Assignment 3

The deadline for the third assignment is May 23rd, before class. Turn in the code, with instructions on how to run it. The report should include detailed observations on the results. For late submissions the late policy on the page of the course will be applied. Details for the turn-in, and how to write reports are on the Assignments web page of the course. The will be an oral examination of the Assignment.

Question 1

A power-law distribution is defined as $P(X = x) = (a - 1)x^{-a}$, where a is the exponent of the distribution. You are given a set of observations $X = \{x_1, \dots, x_n\}$ that are generated from a power-law distribution. Use the Maximum Likelihood Estimation method we described in class to find the exponent of the power-law distribution that best fits the data observations.

Question 2

Consider the following problem:

Given a set X of n real numbers and an integer k, find a partition of X into k clusters $\{X_1, X_2, \dots, X_k\}$, where $\cup_i X_i = X$ and $X_i \cap X_j = \emptyset$, for all $i, j, i \neq j$, such that the sum of diameters of the clusters $\sum_{i=1}^{k} diam(X_i)$ is minimized, where the diameter of cluster $X_i, diam(X_i) = \max_{x,y \in X_i} |x - y|$, is the largest difference between two numbers in the cluster X_i.

Design a linear-time optimal algorithm for the problem. Prove that your algorithm is optimal.

Question 3 (Clustering)

For this question you will practice with applications of clustering algorithms.

You will use the file "clinton_trump_tweets.txt" (**Attention:** This is a new dataset) which you can download from the web page of the course. Do the same preprocessing as in Exercise 2: remove retweets, and do iterative pruning so that we keep the users that have used at least 20 distinct hashtags/handles, and hashtags/handles that have been used by at least 20 distinct users. In this assignment we will also use the frequency with which a user uses a hashtag/handle.

We will examine two different clustering problems.

1. In the first problem, we will look into clustering of hashtags/handles. Represent each hashtag/handle as a vector of integers with the number of occurrences of the hashtag/handle for each user. You can use the python libraries for feature extraction to construct this representation.

First, you will apply the k-means algorithm. Create a plot of the SSE error of the k-means algorithm as a function of the number of clusters, for k up to 20, in order to determine the number of clusters. Run the k-means algorithm for the number that you will select, and examine manually the resulting clusters. From the hashtags/handles in each cluster try to deduce what is the topic it concerns. Include your conclusions in your report.

Then, run the agglomerative hierarchical clustering algorithm for the same number of clusters, as well as the DBSCAN algorithm. Create the confusion matrix for each pair of algorithms and comment on whether the algorithms find consistent results.

2. In the second problem we will look into the clustering of users. Represent each user as a vector of integers with the number a user has used each hashtag/handle. You can use the python libraries for feature extraction to construct this representation.

Our goal in the second problem is to compare the clustering solution against a known ground truth. In the file "clinton_trump_user_classes.txt" that you are given in the web page of the class, we have the "class" membership for each user id in the data. Class 0 corresponds to Trump followers, while class 1 corresponds to Clinton followers.

Run the k-means algorithm and the four different variations of the agglomerative hierarchical clustering algorithm (single-link, complete-link, average, Ward). Compute the confusion matrix with the ground truth, the precision, recall and F-measure for each algorithm and compare their performance.

For k-means look at the two centers and examine the 30 hashtags/handles with the highest values. Can we draw some conclusion from the most frequent hashtags/handles in each cluster about what differentiates the two clusters?

Hand in your code, all the plots, and the results, and the discussion about them.

Bonus: For the users that you have to cluster in the second part of the question, you can obtain additional information from the file with the tweets. Use this information to extract additional features so as to improve the clustering result. Compare with the original solution with respect to precision, recall and F-measure.