Homework #2

CS 6676/7675, Spring 2017

100 points total [5% of your final grade]

Due: March 5, 2017 by 11:59pm
[no submission will be accepted after March 10, 2017 at 11:59pm]

Delivery: Submit via Canvas

Yelp Data Exploration

First download a Yelp dataset (1.8Gb data). Do NOT download "photo auxiliary file". Decompress the downloaded file, and use only three JSON files for this assignment: *yelp_academic_dataset_business.json*, *yelp_academic_dataset_checkin.json*, and *yelp_academic_dataset_review.json*. If you want to load a large file to see the content, you may use "Ultra Edit" tool.

Task 1: Data Download and Descriptive Analysis

In this task, extract business IDs from *yelp_academic_dataset_business.json* if a business's city/location is either "Pittsburgh" or "Charlotte". By using each business ID as a unique key, extract corresponding checkins and reviews from *yelp_academic_dataset_checkin.json* and *yelp_academic_dataset_review.json*. Then, answer the following questions:

- Which types of restaurants are the most popular in each city? Define and describe what "popularity" means.
- In order to understand the popularlity of each city, visualize distribution on the map using tools like Tableau, Google maps API, basemap, D3, etc. *Report your findings including some figures like snapshots of the maps.*
- Report one more interesting finding through descriptive analysis.

Task 2: Text Processing & LDA (Latent Dirichlet Allocation)

Text processing is a fundamental element of creation or manipulation of text. Latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics.

Useful links to topic modeling are listed below:

- ➢ http://nlp.stanford.edu/software/tmt/tmt-0.4/
- ➢ http://mallet.cs.umass.edu/topics.php

- What words are most frequently used to describe Chinese restaurants? To identify a restaurant’s type, refer to a business’s meta info in *yelp_academic_dataset_business.json*
- What are the major themes/topics in the reviews of Chinese restaurants?

What to turn in:

- You should turn in TWO files: your code to solve hw2 (e.g., hw2_yourname.zip) and a PDF report (hw2_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.