Tyler Waltze
Professor Terzi
CS591 - Data Mining
April 19, 2015
Initial Description of Dataset

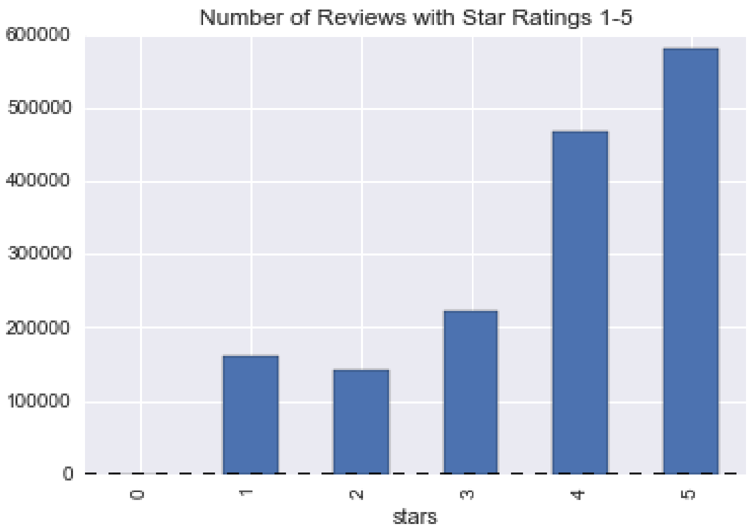
For the project, I am working with Yelp's "Dataset Challenge" data. A breakdown of the data included:

- **1.6M** reviews
- **500k** tips
- **366k** users
- **61k** businesses
- **481k** business attributes, e.g., hours, parking availability, ambience.
- Social network of **366K** users for a total of **2.9M** social edges.
- Aggregated check-ins over time for each of the **61K** businesses

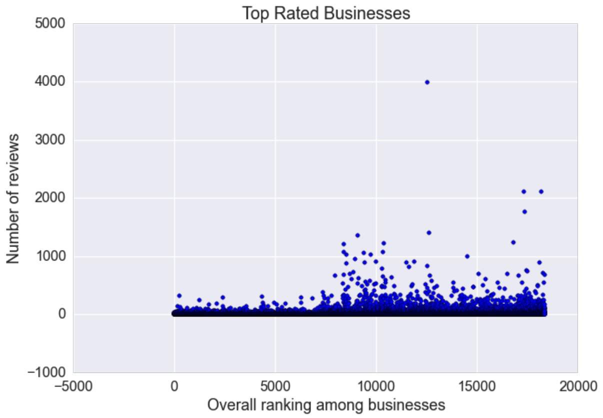
The reviews are what I am focusing on. The goal is to break reviews down into a set of important attributes and determine how words or topics correlate with a review's rating. Reviews themselves contain several pieces of useful data for this, most importantly a review's "star" and "text" attributes. Both of these are required attributes for each review.

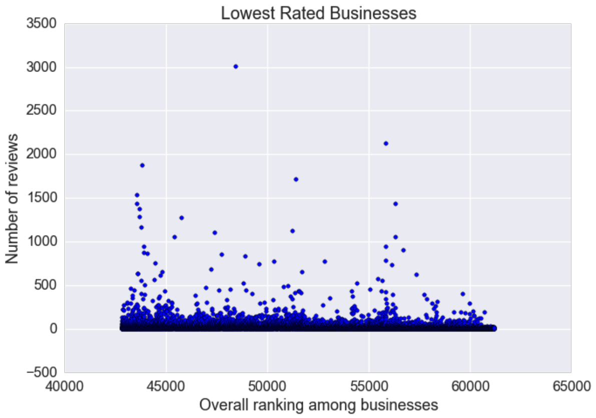
- star: An integer ranging from 1-5 representing a user's rating of a business.
- text: A string representing a user's review of a business.

The first thing I looked at was the general distribution of of ratings.



