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).