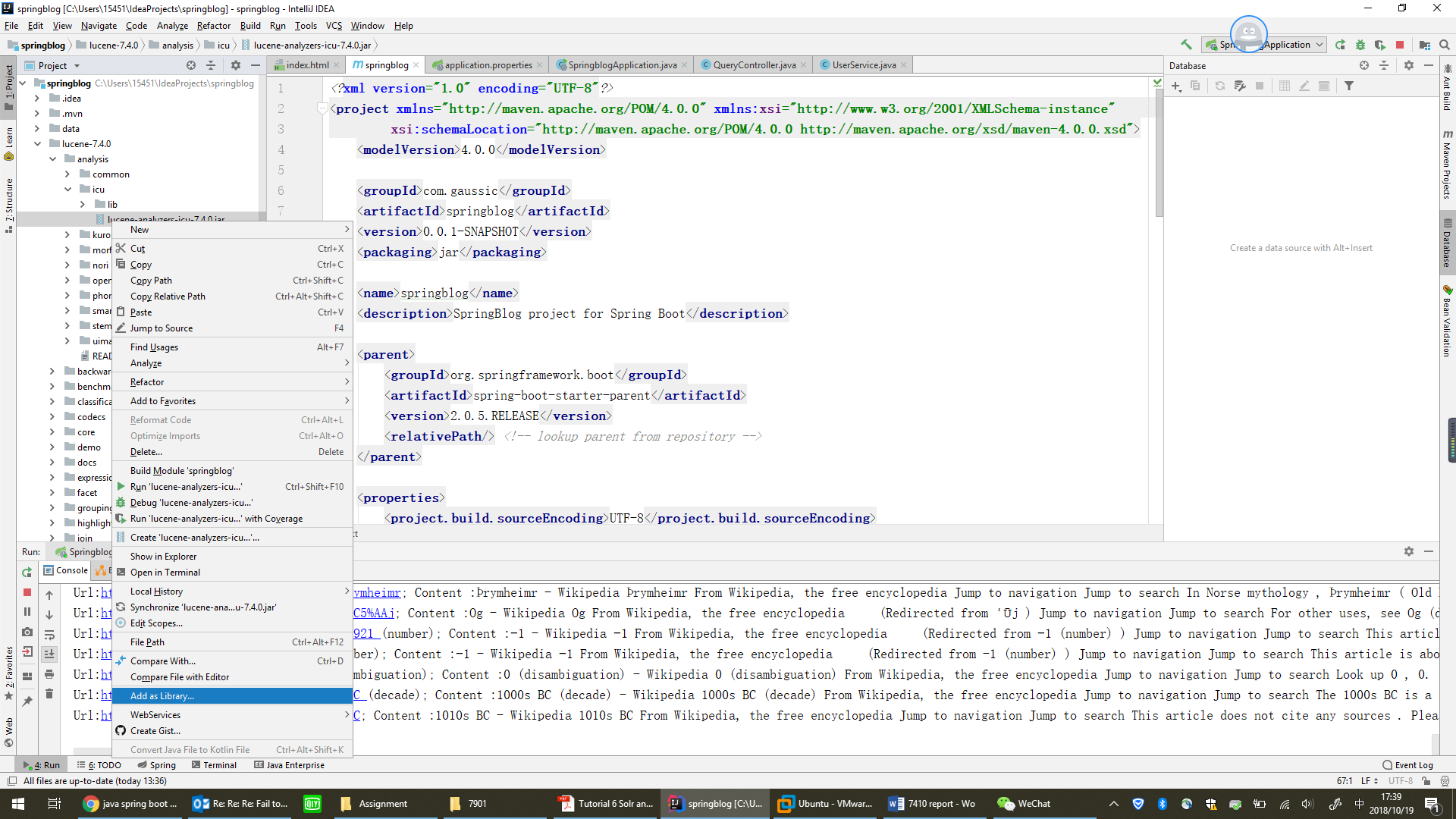
2) The Solr index exists under the path solr-6.5.1/server/solr/nutch/data/index



3) Set up a Spring Boot project in Java and add the index data of this project.

