September 30, 2014
Software Design: Homework 3 Part 3

Project Writeup / Reflection

Project Overview
In this project, I originally wanted to look through the entire Internet on Ebola cases. I wanted to create a timeline of how often Ebola was talked about and written about. However, after talking to NINJAs and Amon, it seemed that this was not an easy task. Every webpage has a different way of noting its date, so it would have been difficult to mine for the code that describes the dates of the webpages. After some discussion, I switched my project to look into just one specific webpage. Now my project looks into the Centers for Disease Control and Prevention webpage. I'm looking at specific search words and looking at the frequencies of words under each description of search results. The general approach I took was just trying to go narrower and narrower into the search space.

Implementation
My code first takes a raw input from the user. The user can enter any word he/she would like to search for. Then by looking into all the results pages, it looks at the descriptions under each result. All the words are then split into individual words and saved in a list. After that, the frequencies of each word are counted. Finally the words are sorted from highest frequency to lowest frequency.
Some of the libraries I used are “urllib2”, “operator”, and “BeautifulSoup”. “urllib2” looks into the specified website I declare. “operator” sorts the list. “BeautifulSoup” parses html into easy to use objects; it also extracts tags so I can use them.
Anyone can run my code. It is easy to implement and read. Once run, the user is prompted to enter a search word. Then my code spits out all the frequencies of the description words.
I used a few lists to better organize my results. I used a dictionary to compile all the resulted words.

Examples
$ python hw3.py
Enter search word: ebola
[('ebola', 262), ('of', 113), ('virus', 105), ('in', 100), ('the', 87), ('disease', 70), ('and', 69), ('to', 57), ('with', 51), ('cdc', 50), ('for', 39), ('is', 37), ('patients', 36), ('or', 35), ('a', 34), ('correlates', 32), ('pediatric', 32), ('survival', 32), ('outbreak', 32), ('biomarker', 32), ('podcast', 29), ('latest', 28), ('1617', 26), ('length', 26), ('download', 25), ('listen', 24), ('now', 24), ('mp3', 24), ('hemorrhagic', 23), ('health', 23), ('fever', 22), ('page', 21), ('1', 19), ('on', 18), ('africa', 17), ('by', 16), ('2014', 16), ('west', 14), ('response', 14), ('cause', 14), ('our', 13), ('blogs', 12), ('\xe2\x80\x93', 12), ('us', 12), ('about', 12), ('not', 12), ('infection', 12), ('an', 11), ('javascript', 11), ('note', 11), ('uganda', 11), ('that', 10), ('disabled', 10), ('cases', 10), ('update', 9), ('you', 9), ('outbreaks', 9), ('your', 9), ('several', 9), ('liberia', 9), ('species', 9), ('public', 9), ('infected', 8), ('blog', 8), ('epidemic', 8), ('supported', 8), ('browser', 8), ('number', 8), ('this', 8), ('animals', 8), ('\xe2\x80\x94', 7), ('pestis', 7), ('anthracis', 7), ('bacillus', 7), ('viral', 7), ('agents', 7), ('anthrax;', 7), ('has', 7), ('plague;', 7), ('guinea', 7), ('including', 7), ('voices', 7), ('exotic', 7), ('yersinia', 7), ('global', 7), ('marburg', 6), ('bats', 6), ('workers', 6), ('severe', 6), ('fighting', 6), ('director', 6), ('august', 6), ('sierra', 6), ('letter', 6), ('travel', 6), ('from', 6), ('requires', 5), ('occurred', 5), ('stories', 5), ("world's", 5), ('travelers', 5), ('matters', 5), ('contact', 5), ('have', 5), ('at', 5), ('tracing', 5), ('are', 5), ('respond', 5), ('de', 5), ('september', 5), ('guidance', 5), ('environmental', 5), ('care', 5), ('zaire', 5), ('diseases', 5), ('procedures', 5), ('through', 5), ('nonhuman', 5), ('how', 5), ('united', 5), ('control', 5), ('sharing', 5), ('updated', 5), ('all', 5) …]

$ python hw3.py
Enter search word: HIV
[('hiv', 239), ('and', 169), ('the', 120), ('hiv/aids', 105), ('of', 90), ('to', 77), ('for', 75), ('prevention', 69), ('about', 57), ('this', 51), ('in', 48), ('with', 40), ('page', 40), ('cdc', 33), ('a', 33), ('infection', 31), ('is', 29), ('tb', 28), ('information', 27), ('are', 26), ('more', 26), ('hepatitis', 25), ('please', 21), ('national', 21), ('cdcgov', 21), ('std', 21), ('health', 21), ('united', 21), ('visit', 21), ('surveillance', 20), ('testing', 20), ('message', 20), ('states', 20), ('on', 19), ('viral', 19), ('among', 19), ('care', 18), ('aids', 17), ('risk', 17), ('living', 16), ('&', 16), ('treatment', 14), ('from', 14), ('men', 14), ('people', 13), ('transmission', 13), ('topics', 12), ('share', 12), ('sharecompartir', 12), ('other', 11), ('screening', 11), ('stds', 11), ('us', 10), ('programs', 10), ('awareness', 10), ('who', 10), ('that', 10), ('1', 10), ('can', 10), ('epidemic', 9), ('tuberculosis', 9), ('how', 9), ('new', 9), ('prevent', 9), ('division', 9), (''cdc's'', 9), ('\xe2\x80\x94', 8), ('at', 8), ('facts', 8), ('gay', 8), ('infections', 8), ('medical', 8), ('day', 8), ('an', 7), ('by', 7), ('provides', 7), ('persons', 7), ('az', 7), ('pregnancy', 7), ('stop', 7), ('project', 7), ('research', 7), ('data', 7), ('learn', 7), ('what', 7), ('all', 7), ('nchhstp', 6), ('fast', 6), ('there', 6), ('services', 6), ('use', 6), ('be', 6), ('interventions', 6), ('do', 6), ('partners', 6), ('have', 6), ('number', 6), ('years', 6), ('or', 6), ('effective', 6), ('monitoring', 6), ('global', 6), ('skip', 6), ('adults', 5), ('incidence', 5), ('some', 5), ('center', 5), ('drug', 5), ('4', 5), ('strategy', 5), ('approach', 5), ('youth', 5), ('recommendations', 5), ('areas', 5), ('report', 5), ('fact', 5), ('disease', 5), (''let's'', 5), ('top', 5), ('sheet', 5), ('sexual', 5), ('status', 5), ('2010', 5), ('its', 5), ('counseling', 5), ('estimated', 5) ...]

Reflection

I think overall this project went pretty well. I would say the only thing I really need to change is I need to start on these projects earlier rather than later. I definitely also need to improve much more on my ability to code. There just seems to be way too many commands to remember and understand. However, I'm sure that once I get more and more used to coding, and more and more comfortable with coding, I'll be able to do much more and understand much more.

I believe my project was appropriately scoped. It could have been a little bit harder, but I decided to go easy on myself. But now that there's a foundation, I can definitely improve on this code and build upon it. For this code, I did not have a good plan for unit testing. :(