INFO290: Data Mining & Analytics  
February 5th - Clustering workshop / homework

Required Tools:

* A k-means clustering implementation (matlab / scikit-learn / R / etc)
* A silhouette coefficient function ( [matlab](http://www.mathworks.com/help/stats/k-means-clustering.html?refresh=true#brah7fp-1) and [scikit-learn](http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html) have these / code your own)
* You may complete this assignment on your own machine or the terminal (the dataset you will be using is called *yelp\_reviewers.txt* on the dlab-matlab machine)
* To complete this in the terminal
  + If you use dlab-matlab, please create your team’s directory in Lab3/
  + OSX: open the “terminal” program in the finder then type:
    - ssh [info290t@dlab-matlab.berkeley.edu](mailto:info290t@dlab-matlab.berkeley.edu)

(*username: info290t password: \*\*\*\* hostname: dlab-matlab.berkeley.edu*)

* Windows: download and then run [putty](http://the.earth.li/~sgtatham/putty/latest/x86/putty.exe), enter in dlab-matlab.berkeley.edu as the host and follow login instructions once connected

**Description:** In class on Tuesday we learned how the k-means clustering algorithm works and metrics that can be used to score the goodness of a clustering. While you could visually judge the goodness of a clustering given 2D data, as the dimensions increase metrics are needed to guide our evaluation. In this Lab you will perform a series of cluster analyses on the *yelp\_reviewers.txt* dataset which has been newly created from the Lab 2 features (questions 3 to 17 plus every group’s submitted question 18 features). The objectives of the Lab are to a) gain experience in running k-means yourself b) quantitatively evaluating the goodness of a clustering c) investigate the answer to your own question about this dataset using these techniques. This Lab is due at 10am next Thursday and will count as a homework. Grading will be based on completion of questions 1-7 (according to google sheet), quality of research question (don’t pick something too simplistic), and appropriateness of techniques used to investigate the question. There will be no quiz or other assignment due next week other than suggested readings. MAX GROUP SIZE: 4

To make it easier to match the feature columns with their meaning, there is a lab3\_header\_lookups.csv file in the info290t folder on the dlab server that maps the feature file column names to their interpretation.

NOTE: When dealing with features that contain -Inf values or other oddities, consider replacing (imputing) those values with the mean of the column or completely removing reviewers with -Info values in any of their features.

**Questions/Objectives**

1. Choose an implementation of k-means and specify it in the google sheet
2. What is the best choice of *k* according to the silhouette metric for clustering q4-q6 (# of cool, funny, useful votes). Only consider 1 <= *k* <= 8.

NOTE: For features with high variance, empty clusters can occur. There are several ways of dealing with empty clusters. A common approach is to drop empty clusters, the prefered approach for this Lab is to treat the empty cluster as a “singleton” leaving it empty with a single point placeholder. This can be accomplished in MATLAB, for example, with the command: [IDX,C]=kmeans(d,4, 'Distance', 'sqeuclidean','EmptyAction','singleton');

1. Answer question 2 but using the log of the features (q8-q10)
2. Answer question 2 but using the percentage of the features (q11-q13)
3. Inspect the [best] clustering generated from question 4
   1. List the number of data points in each cluster (eg. C1: 2,000 C2: 4,200 etc)
   2. Where there clusters that represented very funny but useless reviewers?
   3. How many reviewers were in the cluster that represented relatively equal strength in all voting categories (assuming such a cluster exists in your clustering)?
4. Cluster the dataset using *k = 5* and using features q8-q13 (log and % type votes) and q16 (binarized most active year feature) and the natural log of q17 (avg review chars.)
   1. What is the silhouette metric for this clustering?
   2. What was the average “average star rating (q14)” among the points in each of the clusters (eg. C1: 1.42 C2: 4.23 C3: 3.38 etc)
5. Cluster the dataset using the features described in question 6 + every group’s question 18 features (you may drop features with high incidents of -Inf / blank / or NaN values)
   1. Using the silhouette metric with the elbow method, what was the best *k*?
   2. Using the the sum of within cluster variance metric, with the elbow method what was the best k?
6. For this question please come up with your own question about this dataset and using a clustering technique as part of your method of answering it. This question answer should be submitted to a bCourses assignment in the form a pdf. This report is expected to be between 2 and 3 pages but more is also fine. It is meant to give you practice writing up your results. The report should have the following sections:
   1. DATASET: describe the yelp\_reviews and yelp\_reviewers transformation
   2. RESEARCH QUESTION: Describe what it is you want to ask of this dataset
      1. subsection called FEATURES SELECTED: briefly describe the features from the dataset you choose (at least 1 features should be from a q18)
   3. METHODS: Briefly describe the methods you used (such as k-means) and why you used them
   4. RESULTS: Here you can tell the story of how you investigated the question (plots are always nice) and the conclusions you drew.
7. Bonus question (+15%) - Reviewer overlap: Create a dataset of reviewers x business\_ids where we have a feature for every business ID (there are a lot) and the feature is ‘1’ if the reviewer has written a review for that business and ‘0’ if not. Use the methods described in this assignment to answer the question of how much overlap (of businesses reviewed) of reviewers exists in this dataset. Append this answer to your question 8 document.

Questions asked in class:

Q: Are features clustered individually or joined together? A: joined together

Q: How can I get the file off the server?

A: On your machine type: sftp user@host

Then type: get FILENAME (typing ‘quit’ will exit sftp)

Q: Can scipy and scikit-learn be installed on the server? A: Yep, just installed them

Q: Why are some silhouette scores implausibly good? A: Silhouette coefficient tells you if a clustering has good separation and good density but it does not tell you if objects are equally distributed among clusters. Good coefficients can be calculated if there are clusters that are far away from other clusters but contain very few objects.

Q: Why is the server so slow? A: This server is low memory (16G) and scikit’s implementation is high memory usage. The server memory is exceeded and is using swap (trashing). Please use your own machine if things get too slow. We will look for a potential high-memory replacement server. You are allowed to reduce the dataset size in order to run in a reasonable amount of time on your own computer. Reducing to 10,000 rows (no less) should be fine.