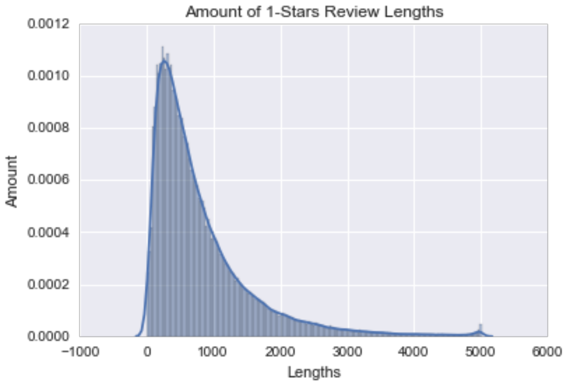
The above graph shows that it is far more common for people to leave positive reviews than negative reviews. Anecdotally, it is surprising to see so few negative reviews, as it might be assumed that equally strong negative and positive feelings of a place would drive an equal number of people to write about their experiences. This could remain the case though, and such disparity could be attributed to fewer people visiting, and thus reviewing, places that consistently have poor ratings. I checked this in the below graphs of the top/bottom 30%.

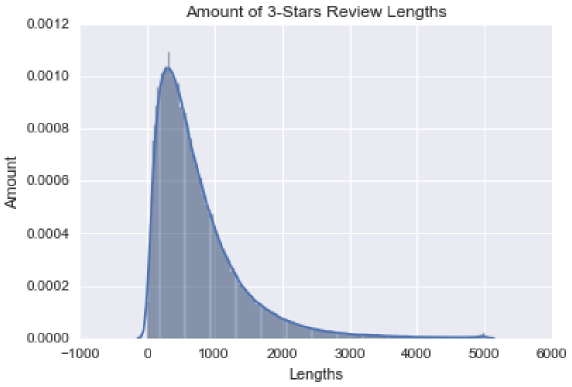


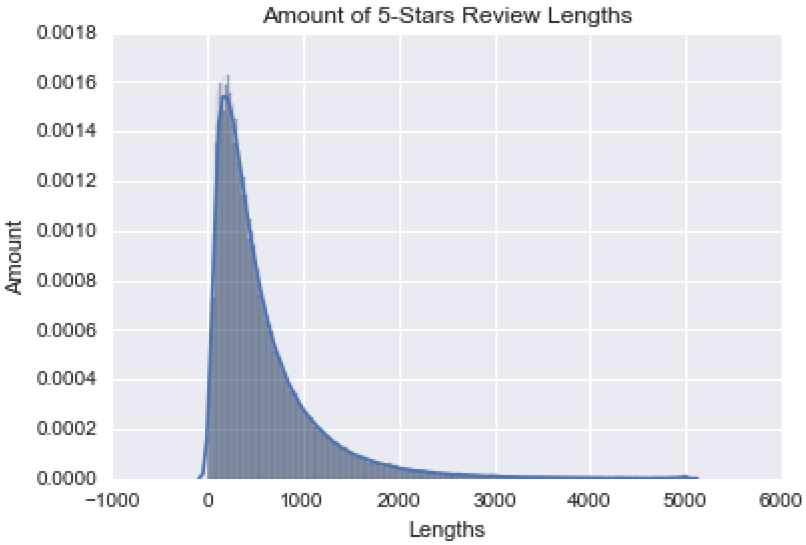


The above two graphs are based on Yelp's business data, which contains the number of reviews (an integer) and the average rating (a float in increments of 0.5) for an individual business. As the above shows, it is true that places given lower average ratings have fewer people review them in general.

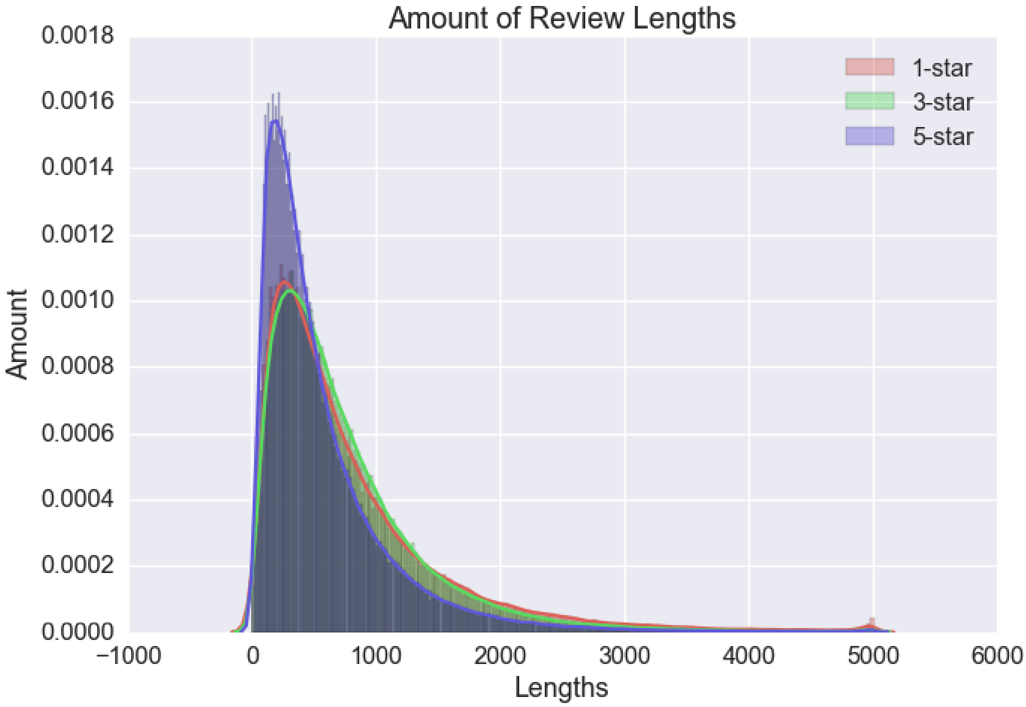
Following this, I checked the relationship between a review's length and its rating.



