

Assignment 2

Fall 2014

CS595 Web Science

Dr. Michael Nelson

Mathew Chaney

October 2, 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 Resources

- None: yet

1.3 Answer

Using the python script in Listing ??, 1000 unique URIs were dereferenced and their raw contents were stored in the `html/raw/` folder as a file with the filename as the md5-hashed URI. These were then stripped of all html elements and their processed contents were stored in the `html/processed/` folder as the same md5-hashed filename. For reference, the URIs were written as the first line of each of their content files.