Assignment III - Due Apr 23, 2017

March 28, 2017

The homework has 100pt. Please submit a single pdf file. You can paste your R code with result into pdf file. Please add comments in your code explaining what the code does. 10pt will be deduced if there is no comment.

Code style reference: `https://google.github.io/styleguide/Rguide.xml`.

You will need to load twitter dataset sql file into your local database for this assignment.

1 Exercise I (40pt)

- Use R and R-SQL API to execute the SQL query and get the result.
- Step 1: use SQL to get a list of unique user_id in twitter_message table as a vector in R.
- Step 2: use R to randomly generate 3 user_id (you can use sample() function or other functions in R).
- Step 3: use SQL to extract all tweets in twitter_message table under those 3 user_id in previous step.
- Step 4: combine tweets from the same user into one variable. As a result, you will have 3 variables(long strings) in R.
- Step 5: remove all numbers, punctuations, non-English words, and stop words using R.
- Step 6: make use of term-document matrix and plot wordclouds. There will be 3 plots in total. You need to put 3 plots into 1 page (refer to L7)

2 Exercise II (40pt)

- Run step 1, 2, 3 in previous exercise.
- Step 4: make a label(index) vector for those tweets/documents. The label is the user_id you have. Thus, all documents posted by one user should have the same label, and you will have 3 unique labels in total.

- Step 5: remove all numbers, punctuations, non-English words, and stop words using R.
- Step 6: get document-term matrix.
- Step 7: apply k-means clustering and hierarchical cluster to get cluster label on each documents. The input matrix is the document-term matrix from step 6. The number of cluster is 3.
- Step 8: use table() function in R showing the difference between cluster results and user_id label.

3 Exercise III (30pt)

- Make a Rshiny app including all your answers for previous questions.
- Make one tab for each exercise, and display your answers as required in following items.
- For exercise I, display the SQL queries you used to get the user_id and the tweets, the 3 user_id you selected and wordcloud plot.
- For exercise II, display a summary of number of tweets, which shows the number of tweets under each user_id. You also need to display the k-means and hierarchical clustering result, and the final comparison.