Stephen Paff

Yelp Project Report

For this project, I conducted SGD Regression analysis on the review data from the 2017 Yelp Dataset challenge (<https://www.yelp.com/dataset/challenge>).

My final analysis included two sections:

1. Part One created the sample of Nevada reviews used in the analysis.
2. Part Two created a SGD Regression model to analyze that sample.

Part One

Part One had three basic stages:

1. Created sample: The following sample was created: 1,000 most commonly reviewed businesses in Nevada and 20,000 most prolific reviewers of those businesses
2. Analyzed sample: The following were variables were calculated:
   1. Polarity and subjectivity of review language
   2. Distance from the Los Vegas Strip
   3. Day of the week and month of review

These were created in a matrix called X.

1. X was stored as a csv file for later use

Part Two

Part Two had three basic stages:

1. Created the X and y matrices for independent and dependent variables respectively. The following variables were analyzed in the model: latitude, longitude, distance from the Vegas Strip, day of the week and month attended, average stars for reviewer, and review polarity. y is their star score.
2. Created and ran SGR Regressor model
3. Calculated coefficient of determination

The coefficient of determination for the model was 0.291 (a measure of accuracy with scale 0 to 1). A more accurate model could be found by providing demographic data for the reviews. As is, only business data was usable.