JavaCrawler

1. Design of the JavaCrawler

The principle of JavaCrawler is easy. Start from a Seed URL, connect that website, check whether there is a Keyword, and continue search by some key links. For example, search “<http://cc.gatech.edu>” with keyword “research” and keylink “gatech.edu”, if the website contains the keyword “research” it will continue downloading the website and import all the links into database.

I used Jsoup library to phrase the html file, which helps me a lot and save a lot of time.

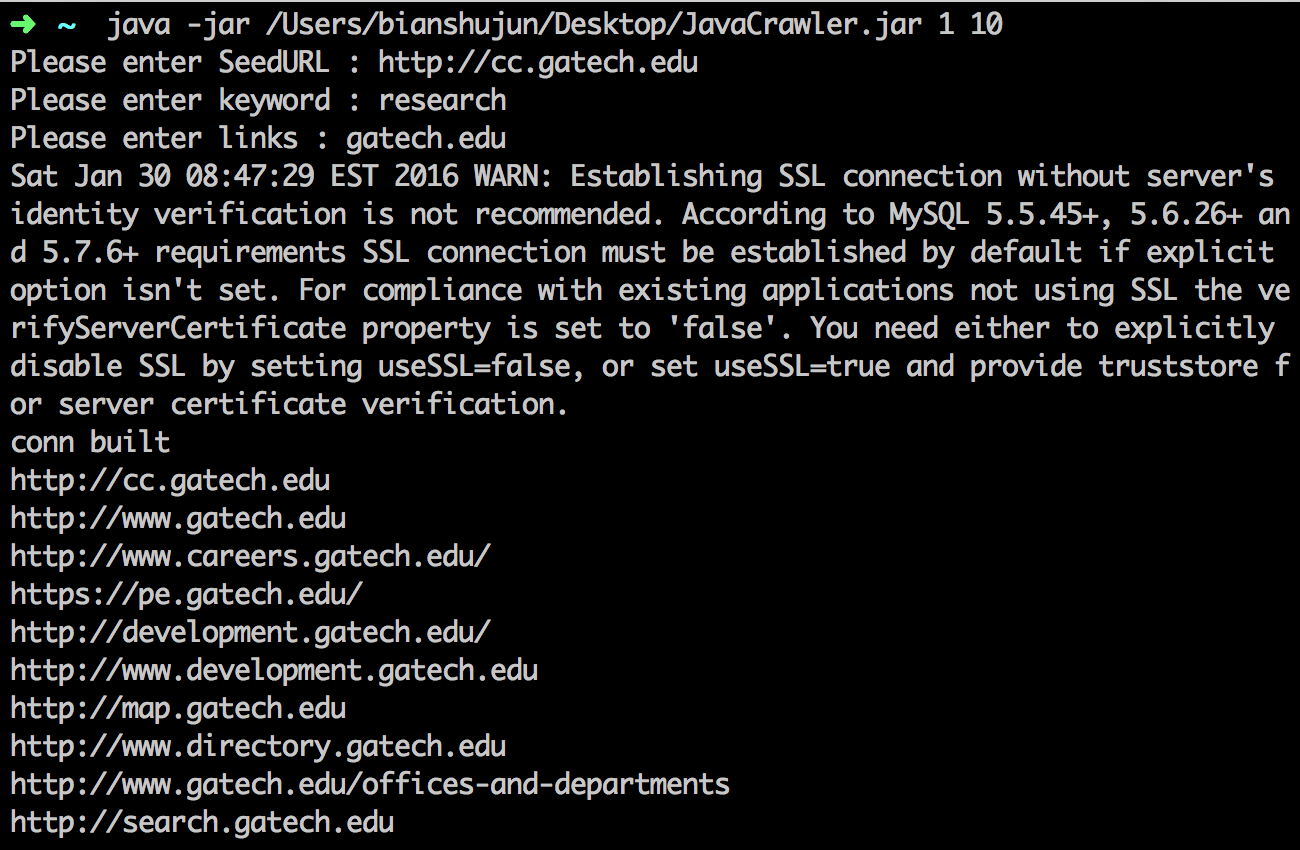
In order to do the permanent store, I used mysql database to save all the URLs.

