

Homework 3: Crawling and Network Analysis COMS 6998: Search Engine Technology, Spring 2011

Instructor : Dragomir Radev (radev@umich.edu)

TA s : Archiman Dutta (ad2839@columbia.edu)
and Abhishek Srivastava (aas2234@columbia.edu)

Since you are all busy working on your final projects, we decided to make homework 3 a non-programming assignment. Note that it will still take a long time to get things working and run the commands, so don't leave it till the last minute.

Part 0. Preparation

Follow the instructions for Clairlib usage given in the pre-assignment.

You may add these lines to your `~/.profile` file to ensure they are set correctly whenever you log in to your CLIC account :

```
export PERL5LIB=/home/cs6998/clairlib/clairlib-fall09/clairlib-core/lib:/home/cs6998/clairlib/perl/lib:  
export PATH=$PATH:~cs6998/clairlib/clairlib-fall09/clairlib-core/util
```

If you are having trouble getting something to run, please double-check your PERL5LIB and PATH variables.

Part 1. Crawling a site and downloading its contents

Try to find a small site (with 100 or fewer nodes). Pick a site that is like an “island” – it has internal links but no external links. Crawl it using the commands below (adapt them as needed). Note that these commands will take a long time to run and may use a fair amount of disk space, so choose your site wisely.

```
% mkdir produced  
% crawl_url.pl --url http://www.mysite.net -o mysite  
% download_urls.pl -c mysite -i mysite -b produced
```

Part 2. Convert the sit download_urls.pl e to a network and get its statistics

Use the following scripts to index the downloaded corpus and convert it to a network.

```
% index_corpus.pl -c mysite -b produced  
% corpus_to_network.pl -c mysite -b produced -o mysite.graph
```

Then generate some statistics about the network with the following script. (Ignore string concatenation warnings if any)

```
% print_network_stats.pl -i mysite.graph
```

Part 3. Network statistics and random graph generation

We have already crawled a large site (www.kzoo.edu - Kalamazoo College in Michigan) for you and converted it into a network. You will analyze one of its departments, namely max.cs.kzoo.edu.

The full network for www.kzoo.edu is stored in the file /home/cs6998/hw3/kzoo.graph. You should use a unix command (like grep) or a script to extract only certain links from kzoo.graph – the links internal to the max.cs.kzoo.edu subgraph (i.e. links with both the source and target URLs within the domain). Let's call this new graph kzoo-maxcs.graph. You can generate statistics for this graph with the same script that you used in Part 2. Note that the script will take around one hour and 40 minutes to run on a network of this size (assuming you set it up correctly).

```
% print_network_stats.pl -i kzoo-maxcs.graph
```

Now generate a random graph comparable to this network using the following script. Replace #nodes and #edges with the number of nodes/edges that you get from the network statistics of kzoo-maxcs.graph. This script will automatically print out the statistics of the random graph as well and should take about 1 hour to finish.

```
% generate_random_network.pl --type erdos-renyi-gnm -n #nodes -m  
#edges -o random_kzoo-maxcs.graph
```

Part 4. Analysis of network statistics

We have calculated network statistics for the entire kzoo.edu network and generated a random graph comparable to this network. This takes very long due to the sheer size of these networks and we don't recommend that you run these scripts on such large corpora for this assignment. Statistics for the kzoo.edu network are in kzoo.stats and statistics for random graph of the same size are in random_kzoo.stats. These files are located in /home/cs6998/hw3/.

Use the provided network statistics to answer the following questions:

- a) Compare the values of the average shortest paths, diameter and clustering coefficients of the two networks. Give reasons for the differences that you observe between the networks with respect to these statistics.
- b) Is the kzoo.edu network a small world? Why (explain in a paragraph)?
- c) Observe the Watts-Strogatz clustering coefficient and Newman clustering coefficient for these two graphs as well as the two graphs that you extracted statistics for in Part 3 (kzoo-maxcs.graph and random_kzoo-maxcs.graph). Explain why the two clustering coefficients match for the random graphs but don't match for the graphs representing actual web networks.

Deliverables

Here is what you need to hand in for each section of this assignment:

Part 1:

Mention the URL that you fed to the crawler script. Turn in the file that contains all the URLs retrieved by the crawler (in the example above, it should be mysite, which will be stored in your working directory). Also, submit a verbal description of the crawling algorithm used in clairlib/poacher – specifically, how does it decide in what order to visit URLs, how many queues does it use, how does it understand what URLs to put in each queue, etc.

(The poacher is at

/home/cs6998/clairlib/clairlib-fall09/clairlib-core/lib/bin/poacher-new.pl.

Instructions on using the poacher:

aas2234@athens /home/aas2234 \$ perl /home/cs6998/clairlib/clairlib-fall09/clairlib-core/lib/bin/poacher-new.pl <http://www.greplin.com>

)

Part 2:

Turn in the statistics of your network that were generated by the script provided (just the output of your script will be sufficient). Briefly comment on the nature and properties of this network that you observe from these statistics.

Part 3:

Turn in whatever script/command you used to generate the subgraph file kzoo-maxcs.graph from kzoo.graph. Also hand in the statistics generated for kzoo-maxcs.graph as well as the statistics for random_kzoo-maxcs.graph.

Part 4: Supply answers to the questions listed above.

Grading

This assignment will be graded out of 50 points which will be distributed as follows:

- Part 1: 15 points
- Part 2: 10 points
- Part 3: 10 points
- Part 4: 15 points

Submission

This assignment is due by 11:59:59 pm Friday, April 15th. For each day that you are late, you will lose 10% of the total points. Please submit files in pdf, html, or txt format. You can upload your submission to the folder HW3 in CourseWorks.