Assignment 2

Fall 2014
CS595 Web Science
Dr. Michael Nelson

Mathew Chaney

October 6, 2014

Contents

Listings

1 Question 1

1.1 Question

```
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 of 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/2013/09/2013-09-09-ms-thesis-http-mailbox.html
etc.

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

1.2 Answer

Using the python requests module [2], the content of each of the 100 randomly selected URIs was downloaded. Then, using the BeautifulSoup module [3], each downloaded page was converted into a BeautifulSoup document tree, searched for all HTML links of the form:

```
<a href="http://www.urlhere.com/path/to/some/resource/">link text</a>
```

2 Question 2

2.1 Question

```
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.)
```

2.2 Answer

Using the module urlparse[1] to extract the top level domain name for use as the label.

3 Question 3

3.1 Question

```
Download and install Gephi:

https://gephi.org/

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 by selecting some portion
of your links (e.g., 25, 50) from the same site.
```

3.2 Answer

4 References

[1] The Python Software Foundation. Python urlparse module. https://docs.python.org/2/library/urlparse.html, 1990-2014.

[2] Kenneth Reitz and Contributors. Python-requests. http://docs.python-requests.org/en/latest/, 1990-2014.

[3] Leonard Richardson. Python-requests. http://www.crummy.com/software/BeautifulSoup/, 1996-2014.