Max Grossenbacher
9/6/17

- Project Proposal -

Project Title: Content Analysis of Yelp Reviews

What questions am I trying to answer?

I am trying to understand how to use NLP on yelp reviews of restaurants to find underlying topics and sentiment. I will use K-means clustering, sentiment analysis, NMF, and neural nets (seq2seq) to classify yelp reviews of restaurants based on rating. Using this information, I will try to reverse engineer these yelp reviews to predict a typical review of a certain rating for a particular restaurant.

What are the data and do you have them? Have you looked at them?

I will be using the Yelp Challenge Dataset made available by Yelp for academic purposes. This file contains 5 separate json files containing business, checkin, review, tip, and user datasets. I will be focusing my project on the review data. I have been able to import and pickle all the datasets. There are 4,153,150 written reviews of 144,072 businesses. Additionally, the review dataset contains rating given for the review and the business ID, which I can match with the corresponding business ID in the business dataset. The business dataset contains the number of reviews per business as well as the average rating of the business. There are 67,437 businesses with at least 10 reviews and 106,691 businesses with at least 5 reviews.

What is the MVP?

I will clean and relate the relevant data. First, I will vectorize the yelp reviews and run NMF, k-means clustering and sentiment analysis in order to determine if there are any latent topics or sentiments conveyed in reviews for different restaurants of different ratings. Furthermore, I will build word clouds that can illustrate common/ important words that describe a particular restaurant, as well as, word clouds that illustrate common words/sentiments dependent on a particular rating.

* I do have a question on whether I should begin with a single restaurant and build models then look to scale them or start with a whole category of business?

MVP +

I will build a seq2seq neural net that will be able to predict a “typical” review for a particular establishment based on their average rating and typical review. Additionally, I hope to expand this to a “typical” review for an establishment given a particular rating.

MVP ++

I will be using a seq2seq neural net to create a content-based spam filter in order to flag possible spam/fake yelp reviews. Spam reviews are reviews that are “faked” /untruthful. Spam reviews can be used to unfairly inflate ratings on yelp or to defame business. Using the “reverse” engineered predictions created using the seq2seq neural net, I hope to find a probability that some review maybe spam based on the sentiment, language and rating of a particular review for a particular establishment.

Why do I think this project is important?

Spam has become an issue of reliability for many websites such as Yelp, which depend on honest user-based reviews of establishments to inform other consumers. If establishments are unfairly paying for higher reviews or defaming their competition, Yelp users may lose trust in the Yelp app. Based on my current research, many spam filters focus on the behavior of users to identify possible spam. For instance, if a user writes more than the typical number of reviews in one day, they may be identified as a spammer. However, this type of spam filtering requires spam to have be written and posted many times before a spammer can be identified, and the spam can be removed. If a reliable content-based spam filter could be developed, ideally, one would be able to identify and flag possible spam before it becomes prevalent.