Capstone Final Report: Predicting Review Ratings for Yelp Automotive Reviews

Java Starchild

October 31, 2015

Introduction

Here is a description of the question/problem and the rationale for studying it. Customer reviews on site such as Yelp have a profound impact on the chances of success of any business. Automotive customers look for a complete and satisfactory experience regarding quality of service and often seek the opinion of patrons when they are choosing a place for their next service. Learning which topics are the most frequent among customer reviews and how they associate to a positive or negative rating can help business improve their services and have a better chance of succeeding.

To achieve this goal, in this report I explore some latent topics in a corpus of Yelp reviews for Automotive Businesses.

Methods and Data

Here I describe how I you used the data and the type of analytic methods that are used. The data used here is part of the Yelp Dataset Challenge.The dataset consists of a set of JSON files that include business information, reviews, tips (shorter reviews), user information and check-ins. Business objects list name, location, opening hours, category, average star rating, the number of reviews about the business and a series of attributes like noise level reservations policy, etc. Review objects list a star rating, the review text, the review date, and the number of votes that the review has received.

I have sampled 10K reviews to allow reasonable run time and filtered the business by category to keep only those businesses in the Automotive category (2965) and reviews related to those businesses (186). The texts from Automotive reviews will form the corpus for this analysis.

I have processed each of the reviews to build a bag of words language model. To create this model I preprocessed each document in the corpus as follows: remove non-writable characters, strip extra white spaces, lower case, remove punctuation, remove numbers, stemming and stop words removal.

After that, each text was tokenized into uni-grams, and the uni-gram frequencies were counted and stored into a document-term matrix of counts. Term counts across all the corpus showed a typical normal distribution. I kept the most frequent terms that, summing all their frequencies, accounted for about 90% of the total number of words in the corpus. The resulting vocabulary has 1592 words.

Results

Here I describe what I found through my analysis of the data. To discover latent themes in our corpus, I run a Latent Dirichlet Allocation algorithm (LDA) using the document-term frequencies matrix as input. To estimate the model parameters we used a Gibbs sampling with a burn-in phase of 1000 iterations and later the distribution was sampled every 100 iterations during 2000 iterations. I tested other approaches (LDA with VME parameter estimation and a Correlated Topics Model) but the topics obtained were less clear than the ones resulting from LDA with Gibbs sampling.

I decided to use 20 topics by fitting a three-segment linear regression and selecting the number of topics about the middle of the second segment. This method, similar to the elbow rule, seeks to get a simple model with enough flexibility. To select the number of topics (k), I ran LDA on 20% of the documents in the corpus (37) using different k values. Figure 1 shows the log-likelihood for a range of values for k.

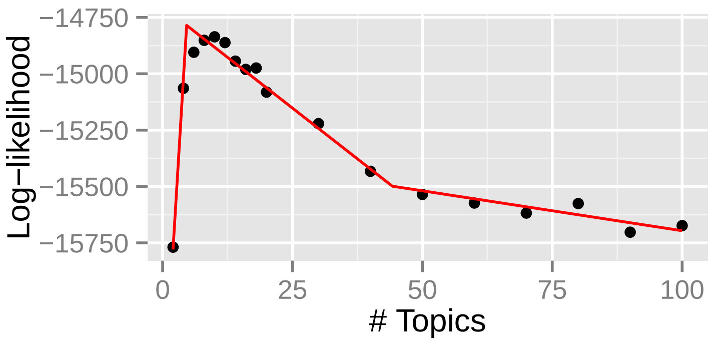

Figure 1: Topic model log-likelihood for a range of topic numbers. The red line is a result of fitting a three segment linear regression to the data.

First, I will examine the results of fitting a topics model to the whole automotive reviews corpus. Second, I'll show the results of running the same analysis over two corpora, one for positive automotive reviews and another for negative automotive reviews.

The overall topic model lists several topics about customer experience. But good and bad experiences are mixed because the corpus includes both positive and negative reviews. I have explored the topics related to positive and negative ratings and can see (not shown) that positive reviews (stars >=3) dominate over negative reviews (stars <3).

