Web Mining

Assignment 1: Twitter Conversations (50 points, 5% weightage)

Date Posted: Aug 13, 2014
Date of submission: Aug 21, 2014. 9pm.

Goal: To introduce Twitter data and analyze conversations on Twitter.

Data: The dataset (768 MB zip file) has been made available by the TAs at /share/Assignments/Assignment1 under the user WebMining@DC
The link for dropbox download for PGSSP students is here:
https://dl.dropboxusercontent.com/u/85901834/tweets_2014_07_31_00.zip

Sharing Twitter data publicly is illegal. Hence please restrict the dataset to yourself. The dropbox link will be removed on Aug 21.

Questions

1. Extract conversations from Twitter feeds. [10 points]
    a. Each tweet has a tweet ID. Also, a specific field describes if a tweet is in reply to another tweet. Note that while most conversations may look like a list, some conversations could look like trees.
2. Get statistics about Twitter Conversations.
    a. Distribution of #participants. Plot freq vs #participants. [2 points]
    b. Distribution of conversation length. Plot freq vs conversation length. Conversation length is the number of nodes (tweets) in a conversation tree. [2 points]
    c. Distribution of conversation duration. Plot freq vs conversation duration. Duration is the difference in minutes between the first and the last tweet in the conversation tree. [2 points]
    d. Manually tag 100 conversations into one of these categories: travel, weather, sports, movies, product review, natural calamity, finance, politics, personal 1-1 conversation, Others. If others, try to specify a category. Plot freq vs conversation category. [10 points]
3. Build a social bot that talks like a human, based on Twitter conversations. Build a software that does the following in a while loop. [15 points]
    a. Waits for the next user query.
    b. Find the best matching tweet t for the query such that the set of conversations contain a reply for the tweet.
    c. Present the reply for tweet t from the corresponding conversation tree as a result of the query.
4. a. What is a good way of improving the above social bot system for continuous multi-turn dialogues? [3 points]

b. Imagine that you have a new search engine which shows Twitter conversations as results rather than webpages as results. What factors should be taken into consideration when ranking the results? [3 points]
c. Imagine that you have a new search engine which shows Twitter conversations as results rather than webpages as results. How would you generate snippets to be shown on search results page for such a search engine? [3 points]

Submission Instructions: Create a directory with the name "<rollno>_as1". Within that create directories q1 and q3. Zip "<rollno>_as1" folder to get <rollno>_as1.zip and upload it. Please do NOT upload the original datasets. Do NOT upload any other extra files, except those which have been asked for below.

In directory q1, put in the code and also a file which contains all extracted conversations. This file of conversations contains 1 conversation per line. A conversation is represented as a tree. The tree must be output in nested brackets notation with tweet ids as the nodes.

In directory q3, put in the code.

Besides q1 and q3, inside the "<rollno>_as1" folder, put in a readme.pdf file that contains 4 sections:

1. A description of your approach for Q1. Also put in instructions on how to run your code.
2. Answers to Q2.
3. A description of your approach for Q3. Also put in instructions on how to run your code.
4. Answers to Q4.