

Assignment 2

Fall 2014

CS595 Web Science

Dr. Michael Nelson

Mathew Chaney

October 1, 2014

Contents

1	Question 1	3
1.1	Question	3
1.2	Resources	3
1.3	Answer	3

List of Figures

Listings

1 Question 1

1.1 Question

Download the 1000 URIs from assignment #2. "curl", "wget", or "lynx" are all good candidate programs to use. We want just the raw HTML, not the images, stylesheets, etc.

from the command line:

```
% curl http://www.cnn.com/ > www.cnn.com
```

```
% wget -O www.cnn.com http://www.cnn.com/
```

```
% lynx -source http://www.cnn.com/ > www.cnn.com
```

"www.cnn.com" is just an example output file name, keep in mind that the shell will not like some of the characters that can occur in URIs (e.g., "?", "&"). You might want to hash the URIs, like:

```
% echo -n "http://www.cs.odu.edu/show_features.shtml?72" | md5
41d5f125d13b4bb554e6e31b6b591eeb
```

("md5sum" on some machines; note the "-n" in echo -- this removes the trailing newline.)

Now use a tool to remove (most) of the HTML markup. "lynx" will do a fair job:

```
% lynx -dump -force_html www.cnn.com > www.cnn.com.processed
```

Keep both files for each URI (i.e., raw HTML and processed).

If you're feeling ambitious, "boilerpipe" typically does a good job for removing templates:

<https://code.google.com/p/boilerpipe/>

1.2 Resources

- None: yet

1.3 Answer