I have fitted two topic models (one for positive reviews and another for negative reviews) with 20 topics each and following the same methodology that I used to compute the overall topic model. In general we can see some of the overall topics also appear in these two new models and we get a finer grain topic distribution.

Discussion

Here I explain how to interpret the results of my analysis and what the implications are for the question. I have explored some latent topics in a corpus of Yelp reviews for automotive businesses. For that, I have fitted to the corpus a topic model using LDA with Gibbs sampling. The topics found display themes related to different customer experience.

I have further explored the customer experience topics by splitting the corpus in two corpora, one for positive experiences and another for negative experiences and fitting a topic model to each corpus. Many themes that appear in the overall topic model also appear in the new topic models. The new models also show a finer grain decomposition of the customer experience. The top topics found in the positive reviews can be used to improve automotive business success.

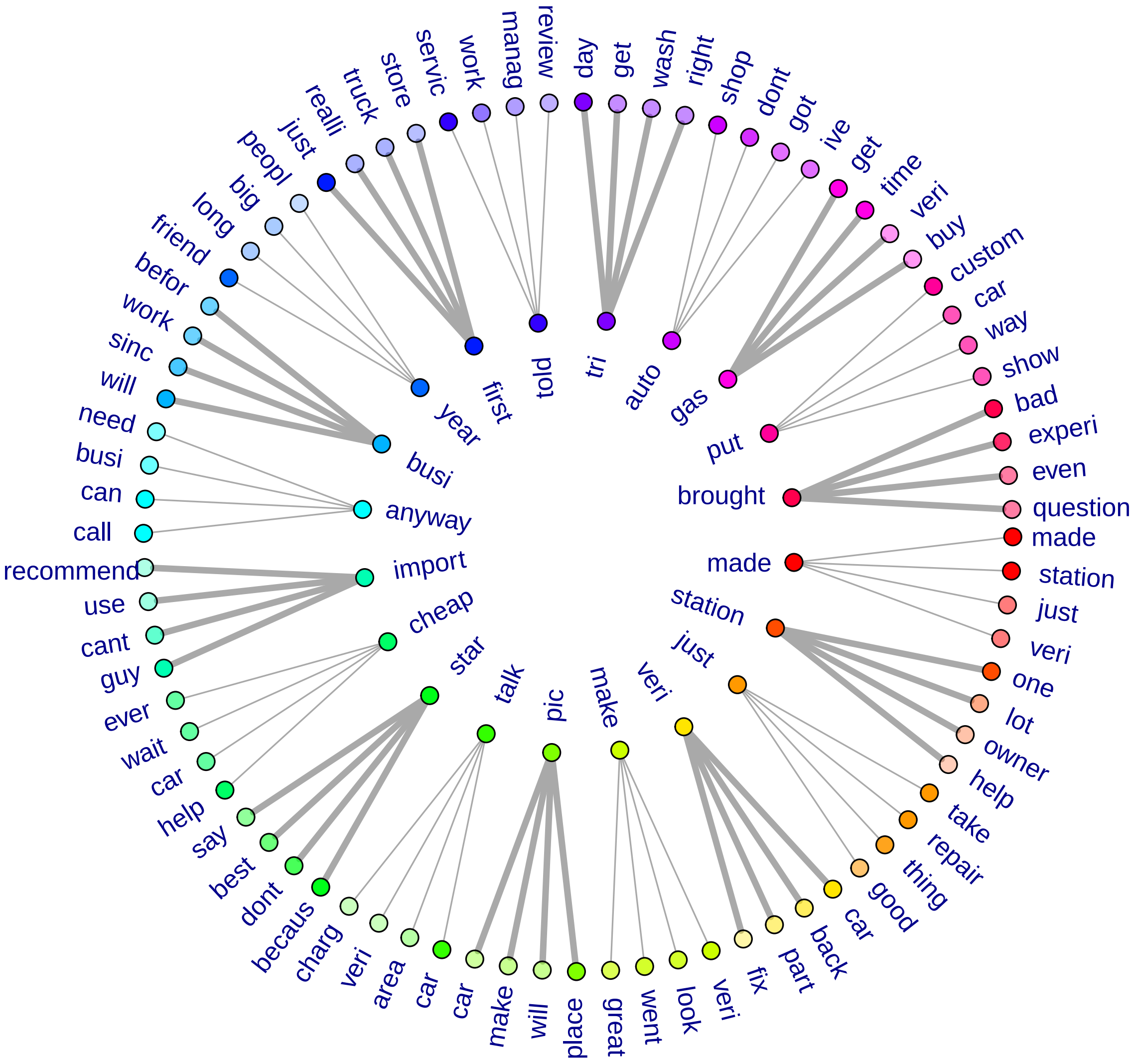

Figure 2: A topic model for Yelp automotive reviews.

