Yelp Data Challenge Final Project

11/22/2016

Yelp Dataset Challenge

Description

Yelp is hosting a data challenge. The data includes

- 2.7M reviews and 649K tips by 687K users for 86K businesses
- 566K business attributes, e.g., hours, parking availability, ambience.
- Social network of 687K users for a total of 4.2M social edges.
- Aggregated check-ins over time for each of the 86K businesses
- 200,000 pictures from the included businesses

From 4 countries and multiple cities:

- U.K.: Edinburgh
- Germany: Karlsruhe
- Canada: Montreal and Waterloo
- U.S.: Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison

More details can be found here: https://www.yelp.com/dataset_challenge

Your job is to work on a project that you can potentially apply for the challenge using method you learned in class. But more importantly, I want you to have fun with this project since it's a really cool data. How you formulate the problem is up to you and you can try multiple things but for the final project evaluation, I'd like you to have at least one model with random effect(s).

This is a great opportunity to practice the skills you have learned and even earn money if what you do gives interesting outcome. Evaluations for the class will be based on the following criteria.

- How interesting your report is or how well you can sell your result.
- How attractive your EDA figures are.
- How well you've formulated the problem to tangible but still interesting level.
- The correctness of the random effect model.
- Model checking and other validations to solidify your analysis.
- Well formatted final result.

The data size may seem daunting at first but you can focus on specific subset of the data. Here are information you can use to get your data working. I have personally reduced the json into csv using one of the script I've found below and it worked fine. If you'd like a pointer I will be happy to give you one but I'd like to see what you are capable of doing on your own.

- Examples of past challenges http://www.ics.uci.edu/~vpsaini/ http://blog.nycdatascience.com/r/project-1-exploratory-visualizations-of-yelp-academic-dataset-draft/ https://www.springboard.com/blog/eat-rate-love-an-exploration-of-r-yelp-and-the-search-for-good-indian-food/
- Python script to convert json into csv https://raw.githubusercontent.com/rchen314/SpringboardCapstone_YelpAnalysis/master/ConvertYelp.py https://gist.github.com/paulgb/5265767
- dplyr and MySQL http://stackoverflow.com/questions/29878227/write-table-in-database-with-dplyr
- Using Hue and Hive http://blog.cloudera.com/blog/2013/04/demo-analyzing-data-with-hue-and-hive/
- Using MongoDB http://blogs.uoregon.edu/rclub/2016/01/27/notebook-playing-with-the-yelp-challenge-data/