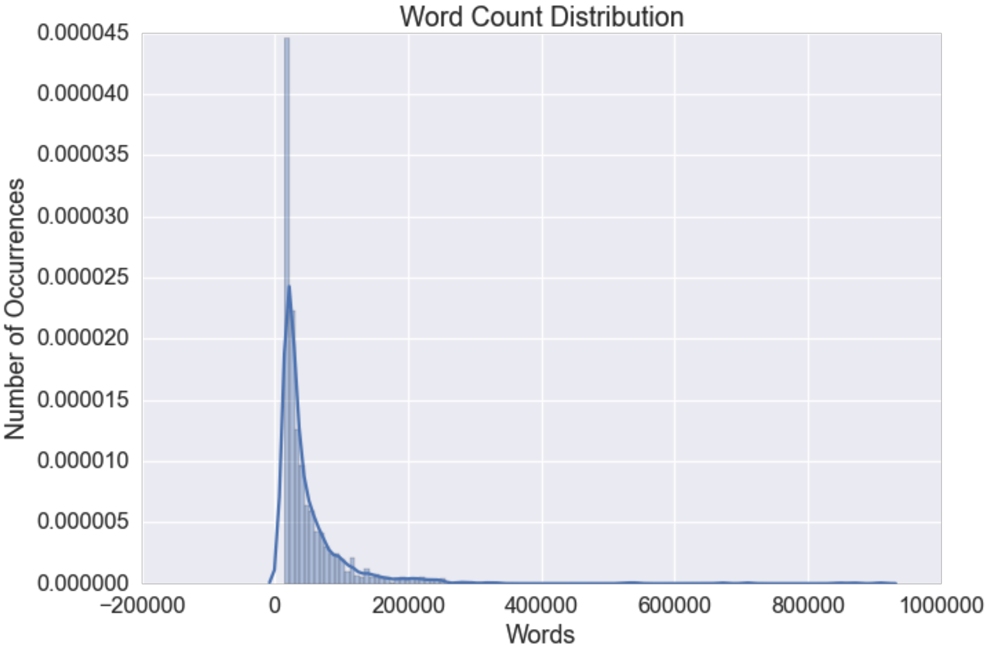




This showed that there is a fairly consistent distribution in review length among the different ratings. This was surprising because one might expect that those who who feel strongly about a place (a very high, 4-5, or very low, 1-2 star rating) would spend more time on their review and have more to say. This is not the case.



The final thing I looked at was word use, particularly how frequently words were used.



Word breakdown:
Number of words: 80618734
Number of unique words: 341183
Number of words which occur only once: 170663
Number of words which occur 5 or fewer times: 255496
Number of words which occur 100 or fewer times: 318729
Number of words which occur 10000 or fewer times: 339792
Number of words which occur 10000 or more times: 1391

Out of the top 10 most frequently used words, 3 of them (place, food, and service) describe a particular attribute of a business. 4 of them (good, great, nice, and best) describe a business positively.

The goal with this data is to see if it is possible to guess a review’s rating based on its content, and to break down a business’ average rating into separate categories rating its attributes, such as food, service, etc. This will be accomplished with more intelligent topic modeling using gensim and sentiment analysis.