Homework #1

CS 6676/7675, Spring 2017

---

100 points total [5% of your final grade]

Due: February 2, 2017 by 11:59pm
[no submission will be accepted after February 7, 2017 at 11:59pm]

Delivery: Submit via Canvas

Support Donald Trump after Election

In this homework, you are going to learn the basics of the social media mining by collecting Twitter data, preprocessing the data, and conducting exploratory analysis.

Task 1: Collecting Twitter Data

In this task, you will collect some Twitter data by using Twitter APIs, especially, using Rest APIs.

By using a search API, collect at least 3,000 tweets containing a keyword "donald trump", and store JSON format. Add following conditions when you use the search API:

- A tweet's geolocation should be inside USA.
- Tweets should be posted after election date (November 8, 2016).
- Tweets should be written in English.

Include a few sample tweets in your report. Also, write the detailed steps that you followed in order to collect your Twitter data. Report the size of your dataset.

Task 2: Preprocessing the data

Preprocess the collected dataset by extracting the relevant properties/fields from the JSON format.

- Extract following properties from each tweet: created_at, tweet_id, text, user_id, geo, coordinates, user_id, user_name, user_location, place, country,friends_count, followers_count and language. Useful links to parse a JSON file are listed below:
  - http://stackoverflow.com/questions/15284194/easiest-way-to-extract-fields-from-json
  - http://stackoverflow.com/questions/12934699/selecting-fields-from-json-output
- Determine sentiment infromation of each tweet: sentiment analysis is the process of identifying opinions expressed in a text. It determines whether the attitude towards a particular topic (positive, negative, or neutral)
  Useful links to sentiment analysis are listed below:
  - http://text-processing.com/demo/sentiment/
  - http://nlp.stanford.edu/sentiment/

Include the steps you followed to parse the JSON data and apply sentiment analysis with some examples in your report.

Task 3: Exploratory Analysis through *k*-means clustering

In order to gain clear understanding of the data that you preprocessed, you will conduct exploratory analysis. Based on knowledged regarding on *k*-means clustering algorithms learned in CS5665, you will conduct cluster analysis.

- Explain the reason you selected "k" number for your clustering and how each cluster is different from each other.
- In order to understand the support of Donald Trump in a different state, visualize the tweets distribution based on sentiment information on the map using tools like Tableau, Google maps API, basemap, D3, etc. *Report your findings including some figures like snapshots of the maps.*
- Perform one more interesting analysis of your choice.

Report what you analyzed, found and learned.

What to turn in:

- Submit to Canvas your report: a PDF document containing the answers regarding the above tasks. (hw1_yourname.pdf).
- This is an individual assignment, but you may discuss general strategies and approaches with other members of the class (refer to the syllabus for details of the homework collaboration policy). At the top of your report, please write the names of classmates you consulted and the nature of your discussion.