Stephen Paff

Yelp Dataset Challenge

Data Wrangling

6/7/17

**Description:** In this report, I describe how I initially cleaned up the data for the Yelp Dataset Challenge.[[1]](#footnote-1) I will outline issues I encountered with the data and how I addressed them. I used Jupyter for this process.

**Problem 1:** The datasets are extremely large, so after importing the json files onto the Python Notebook, the system experiences memory issues and ran extremely slowly.

**Solution**: I created a smaller version of each file, with only a hundred entries each for initial exploration into the data. I then exported them to a csv file so that they could be imported without having to use the larger, memory-draining original files.

**Problem 2:** I needed to develop a method to analyze the sentiment of reviews.

**Solution:** Using TextBlob, I added two new columns to the reviews, which calculated and stored the review’s polarity and subjectivity rating.[[2]](#footnote-2)

**Problem 3:** Each business has a stored list of business categories. Most businesses had several potential categories (e.g. restaurant, entertainment, nightlife, etc.]. To conduct analysis on the preferences by categories, I had to separate these. Some businesses, however, had no categories (in which case I filled them with an empty list).

**Solution:** I created a new DataFrame which listed out each potential business category and for each business, indicated whether the business matched that category. The column with the list was initially stored as a string (written in the format of the list), so I had to convert it to a list type. By first, auto-filling empty values as an empty list [], this prevented null pointer errors in that process. I then created new DataFrame files by merging these with DataFrames for reviews and for tips so that I could analyze reviews and tips by categories. I finally exported these merged DataFrames to a csv file for future use.

1. See this link for a description of the challenge: <https://www.yelp.com/dataset_challenge>. [↑](#footnote-ref-1)
2. For a basic description of Sentiment Analysis on TextBlob, see <https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis>. [↑](#footnote-ref-2)