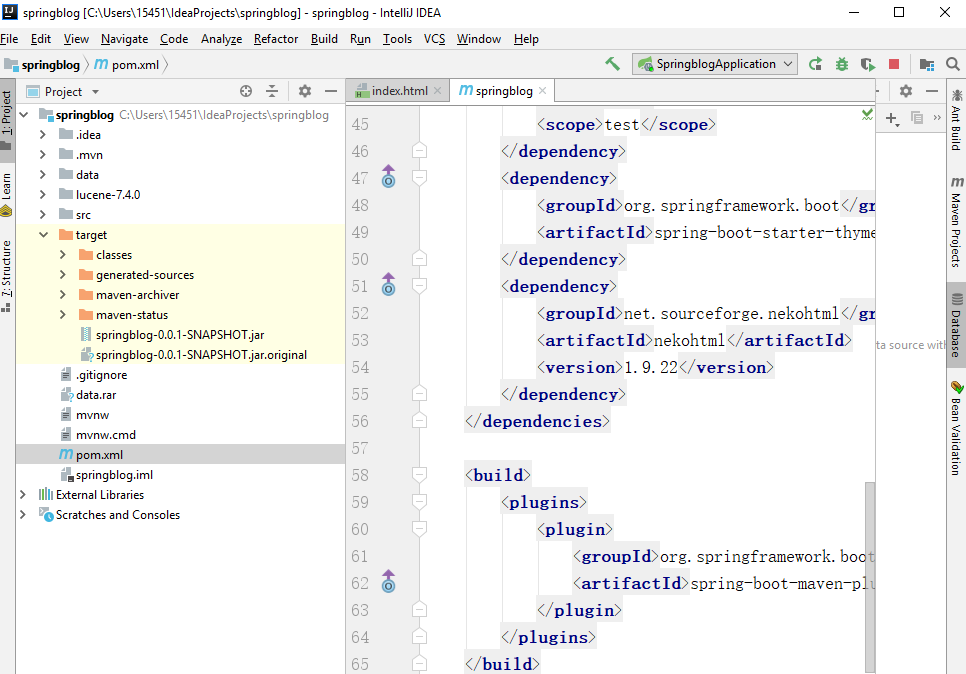
4) Download lucene-7.4.0 from the following website and add library for the four library applied by this process.

<http://archive.apache.org/dist/lucene/java/7.4.0/>

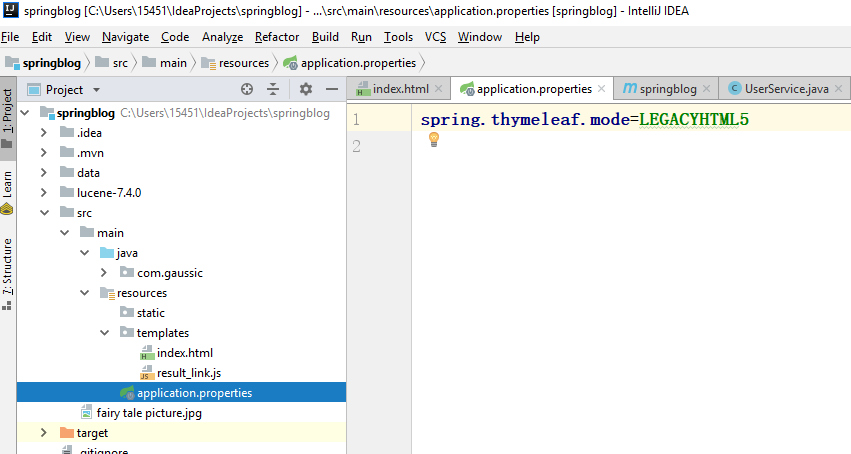


5) Run the Maven Projects and the maven snapshot is generated out under ‘target’.

Besides, the project inserts the xml file pom.xml with various dependencies.