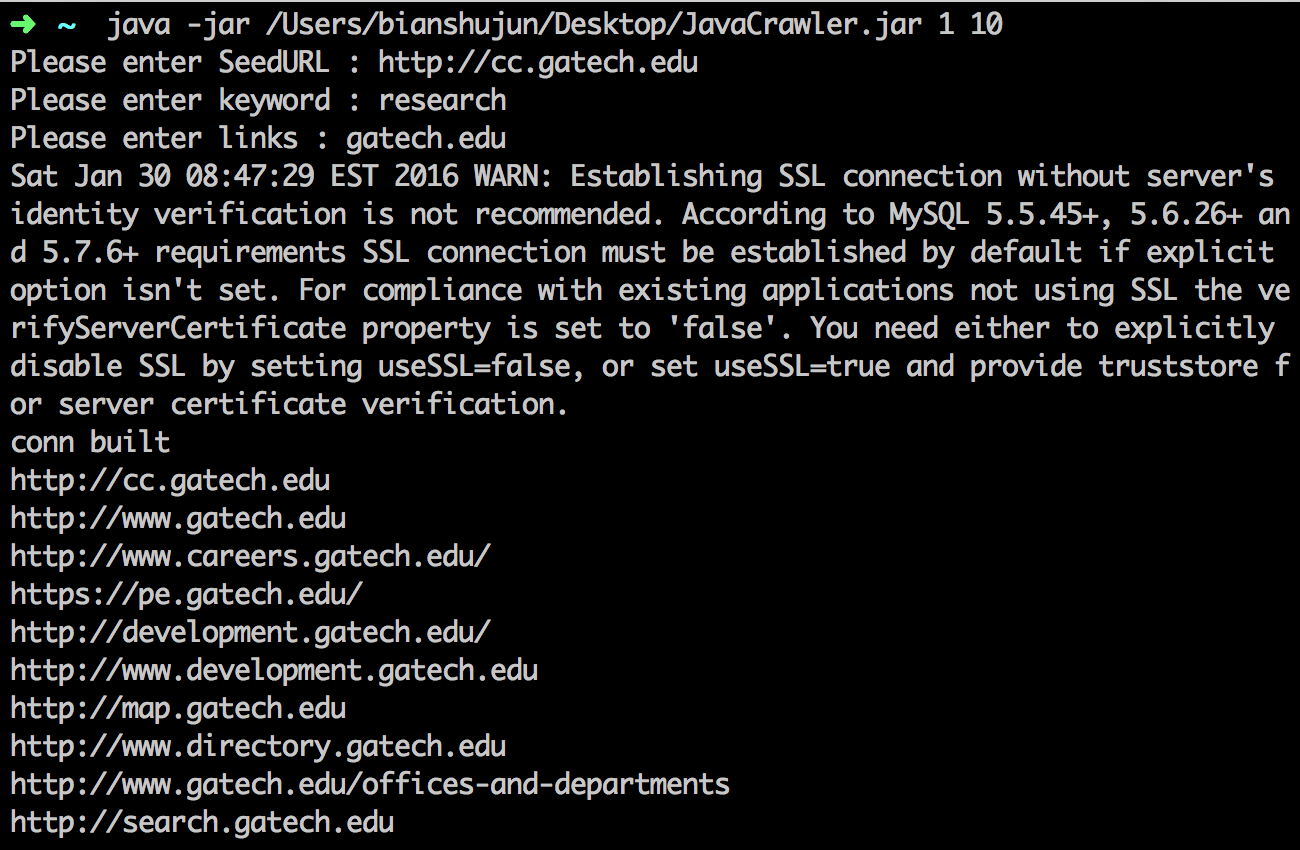
Also in order to improve the performance, I used the multi-thread technology in order to get different URLs at the same time.

