Electives

1. Data Science (STAT 231)
2. Non-Parametric Statistics (STAT 225)

Off-campus courses: Not applicable

Project Summary

Project Overview:

Yelp provides its multitude of users with the ability to find particular businesses that meet particular requirements by utilizing a review-based forum. Within each review lies an abundance of information that can be analyzed using a variety of techniques, such as text and sentiment analysis, in order to extract trends underlying the Yelp framework. Our goal is to capitalize on the abundance of Yelp reviews written for a large number of businesses in order to obtain novel insights that may better inform Yelp users and to utilize visualization techniques, such as shiny apps, to better communicate the trends we observe and the results we obtain to the Yelp audience.

Additionally, this project qualifies for Yelp’s dataset competition, open to undergraduate and graduate students in a related field of study. All submissions are judged on a variety of criteria, including technical rigor, novelty, and relevance to Yelp’s users. Ten awards are available for the top submissions chosen by Yelp.

Project Data:

For their data competition, Yelp has provided a downloadable subset of their business, review, and user data. More specifically, the data includes access to 5,996,996 reviews, 188,593 businesses, 280,992 pictures, 1,185,358 tips (suggestions to the business), 1,518,169 users, aggregated check-in times for each business included, and over 1.4 million business attributes, including parking, atmosphere, etc. Following agreement to Yelp’s terms of use, this data becomes available for download as a set of JSON files.

The included JSON files are as follows:

1. **business.json** - data pertaining to a particular business, including attributes, location, and categorical classification (i.e. food, drinks, etc.).
2. **review.json** - data pertaining to a particular review, including the review text, user id, and business id for which the review was written.
3. **user.json** - data pertaining to a particular user, including the user’s friend mapping, name, and number of reviews written.
4. **checkin.json** - a series of times the mark checkins on a particular business.
5. **tip.json** - data pertaining to tips, including the time, number of likes, and the actual text of the tip.
6. **photo.json** - data pertaining to the images on a business’ Yelp page, including the caption and classification (i.e. food, drink, etc.)
Given the large size of the data and our limited time frame, we decided to reduce our data set by focusing on food-related business (i.e. restaurants, bars, etc.) and by selecting the following cities: Toronto, Phoenix, Pittsburgh, and Cleveland. Toronto and Pittsburgh were selected to address possible differences between Canada and the United States. In total, there are 287,169 reviews for food-related businesses. Cleveland and Phoenix were selected to address possible differences between the East and West of the United States. In total, there are 364,336 reviews for food-related businesses.

**Project Objectives:**

Using the subset of the competition data provided by Yelp, we hope to provide novel insights that may better inform Yelp business suggestions to its users. More specifically, for our capstone project, we plan to address the following questions:

1. Can we use sentiment and text analysis (i.e. identification of keywords) in reviews to predict the star rating given? If we compile a star rating for each review, does the average accurately predict the overall rating a business receives?

2. There are multiple sentiment lexicons available, such as AFINN and nrc (in the lexicon package). Does the sentiment assigned to a review by a particular lexicon better predict the star rating over another lexicon?

3. Are there any differences in sentiment and star rating between the East and the West of the United States? Similarly, does the relationship between sentiment and star rating suggest that Canadians are, in fact, kinder than Americans?

To communicate the answers to these questions, we plan to produce a shiny app or website with interactive components. To best express our results, we hope to include a variety of visual components in addition to text, graphs, and data. Additionally, we plan to provide a complete write-up to outline our extensive data wrangling process, analyses, and findings along with some general information on sentiment analysis and the lexicons used.

**Proposed Tasks**

The primary focus of my Capstone project is on utilizing sentiment and text analysis of a subset of the Yelp data to uncover novel relationships and to potentially predict rating or another feature of a particular food-related business. There is still an abundance of data provided by Yelp left virtually untouched by our analyses. In particular, the business.json file includes location (longitudinal and latitudinal) data for businesses. Using network analysis, particularly Neural Networks, and predictive models, I would like to explore the relationship between restaurants within a defined geographical distance and food type, price, and potentially other business attributes.

More specifically, for my comprehensive project, I propose the following objectives:

1. Create a function that uses the latitudinal and longitudinal data of each restaurant to find and return all restaurants within a defined radius. The geosphere package in R provides specific commands to convert such data into distances.

2. Provide an introduction to Neural Networks including basic background, related conditions, applications, and a justification for applying to the Yelp data. Additionally, I plan to demo Neural Networks on a toy data set as an example of its application.

3. Use Neural Networks to predict the overall rating and/or the primary sentiment (the most common sentiment observed) of reviews received by a restaurant using information from the restaurant of interest and the restaurants within a specified area. The neuralnet package in R provides a function to easily fit Neural Networks.
4. Compare the predictive power of Neural Network models for star rating or overall review sentiment to the predictive power of a Simple Linear Regression model.

This list of objectives will incorporate some of the results of the Capstone project, as our primary focus is on sentiment and text analysis. On the other hand, the incorporation of geographical data and more in depth Neural Network analysis expands beyond the goals of the Capstone. As I have not utilized any type of network analysis before, these objectives will allow me to demonstrate my ability to independently learn and apply statistical techniques and to communicate those results.