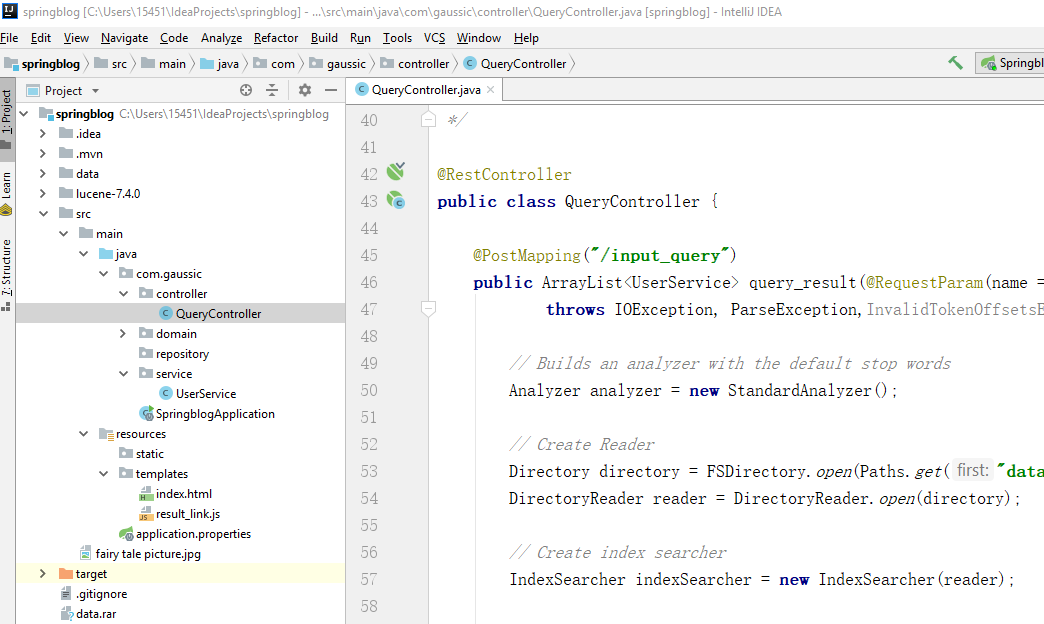


6) Set the thymeleaf environmental in the ‘application.properties’ as below.



