CS 495/595 Introduction to Web Science

Fall 2013

http://www.cs.odu.edu/~mln/teaching/cs595-f13/

Assignment #9

Due: 11:59pm Dec 5 2013

(10 points; 2 points for each question and 2 points for aesthetics)

Support your answer: include all relevant discussion, assumptions,

examples, etc.

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.

2. Create an ASCII and JPEG dendrogram that clusters (i.e., HAC)

the most similar blogs (see slides 12 & 13). Include the JPEG in

your report and upload the ascii file to github (it will be too

unwieldy for inclusion in the report).

3. Cluster the blogs using K-Means, using k=5,10,20. (see slide 18).

How many interations were required for each value of k?

4. Use MDS to create a JPEG of the blogs similar to slide 29.

How many iterations were required?

===================================================================

========The questions below is for 5 points extra credit===========

===================================================================

5. Re-run question 2, but this time with proper TFIDF calculations

instead of the hack discussed on slide 7 (p. 32). Use the same 500

words, but this time replace their frequency count with TFIDF scores

as computed in assignment #3. Document the code, techniques,

methods, etc. used to generate these TFIDF values. Upload the new

data file to github.

Compare and contrast the resulting dendrogram with the dendrogram

from question #2.

Note: ideally you would not reuse the same 500 terms and instead

come up with TFIDF scores for all the terms and then choose the top

500 from that list, but I'm trying to limit the amount of work

necessary.