Natural Language Processing, Sentiment Analysis and Yelp’s Best Burrito in America
Nick Campanelli

Problem:

Reviews and overall star ratings on Yelp often contain irrelevant information to the specific question a potential customer may have. In particular, I am interested in finding the restaurant with the best burrito in America. Selecting the Mexican restaurant with the highest overall star rating could be a good place to start, but necessarily ignores whether the majority of reviews praise the quesadillas, service, or atmosphere. Perhaps the restaurant serving the best burrito in America is nowhere near the top of the overall star rating because they don’t have a parking lot or vegetarian options.

Developing a classifier that more strongly weighs relevant reviews using NLP methods could help solve this problem. This is an invaluable method because, if successful, it will be extremely generalizable. The scope of this project is to find the best burrito in America, but other applications can benefit from similarly specified queries. Interested in yoga studios with the most popular teachers, laundromats with the fastest machines, or bars with the cheapest happy hours? The possibilities are limitless and hold great potential for identification of underrated establishments in all realms of business.

Data:

Yelp, the massive review platform, is kind enough to publicly offer reviews for a huge swath of businesses across the country and world. Split into several large JSON files, you can filter review information based on user, business, category, etc. The biggest problem is the sheer size of the data. There are around 5 million reviews contained for over 100,000 businesses. These datasets are acquired from Yelp.com.

Approach:

1. The first portion of this project will be working with the sheer size of the data. The full dataset is too large to hold in local memory so it will have to be parsed line by line to pull out only what I need. This will be reviews for every restaurant in the dataset that contains the ‘Mexican’ tag.
2. Using NLP tokenization and vectorization I will train a classifier that predicts Yelp’s star ratings. The classifier will attempt to use a custom built sentiment analysis score to predict how many stars the reviewer has given the restaurant. This rating system may be compressed (from 5 stars down to good or bad) to improve prediction. This sentiment analysis could potentially be informed by accessing the database of reviewers which contains features such as total number of reviews and average ratings.
3. When the classifier performs well enough, I will turn it loose on a predefined test set of all reviews that talk about ‘burritos’ (~60,000 reviews). Testing the classifier on this set will return a group of predictions that are incorrect. I will use these incorrect predictions to modify (boost) my custom sentiment analysis score in an attempt to accommodate

those reviews where the opinion of the burrito differs from the rest of the review (for good or bad).

4. To find the best burrito in America I will have to invent a new score to compare restaurants. This will have to include some sort of normalization for location, number of reviews, local competition and chain restaurants. Multiplying this normalizing score by the predicted star reviews from the sentiment analysis will give me a ranking of the best restaurants serving burritos in America.
5. Visit #1 on the list.

Deliverables:

At the end of this project I will have code that is (hopefully) relatively generalizable to different queries, a slide deck to present my findings, a medium blog post to describe my methods, and a good starting point for my burrito bucket list.

,