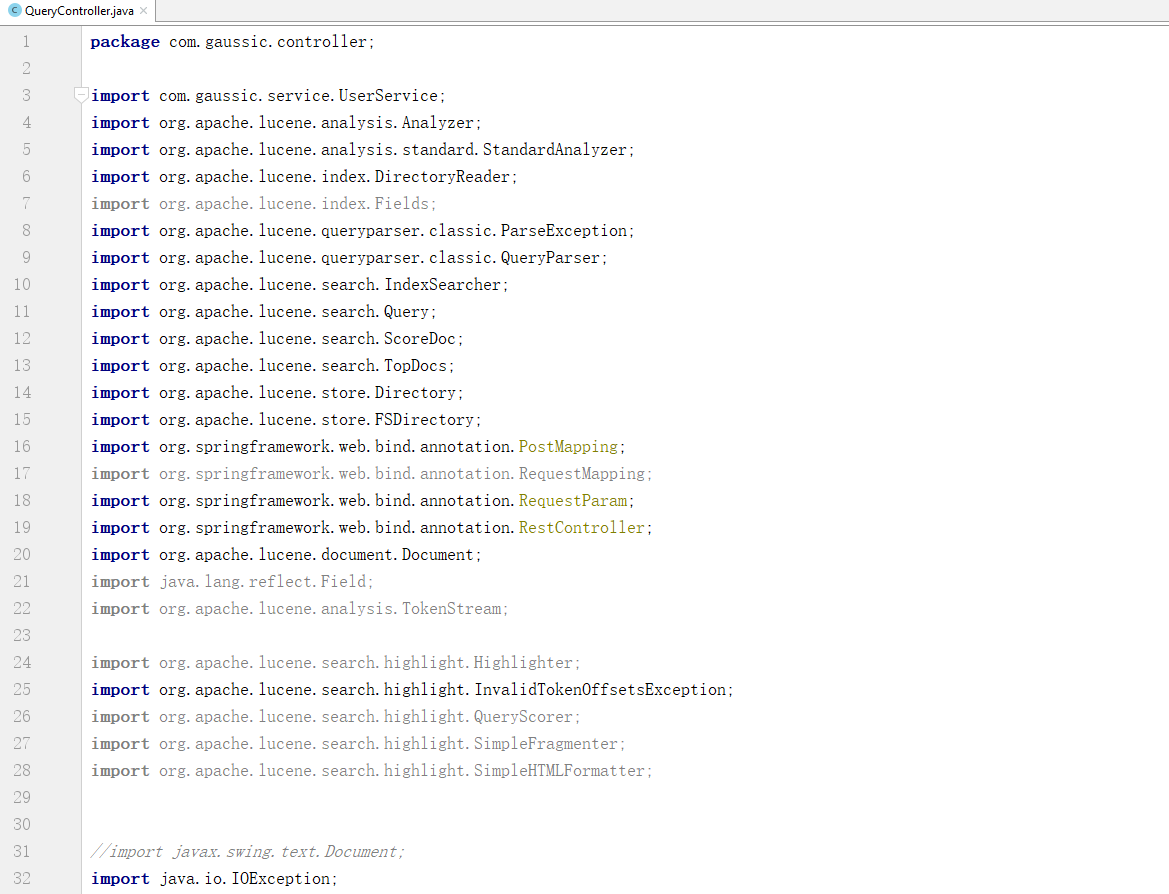
7) Main code part

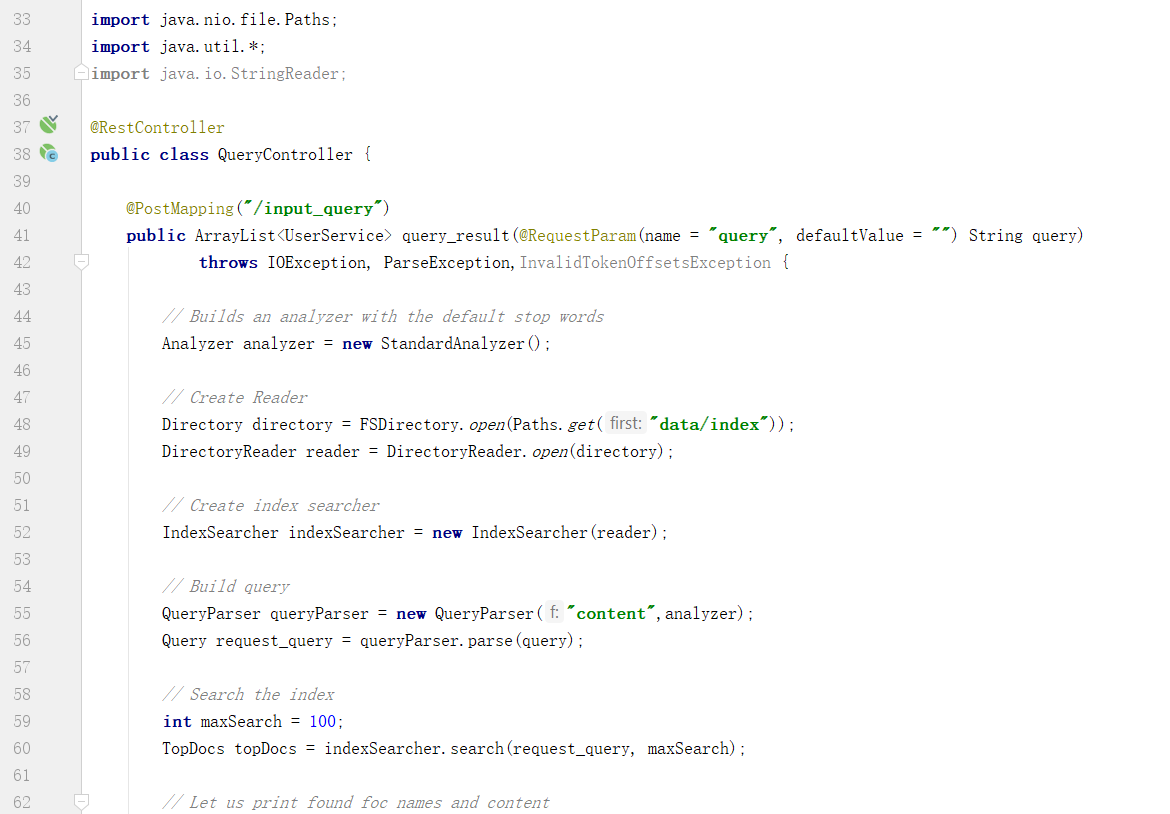
There is 3 main code parts involved in the searing engine. Firstly, the first one is ‘QueryController’, which is used to return the query result as a Document Class list to the website according to the query; the second one is ‘UserService’, which is used to define the Document Class as ‘UserService’, which including the attribute of the class; the third one is ‘index.html’ it is the front service to generate out a web page for client searching and result exhibition.



