Exploring the Yelp Dataset

– provided by and the intellectual property of Yelp

Group Members:

Carter Reid reidcc@Colorado..edu

Garrett Lubin galu7424@Colorado.edu

Maxwell Reynolds mare3521@Colorado.edu

Surya Jatavallabhula suja3865@Colorado.edu

What we have:

- 8.69 gigabytes of Business, User, and Review data
- Another 7.67 gigabytes of Business and Customer photos
- Data currently stored as json and csv format
- A potential chance to win $5,000 cash prize from yelp!

Tools we intend to use:

- Python – pandas and numpy libraries
- Github – documentation, and version control
- SQL – data is organized in a format compatible with SQL relational databases. If performance requires, SQL may be used
- Tableau – if needed during data presentation and visualization

A more in depth look at the data:

6 large json files:

- Business.json – contains unique business id, name, geographic data, other classifiers
- Review.json – contains review information, including user and business id, useful, funny, cool, and stars awarded
- User.json – contains user id, review count and other review data
- Checkin.json – contains business ID, reservation date, user ID
- Tip.json – contains business id, user id, date, and a text field
- Photo.json – contains business id, photo classification, caption, photo id

What we intend to do:

- We intend to look for interesting patterns among yelp reviews and their authors. Are funny reviews considered to be more helpful? Do funny reviews usually have low review scores?
- We also intend to look for relationships between business attributes and the reviews that they receive.
- We also have date attributes, which may allow us to predict rating based on time of year.
- Open ended exploration. We intend to look at many different relationships in this dataset.
- We might combine external variables such as weather data or health of the economy to view correlations with review volume or sentiment.

Prior work

- Prior work by previous contest winners created a category predictor given review text
- A positivity estimator based on review text
- An automatic review generator based on a small initial text such as ‘They have the best’
- A json to csv converter
- Prior work can be found here: https://github.com/Yelp/dataset-examples

Next Steps:

- Data cleaning/integration: one of our json files is not importing correctly in our python environment. This happens to be the business table, which contains many important attributes. We currently believe that the problem exists only in one attribute, so we may just try to remove the single attribute if we need to. If this problem persists, there are still many relations to evaluate.
- Data has currently been mostly preprocessed into csv files to make importuning them into pandas easier, though we may later decide to import them into an SQL environment if that proves to be faster

Where our data comes from:

- We received our dataset from yelp (link below) distributed through their dataset challenge. Yelp is the sole owner of this dataset and therefore has intellectual property over any discovery or insight gained in this project.
- https://www.yelp.com/dataset/challenge