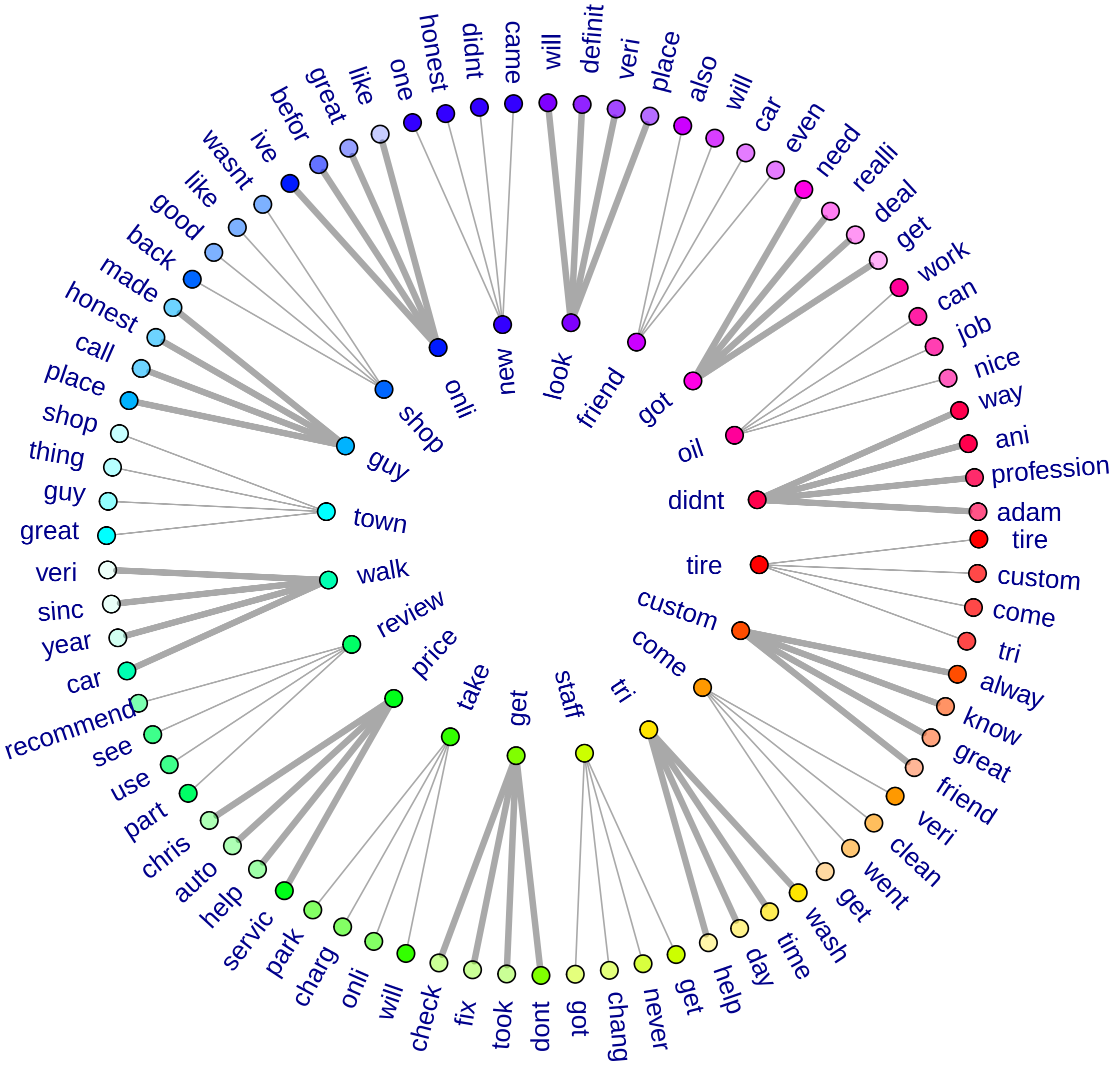

Figure 3: A topic model for positive Yelp restaurant reviews (stars greater than or equal 3)

