Stanley Zheng
Due Dec 05
CS495/595
Assignment 9

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.

using the first two blogspot rss feeds, and then using 100 other feeds grabbed from aggregator services, a blog matrix was generated. The code was obtained from the Programming-collective-intelligence chapter 3 code; it was used verbatim and fed rss feeds from the aggregator ‘feeds.txt’ From the blog matrix the following words were produced as top

Top Ten Words

- Blog
- how
- best
- life
- zich
- had
- people
- for
- when
- time

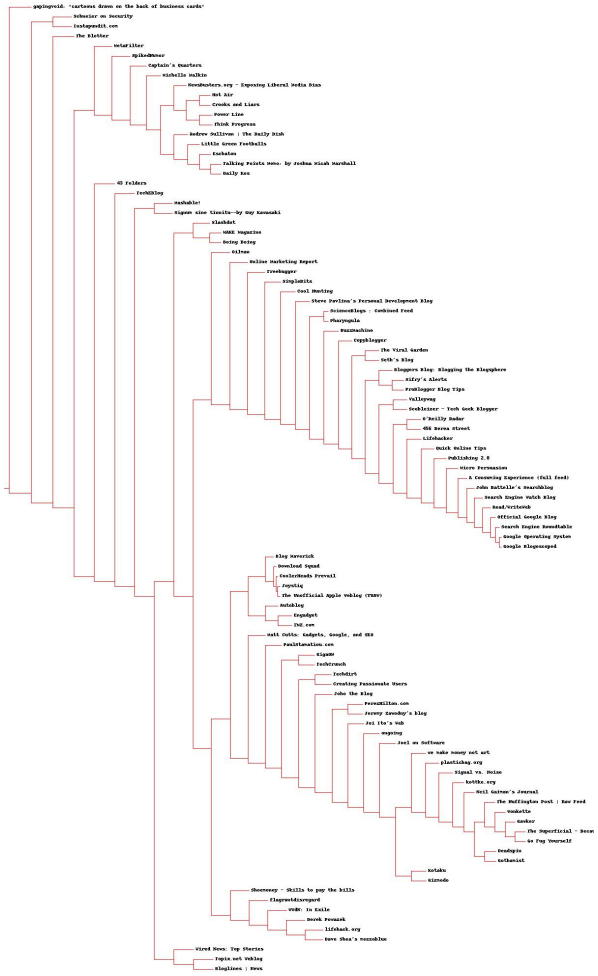
Use the blog title as the identifier for each blog (and row of the matrix). Use the terms from every item/title (RSS) or entry/title (Atom) for the columns of the matrix. The values are the frequency of occurrence. Essentially you are replicating the format of the "blogdata.txt" file included with the 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).

Code from the powerpoint slide 12. Received many errors regarding self-generated ‘blogdata.txt’ from question 1. My assumption is the number of keywords is too few, or the feeds failed to generate enough descriptive text for it to index.

See figure 1. The generated dendrogram from the sample data found with programming college intelligence.
See figure 2 for ‘q2/ascii..’ text version found in file
Figure 1.

Stanley Zheng
Due Dec 05
CS495/595
Assignment 9



Stanley Zheng
Due Dec 05
CS495/595
Assignment 9

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?

Each K-Means was ran as follows with the following output:
K=5: 6 iterations,

Stanley Zheng
Due Dec 05
CS495/595
Assignment 9

im
K =10: 5 iterations,
K = 20: 5 Iterations

Iteration 0
Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5
#break for 5

Iteration 0
Iteration 1
Iteration 2
Iteration 3
Iteration 4
#break for 10

Iteration 0
Iteration 1
Iteration 2
Iteration 3
Iteration 4
#break for 20

Stanley Zheng
Due Dec 05
CS495/595
Assignment 9

4. Use MDS to create a JPEG of the blogs similar to slide 29.
How many iterations were required?

Here is the JPEG created from the blogfile. See figure: 3

Iterations were counted by piping the output into a secondary ‘numbers.txt’ file and then counting the number of lines. The counting in done in ‘q4/count.py’ .

258 Iterations were required to create the ‘blogs2d.jpg'

Figure 3: