Data Mining: Yelp Reviews

Tyler Lugger, Amogh Jahagirdar, Elias Bezanis

Project Goal

- In this project, we plan to use the yelp review data to predict user behavior, specifically for restaurants, in different aspects.

- Some core questions we want to answer
  - What day of the week is a user more likely to eat out?
  - What type of restaurant is a user most likely to review next?
  - Are there some factors outside of restaurant's control that may bias the rating? (i.e weather)

Prior Work

- 9th year of the Yelp Dataset Challenge
  - 8 years of winning projects to review
- 1st Year Winners of the Competition:
  - One group built an algorithm that isolated what factors customers specifically cared about, when reviewing their restaurant.
- 4th Year Winners:
  - Analysis of different text mining techniques to extract most meaningful phrases.
- Marowen Ng
  - Find people who consistently give only 1 or 5 star reviews
  - Predict ratings for different businesses based off of specific user behavior

Dataset Downloaded

- https://www.yelp.com/dataset_challenge
- Enormous dataset (5GB) available for academic purposes
  - Potential $5,000 award depending on what we can find with this data
- Contains:
  - 4.1 Million Reviews and 947 Thousand Tips
  - 1 Million Users and 144 Thousand Businesses
  - 1.1 Million Business Attributes
- Focus on businesses from 11 cities in 4 countries

Work To Be Done

- Data Cleaning/Preprocessing
  - First have to filter data to only get restaurants
  - Store the resulting data into a database such as MongoDB
- Data Integration
  - Since part of one of the questions we're interested in is how outside influences at a given time might influence rating, it's necessary to have this data.
  - So, for example, to see if there is a correlation between weather and ratings, it will be necessary to store data from a weather api.
- Actual Processing
  - Will have to try out different ML techniques for predicting user behavior (e.g Decision Trees, SVM, etc.)

Tools To Be Used

WEKA for precursor, basic analysis.

Database: MongoDB

Language: Python

Frameworks: Pandas, numpy, scikit-learn, matplotlib

Evaluating Our Results

Statistical tests to determine if our predictive models are effective:

- Can be as simple as percentage of test cases that the predictive model properly labeled.
- More involved can be cross-validation of the predictive model
- Could be interesting to see growth of accuracy of model as sample size increases, to assess sensitivity to new data.