CSC 177-01 Data Warehousing and Data Mining (Spring 2018)

Mini-Project 1: Clustering

Due at 10:00 am, Friday, March 9, 2018

Demo Session: class time, Friday, March 9, 2018

In this project you will practice with applications of clustering algorithms.

You will use the file "clinton_trump_tweets.txt". The file contains the tweets on Twitter collected during 2016 US presidential election. The file contains tab-separated entries with 14 columns that correspond to the following fields:

Name, ScreenName, UserID, FollowersCount, FriendsCount, Location, Description, CreatedAt, StatusID, Language, Place, RetweetCount, FavoriteCount, Text

Note: the file contains ISO-8859-1 encoded data. If you use *read_table* in pandas to read the file in, set parameter "encoding" = "ISO-8859-1".

1. Preprocessing (30 pts)

Step 1: First, you need to clean up the data. Each line of the file is a tweet. Throw away all tweets that are retweets (the text starts with RT), and from the text keep only the hashtags (words that start with #) and the handles (words that start with @). Create a "basket" for each tweet that contains at least one hashtag or handle.

Step 2: Do iterative pruning so that we only keep **the users that have used at least 20 distinct hashtags/handles**, and **the hashtags/handles that have been used by at least 20 distinct users or 20 tweets**. In this project, we will also use the frequency with which a user uses a hashtag/handle (i.e., how many times a user uses a hashtag/handle).

2. Clustering (70 pts)

We will examine two different clustering problems.

Problem 1: Clustering of hashtags/handles

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.

(10 pts) 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.**

(10 pts) Then, run the SSE-based agglomerative hierarchical clustering algorithm for the same number of clusters. Setting the parameter "linkage" to "ward" gives you SSE-based agglomerative hierarchical clustering. From the hashtags/handles in each cluster, try to deduce what is the topic it concerns. **Include your conclusions in your report.**

(10 pts) Compare the clustering result of K-means algorithm with that of the agglomerative hierarchical clustering algorithm. **Include your conclusions in your report.**

Problem 2: Clustering of users

In the second problem we will look into the clustering of users. Represent each user as a vector of integers with the frequency (i.e., how many times) 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", 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 (K=2) and the two different variations of the agglomerative hierarchical clustering algorithm (MAX-based and SSE-based). Setting the parameter "linkage" to "ward" gives you SSE-based agglomerative hierarchical clustering while "complete" gives you MAX-based agglomerative hierarchical clustering.

(30 pts) Compute the confusion matrix with the ground truth, the precision, recall and F-measure for (1) the k-means algorithm, (2) MAX-based agglomerative hierarchical clustering, and (3) SSE-based agglomerative hierarchical clustering. **Compare their performance and include your conclusions in your report.**

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

Teaming:

Students must work in teams of 2 or 3 people. Think clearly about who will do what on the project. Normally people in the same group will receive the same grade. However, the instructor reserve the right to assign different grades to team members depending on their contributions. So you should choose partners carefully!

Deliverables:

(1) **All your source code** in terms of Python Jupyter notebook or Python source files.
(2) **Your report in PDF format.** In the report, include a section "**Task Division and Project Reflection**", where you should describe the following:
- Name and Id of each member in your team,
- who is responsible for which part,
- and what you have learned from the project as a team.

10 pts will be deducted for missing that section.

All the files must be submitted **by team leader** on Canvas before

10:00 am, Friday, March 9, 2018

NO late submissions will be accepted.

Demo Session:

Each team member must demo your work during the scheduled demo session. Failure to show up in demo session will result in **zero** point for the project.

Datasets:

Please use the following fixed link on google drive:

https://drive.google.com/open?id=1SXio43zM6JUiCtiCthwJVnE-CueRLrz8