CS595 Intro to Web Science, Assignment #4

Valentina Neblitt-Jones

October 10, 2013

Question 1

From your list of 1000 links, choose 100 and extract all of the links from those 100 pages to other pages. We're looking for user navigable links, that is in the form of:

```
<a href = "foo">bar</a>
```

We're not looking for embedded images, scripts, ¡link¿ elements, etc. You'll probably want to use BeautifulSoup for this.

For each URI, create a text file of all the outbound links from that page to other URIs (use any syntax that is easy for you). For example:

site: http://www.cs.odu.edu/~mln/
links: http://www.cs.odu.edu/ http://www.odu.edu http://www.cs.odu.edu/~mln/research/ http://www.cs.odu.edu/~mln/pubs/ http://ws-dl.blogspot.com/ http://ws-dl.blogspot.com/2-13/09/2013-09-09-ms-thesis-http-mailbox.html etc.

Upload these 100 files to github (they don't have to be in your report).

Answer to Question 1

Question 2

Using these 100 files, create a single GraphViz "dot" file of the resulting graph. Learn about dot at:

Examples:

- http://www.graphviz.org/content/unix
- http://www.graphviz.org/Gallery/directed/unix.gv.txt

Manual:

- http://www.graphviz.org/Documentation/dotguide.pdf

Reference:

- http://www.graphviz.org/content/dot-language
- http://www.graphviz.org/Documentation.php

Note: You'll have to put explicit labels on the graph, see: https://gephi.org/users/supported-graph-formats/graphviz-dot-format/

Note: Actually, I'll allow any of the formats listed here: https://gephi.org/users/supported-graph-formats/, but "dot" is probably the simplest.

Answer to Question 2

Question 3

Download and install Gephi:

Load the dot file created in #2 and use Gephi to:

- visualize the graph (you'll have to turn on labels)
- calculate HITS and PageRank
- avg degree
- network diameter
- connected components

Put the resulting graphs in your report.

You might need to choose the 100 sites with an eye toward creating a graph with at least one component that is nicely connected. You can probably do this be selecting some portion of your links (e.g., 25, 50) from the same site.

Answer to Question 3