Iteration Cycle 3

Date: 11.02-11.04

Team member: Bangjie Liu, Weixiang Ding, Qing Li, Yi Wang, Wenyan Hu.

1. Get a Database Instance from the Amazon AWS

AWS Account Information:

<http://argonne.pc.cc.cmu.edu:8080/11601F15_Team3/11601F15_Team3_Python_Webcrawler/blob/master/docs/AWS_Account_infos.docx>

Try to connect from local to the remote database.

2. Decide the website and the chosen information from the website for crawling and begin to crawl

Website: <http://www.imdb.com>

Algorithm: Breadth First Search algorithm

We fetch the actor’s name and photo from website, and then get the needed information of other actors who ever cooperated with the actor according to the movies he acted. Finally, we can get a connection information data network of the actors.

Stage: Have put the first version of the BFS python code.

<http://argonne.pc.cc.cmu.edu:8080/11601F15_Team3/11601F15_Team3_Python_Webcrawler/blob/master/code/WebCrawler-0-1/WebCrawlerIMDB.py>

Tasks for Cycle 4

Web Crawler mini team: Yi Wang, Weixiang Ding, Wenyan Hu

* Develop a new iterated version of python crawling.

Database mini team: Bangjie Liu, Qing Li

* Create the tables for the information crawled.
* Continue to test the connection of remote database.