COSC 481 – Miniproject #1

Task: create a web text scraper and do some rudimentary statistical analysis on the text that you find.

For this project you will create a python program that will do the following things:

1.) Take a significant amount of text from one of the following news (or news-like) web sources:

espn.com
cnn.com
huffingtonpost.com
ign.com
theonion.com

You should select 20 articles from a single web source that are all reporting on a similar topic (for instance, 20 articles on the NFL from ESPN). You can either hard-code the urls into your scraper, or create a web crawler that'll pull the information for you. You may not pull the text by hand to feed in to the remainder of the program.

2.) Print a list of the web sources you are pulling from.

3.) Print the following statistics on the web sources:

mean # of words/article
median # of words/article
most frequent word(s)
frequency list of all words

4.) Also, as an interesting bit of data to pull from web sources, develop a script to identify 10 different speakers quoted across all the articles you scraped, and list them after the statistics.

5.) Submit your sufficiently commented source file via Blackboard by 11:59pm on 2/15. This project may be worked on in pairs. If you choose to work in a pair, you must include that information as part of your comment block at the start of your source.