ENGR 102 – Programming Practice
Practice Session (Week 9)

delicious (formerly, del.icio.us) is a web site that allows users to save their favorite links (bookmarks) online. Each link has also one or more "tag"s that represent the categories or topics of the website, such as "programming", "cooking", "research", etc.

In this practice session, you will work with a tiny fraction of the *delicious* data, and practice hierarchical clustering on this data. More specifically, you will do the following:

1. Create a dataset of bookmarks suitable for clustering. That is, you are going to create the matrix representation of bookmarks data similar to blogdata.txt that we used in the class. In this matrix representation, rows will be URLs and columns will be tags, and the cells of the matrix will state how many times a given URL is tagged with a particular tag.
2. Once you construct the matrix, run hierarchical clustering on it, and see how bookmarks are clustered. Draw pictures of the clusters by employing the functions that we used in the class.
3. Now, modify your matrix, and cluster tags instead of URLs, and see how tags are clustered together. Draw pictures of the clusters by employing the functions that we used in the class.

DATASET

This dataset consists of 100,000 popular URLs bookmarked on Delicious within a past time window. Each URL includes the date first saved, the number of saves, and the top 10 tags used and their respective counts. The dataset is available on LMS. You are provided different size datasets (data.zip). You may start working with the smallest one (delicious.tiny.txt) initially.

DATA FORMAT

The file is tab delimited; the columns are (from left to right):

- URL
- Number of saves
- Date of first save
- Tag
- Tag count
- [last two repeat to represent the top 10 tags and their counts]

Here is an example:

```
http://boingboing.net/  11053   2002-11-15      blog    5018    news    2763
culture 2542
blogs   2475    technology 2166 fun 1525 tech 1436 daily 1016 art 641 geek 464
```