Stephen Paff

Yelp Dataset Challenge

Data Story and Visualization

6/8/17

**Description:** In this report, I describe what type of statistical analysis and visualization techniques to understand the data from the Yelp Dataset Challenge.[[1]](#footnote-1) I will outline my initial approaches and strategies.

**Statistical Tests:** I intend to conduct the following statistical tests to learn more about the data:

* Cluster analysis of textual reviews: Determine common word associations to categorize restaurants
* Sentiment analysis of customer reviews: Use the polarity measure in TextBlob’s sentiment analysis to analyze how positive or negative a review is. A few tests will be conducted to determine the validity of this metric. For customers who both reviewed and rated, does this metric yield similar results to their number rating, and how would they compare? Is there a relationship between how positive and negative the language of a review is with objective or subjective language?
* Test for statistically significant differences leading to positive reviews between and within subcategories: Within each style of restaurant, are there statistically significant variables potentially influencing a positive or negative review and rating. Factors to consider are the time of year, location, regional variation, and style.
* Cluster analysis of business by user: Are there patterns in what type of businesses a given user enjoy? For example, do users who enjoy Restaurant A typically also enjoy Restaurant B (enjoyment would be measurement by their rating and/or the positivity of their review)? Attempt to group businesses by user preferences, so that if a customer presents an interest in one business, Yelp can helpfully recommend other restaurants as well.

**Visualizations:** These are initial graph ides to use to convey these ideas:

* Bubble Scatterplots: These would analyze the average positivity of reviews/ratings for a business. I am interested in compared within subcategories, particularly common and pertinent restaurant subcategories such as “fast food,” “Indian food,” etc.
* Density Scatterplots: These would analyze the distribution of sentiment polarity across a various variables spelled out in the second bullet above (particularly subjectivity and review) to determine correlation. Traditional scatterplots with distributions histograms could work, but density mappings would be better, although require more processing and take longer. I intend to plot these as a bivariate distribution, deciding between one of the forms described on this website: <http://seaborn.pydata.org/tutorial/distributions.html>.
* Cluster Analysis Graph: In addition to simple scatterplots to represent cluster analysis, I also intend to use network graphs to represent the connections between different clusters and categories. Above I outline a few different methods to connect business and create clusters (by users, by product categories, etc.), and for each, I will use network graphs with nodes based on the variable or feature I am using to connect them.

1. See this link for a description of the challenge: <https://www.yelp.com/dataset_challenge>. [↑](#footnote-ref-1)