OLD DOMINION UNIVERSITY

CS 495: Introduction to Web Science
Instructor: Micheal L. Nelson, Ph.D
Fall 2014 Thursdays 4:20pm – 7:10pm ECSB 2120

Assignment # 9
Joseph Elder UIN: 00844802

Honor Pledge
I pledge to support the Honor System of Old Dominion University. I will refrain from any form of academic dishonesty or deception, such as cheating or plagiarism. I am aware that as a member of the academic community it is my responsibility to turn in all suspected violations of the Honor Code. I will report to a hearing if summoned

December 4, 2014

1.1 Question 1

1.1.1 The Problem

1. Create a blog-term matrix. Start by grabbing 100 blogs; include:

http://f-measure.blogspot.com/
http://ws-dl.blogspot.com/

and grab 98 more as per the method shown in class.

Use the blog title as the identifier for each blog (and row of the matrix). Use the terms from every item/title (RSS) or entry/title (Atom) for the columns of the matrix. The values are the frequency of occurrence. Essentially you are replicating the format of the "blogdata.txt" file included with the PCI book code. Limit the number of terms to the most "popular" (i.e., frequent) 500 terms, this is *after* the criteria on p. 32 (slide 7) has been satisfied.

Create a histogram of how many pages each blog has (e.g., 30 blogs with just one page, 27 with two pages, 29 with 3 pages and so on).

1.1.2 The Solution

Using python version 2.7.6 the text file 'blogs.txt' is generated which contains the web address to the 100 RSS feeds for all the blogs to be used for this assignment. The two given blogs are hard coded into the list then 98 more random blogspot blogs are added to the list. The web address of the XML page for RSS feed is what is being added to the list for each entry in the format “http://f-measure.blogspot.com/feeds/posts/default?alt=rss” the XML is then downloaded by the urllib2 python module. The XML is then parsed by the Beautiful Soup python module to make access of the blog titles and item titles much easier.

Each title for each blog is processed and all of the words in every title are counted by occurrence. Each word which occurs in a title is stored in a dictionary with value representing the number of occurrences for each word. Stop words are removed from the list. The 500 most popular terms are kept in a list, this list is sorted and will be used as a list of keys to the dictionary of word occurrences. The blog term matrix is generated in the same format as the PCI example 'blogdata.txt' the matrix will be found in a file also called 'blogdata.txt'.

References

1. https://docs.python.org/2/