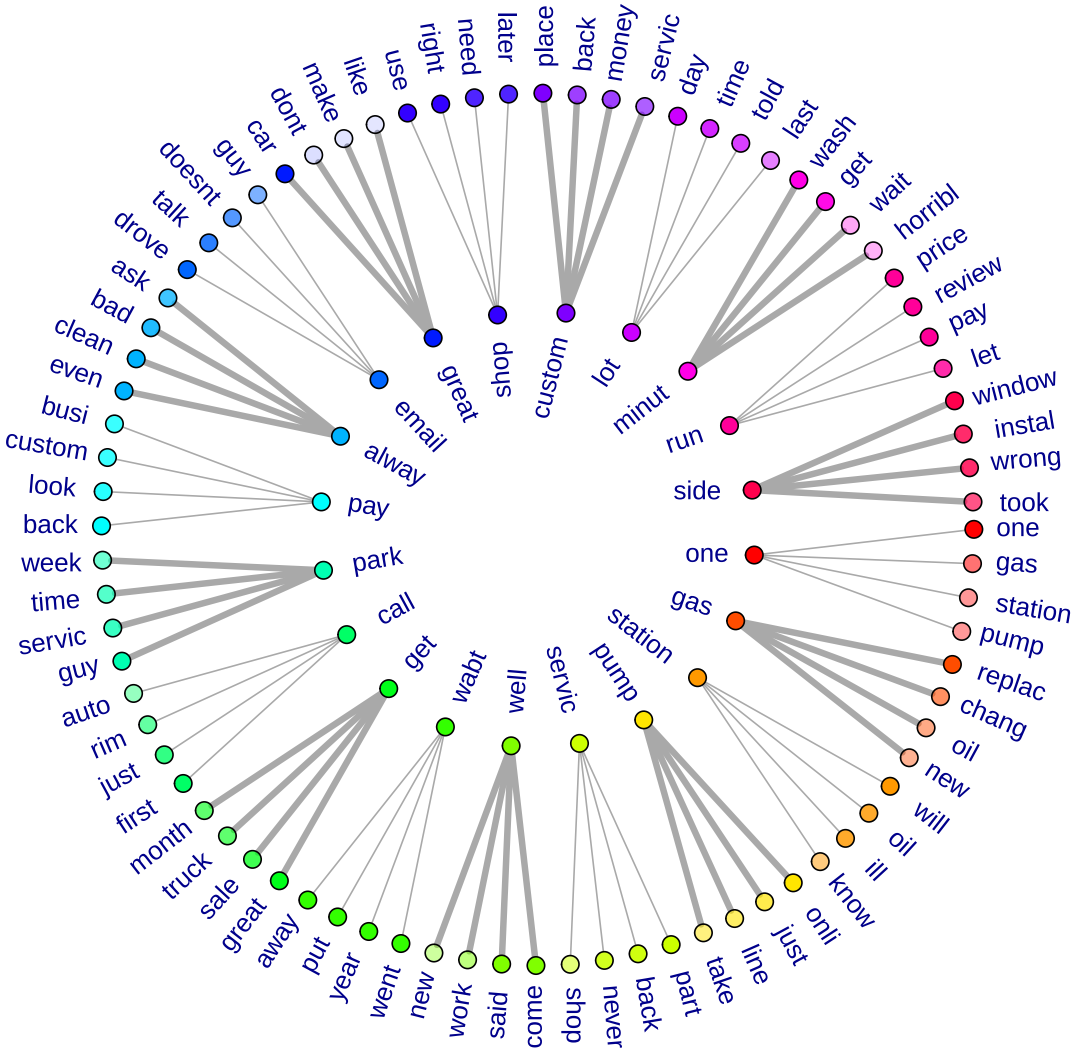

Figure 4: A topic model for negative Yelp automotive reviews (stars less than 3)