a. QueryController

By the Lucene package, a great amount of functions could be used in the process. The process of query control is as below: firstly, the dictionary reader reads index data from raw file; secondly, obtain the query from the website; thirdly, generate an index searcher to obtain the documents class from raw file by query, and save the documents class to topDocs; fourthly, adopt a loop to fetch the related information from the documents class, such as url, title, score, content, and save them into a new class named UserService. Finally, all the documents information will be returns to the website in the format of UserService List.



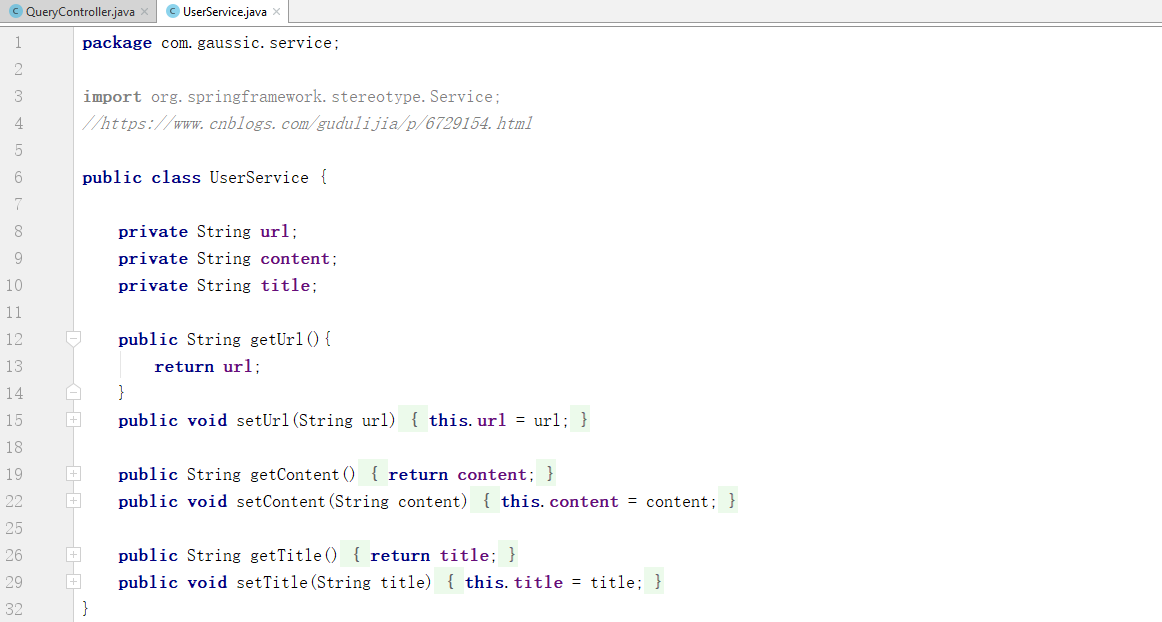






b. UserService

The user service part is used to define the class, it include three attributes, including url, content, and title;