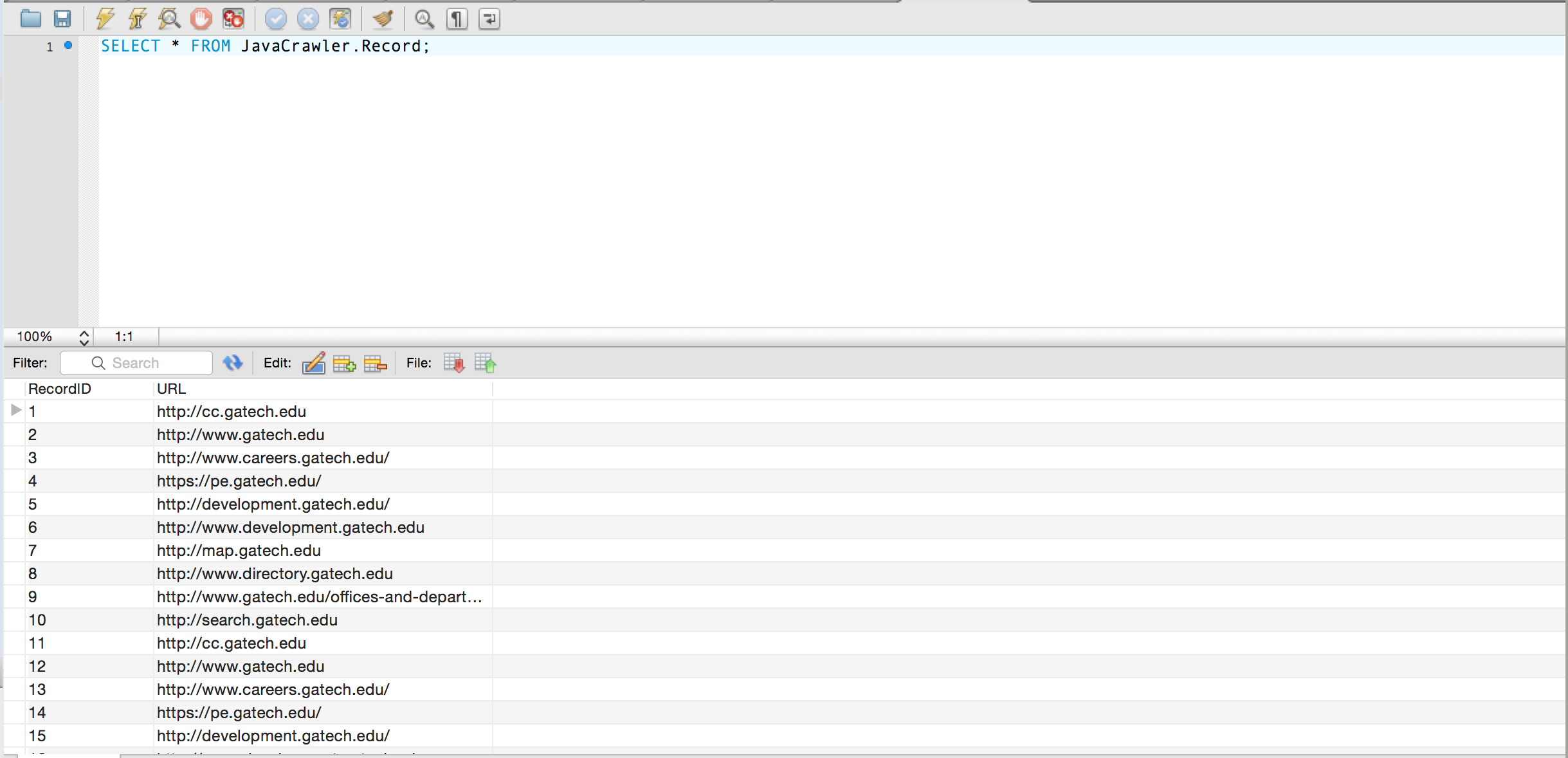
This application still has some places to improve, especially the performance part. I think currently the bottleneck is how to phrase the html file quickly and how to save data to the database quickily. For example, I have tried something like saving while downloading. Or maybe we can design a new kind of data structure.

Also in this version, I didn’t consider the problem like how to avoid the crawler-check of those website.

2. Screenshots







3. Experience and lessons learned

First, I learned the experience with dealing with the mysql. Combining mysql and eclipse cost me a lot of time because there is always some weird problem like Acessing to database denied.

Also I reviewed some basic knowledge of the Java, as I haven’t written any Java in 2 years, I reviewed the operation of ArrayList and DataBase as well as some multi-thread programming.

And through this project, I learned another important tool: Jsoup, which could help me dealing with the html file powerfully and quickly.

Last not the least, I have learned some principle of Web Crawler like how to design it, how to improve it and so on.