Individual Assignment 2
Due at the beginning of the class on Nov. 14
Total: 8 points

Document clustering, or text clustering, is a very popular application of clustering algorithms. A web search engine, like Google, often returns thousands of results for a simple query. For example, if you type the search term "jaguar" into Google, around 200 million results are returned. This makes it very difficult to browse or find relevant information, especially if the search term has multiple meanings. If we search for "jaguar", we might be looking for information about the animal, the car, or the Jacksonville Jaguars football team.

Clustering methods can be used to automatically group search results into categories, making it easier to find relevant results. This method is used in the search engines PolyMeta and Helioid, as well as on FirstGov.gov, the official Web portal for the U.S. government.

In this problem, we'll be clustering articles published on Daily Kos, an American political blog that publishes news and opinion articles written from a progressive point of view. Daily Kos was founded by Markos Moulitsas in 2002, and as of September 2014, the site had an average weekday traffic of hundreds of thousands of visits.

The file dailykos.csv contains data on 3,430 news articles or blogs that have been posted on Daily Kos. These articles were posted in 2004, leading up to the United States Presidential Election. The leading candidates were incumbent President George W. Bush (republican) and John Kerry

(democratic). Foreign policy was a dominant topic of the election, specifically, the 2003 invasion of Iraq.

Each of the variables in the dataset is a word that has appeared in at least 50 different articles (1,545 words in total). The set of words has been trimmed according to some of the techniques in "text mining" we will learn later in this course (e.g., punctuation has been removed, and stop words have been removed). For each document, the variable values are the number of times that word appeared in the document.

Answer all the questions in Word, and upload it to Canvas. You do not need to paste your R codes.

Questions:

First, read the data set into R. You should use k-means algorithm to cluster on **all of the variables**.

Note 1: NO need to normalize data as all are on the same scale.

Note 2: It may take while to run the clustering algorithm as we have a lot of observations and a lot of variables!

1. Run K-Means Clustering, setting the random seed to 1000 and the number of clusters equal to 7. List the number of observations in each cluster.

MSIS 510
Individual Assignment 2

2. In class, we interpreted the clusters with respect to the centroids defined by the quantitative variables that were used in forming the clusters. Now, since there are too many variables, instead of looking at the average value in each variable individually, we'll just look at the top 6 words in each cluster. Write down the 10 most frequent words in each cluster, for each of the clusters.

   To do this, you need to note that **km$centers** gives you a matrix (2-dimensional structure). You can use a specific row index to grab a row that is corresponding to each cluster. Then for each cluster, you can sort the mean frequency values of each of the words.

   (**HINT:** you can use the sort() function to sort a vector. See the documentation, or search its usage on Google!)

3. Looking at the output of Problem 2, which cluster could best be described as the cluster related to the Iraq war?

4. In 2004, after beating running mate John Edwards, Howard Dean, Wesley Clark, and other candidates in the primaries, John Kerry became the Democratic nominee. Given this information and output of Problem 2, which cluster best corresponds to the Democratic Party?