CSCI 452 (Data Mining)
Dr. Schwartz
HTML Web Scraping
150 pts

Overview

For this assignment, you'll be scraping the White House press briefings for President Obama's terms in the White House to see which countries have been mentioned and how often. We will use a mirrored image (originally so that we wouldn't cause undue load on the White House servers, now because the data is historical). You will use Python 3 with the following libraries:

- Beautiful Soup 4 (makes it easier to pull data out of HTML and XML documents)
- Requests (for handling HTTP requests from python)
- lxml (XML and HTML parser)

We will use Wikipedia's list of sovereign states to identify the entities we will be counting, then we will download all of the press briefings and count how many times a country's name has been mentioned in the press briefings during President Obama's terms.

Specification

There will be a few distinct steps to this assignment.

1. Write a Python program to scrape the data from Wikipedia's list of sovereign states (link above). Note that there are some oddities in the table. In most cases, the title on the href (what comes up when you hover your mouse over the link) should work well. See "Korea, South" for example -- its title is "South Korea". In entries where there are arrows (see "Kosovo"), there is no title, and you will want to use the text from the link. These minor nuisances are very typical in data scraping, and you'll have to deal with them programmatically. You will want to save the list of countries to a file so that you can use them later. **Be sure to alphabetize the countries in countries.py**.

   ** Name this program countries.py **
   ** Save the list of countries as countries.txt **

2. Write a Python program to download and store the index pages listing the press briefings. If you go to the landing page of the press briefings, you'll see that you can do a continuous scroll to keep loading more briefings. This dynamic method doesn't work as well for us, but there are built in parameters we can use for more old-fashioned navigation. Specifically, you can use a parameter on the url to specify a page. The pages start at 1 and go to... (well, you'll have to figure that part out).

   To get the first page, use
   http://cs.millersville.edu/~sschwartz/mirror/www.whitehouse.gov/briefing-room/press-

briefings%3fpage=1 then just keep changing the page number in the url. When there are no more briefings, a 404 (page not found) error is returned. You should play with the urls to see where the pages stop, but **your program should NOT hardcode the last page number** – you should programmatically detect the last page. This would be important if you were writing a scraper that you would use on changing content.

You should store the index pages in a separate directory for easier processing and organization. If you use the makedirs from the os package in python, you can do this from within your program. It would make sense to create filenames corresponding to the page numbers (0.html, 1.html, etc.)

** Name this program indexPages.py **

3. Write a Python program to process the index pages and download all of the press briefings (to a separate directory, for your own sanity!) Start this slowly, just trying to do one index page and download a few briefings. Store both the title and the text of the briefing.

   ** Name this program getBriefings.py **

4. Write a Python program that will use the list of countries you created in step 1 and process all of the press briefings you downloaded in step 3. You want to count how many times each country occurs in the press briefings. Count the number of times the word occurs, not just how many press briefings it occurs in (if Albania occurs 3 times in 1 briefing, the 3 should be counted). There are at least two different formats for the press briefings (the html format changed at some point over the years). So be sure to examine a wide selection and either detect the different formats or find a common tag/cue (I believe there is one). Your output from the program should be a list of the countries along with the count of how many times the names occur in the press briefings. Write the output to a text file – on a line, it should be count then country name.

   ** Name this program countryCount.py **
   ** Save the output to file countryCount.txt **

Create a .zip file containing your four python programs and two text files, named as specified above. Submit your .zip file as the Press Briefings lab to Autolab.