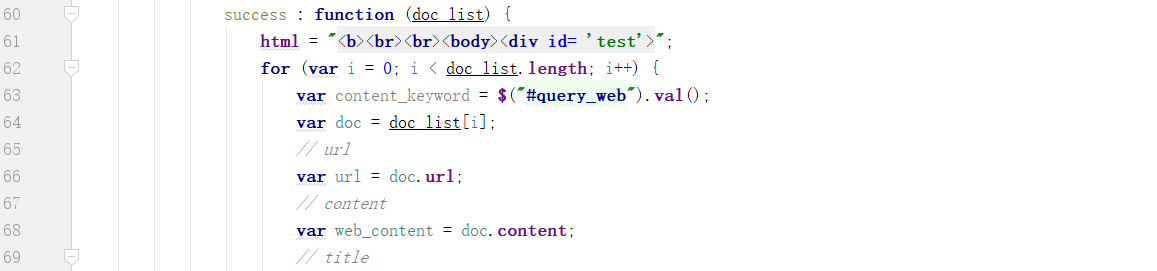


c. website (html)

The following code is used to design the web interface. After input the query, the ajax method send the query to the java code by a special url ‘/input\_query’, the special java function in QueryController accept the query as a parameter by the identification ‘@PostMapping("/input\_query")’. After that, the process sends java function running result to the website, and shown on a part of the web, with part id as “#content”.

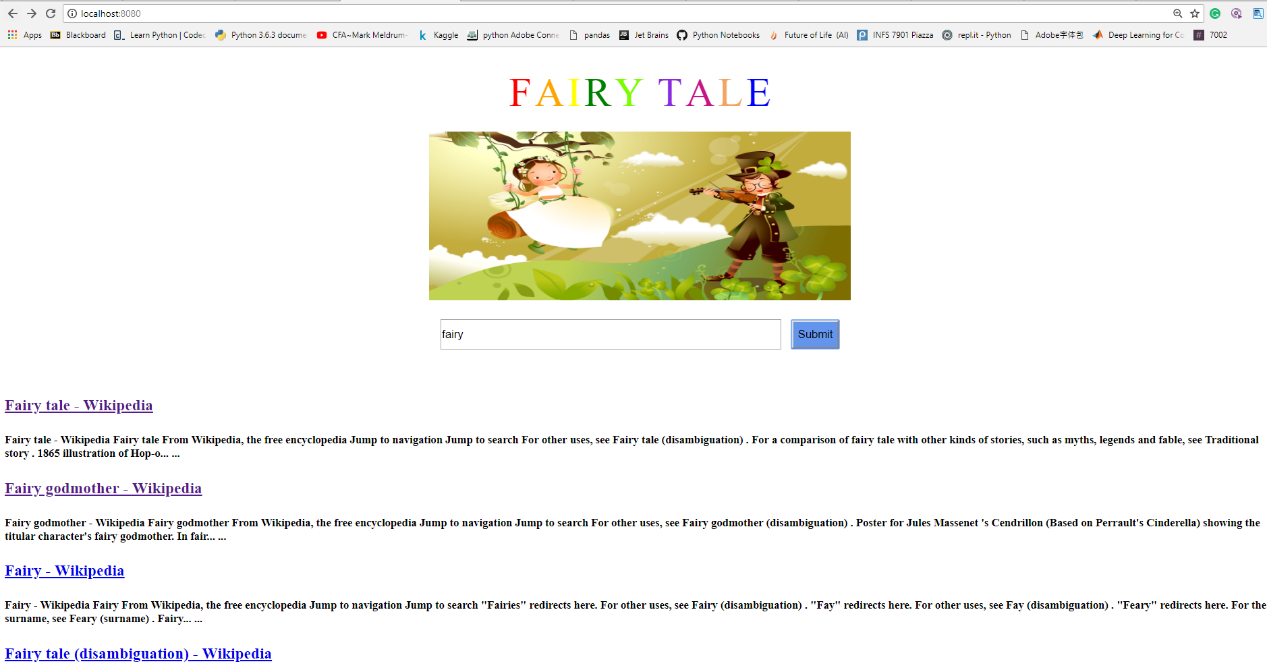








8) Interface



In conclusion, this process used Spring Boot to realize the link between web page and java code, and used Lucene and Solr to realize the simple interface search engine.