To Write a Useful Yelp Review in Charlotte, NC, Say "Cheese". Or Write a Longer Review.

Michael Green

November 20, 2015

Introduction

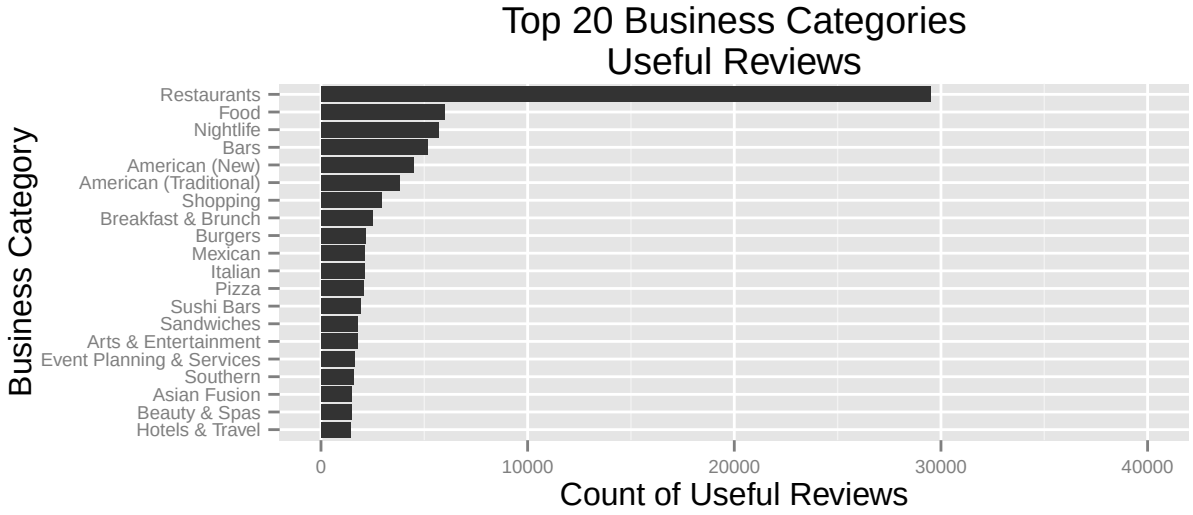
47% of the Yelp reviews in the data set provided inYelp Dataset Challenge number 6 have at least one vote for "useful". I wanted to use topic modeling to determine if content or words used in the text of the reviews could help determine what makes a review useful. To reduce the dataset to a more manageable size, and to account for differences in regional preferences, I chose to use reviews written only for businesses in the Charlotte, NC area. The results of this exploration, documented below, show that topic modeling and many other statistics show very little difference between the useful and not-useful reviews. However, examination of the document term matrices created to perform topic modeling show that there are differences in the variety of words and average numbers of words in a review that may correlate to usefulness. In addition, the word "cheese" occurs 416 times in a sample of 2000 useful reviews and only 197 times in 2000 reviews that were not marked useful.

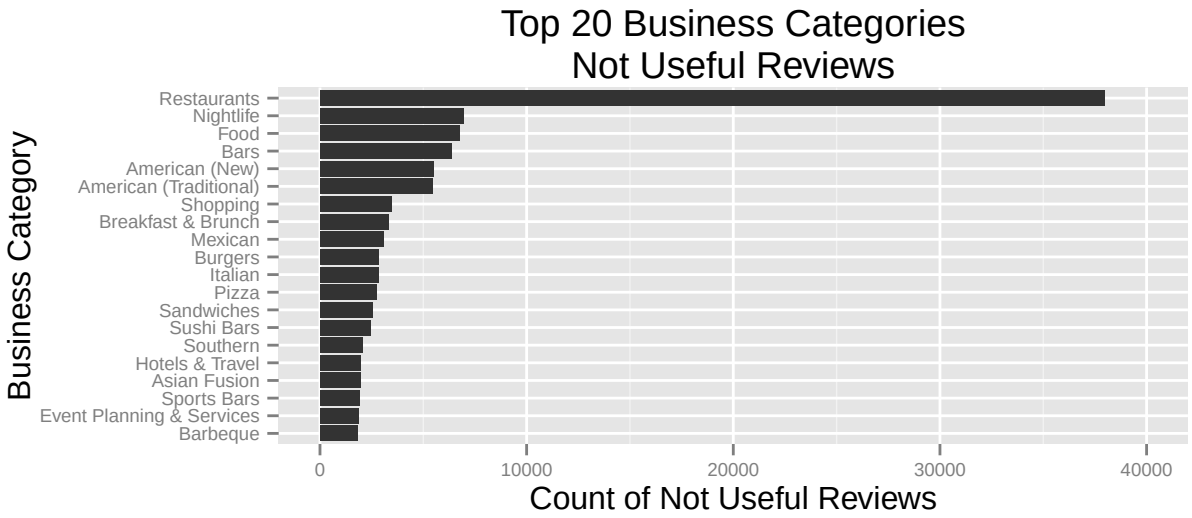
Methods

The data set used for this paper can be downloaded here. The R code I created to explore this data and create this report can be found at my github page.

There are several files in the yelp data set. I used the yelp_academic_dataset_business.json file and the yelp_academic_dataset_reviews.json file. In order to select review data from Charlotte, NC, I merged latitude and longitude values from the business data with the review data using business id from the review data to identify the correct values. Then I selected data from Charlotte (35.2269° N, 80.8433° W) by choosing all records with latitudes between 34° and 35° and longitudes between -82° and -90°.

I then separated the Charlotte data into useful and non-useful reviews. Exploring the useful and non-useful review sets revealed little difference in the distribution of star ratings. Also, when merging business category information into the review datasets, the frequency distribution of categories was similar. Most reviews pertain to restaurants and food. The top 20 categories in each review set are shown below:





No useful insight into what makes a review useful yet. Perhaps topic modeling will be illuminating.

A topic model is a statistical model that helps to discover abstract topics in a body of documents by exploring relationships between the probability of occurrence of the different words in the documents. The output of a topic model is a pre-determined number of topics and the words that make up each topic, ranked by relevancy. A document may be described by one or more of the topics generated. More on topic modeling can be found here. I used the topicmodels package in R to model the yelp review data. Documentation for this package is here. I created two models, one for the useful reviews and one for the not-useful reviews.

In order to create the topic models using the limited capacity of my eight year old laptop computer, I randomly sampled 2000 rows of data from each set of reviews. From each sample set, I created a text corpus and a document term matrix using the functions in the tm package in R. I experimented quite a bit with the document term matrix, exploring ways to reduce the vocabulary sets and remove words common to most documents. One method of reducing the vocabulary is term frequency – inverse document frequency (tf-idf). Tf-idf weights each term proportionally by the number of times it appears in a document offset by the frequency of the word in the entire document set, thus eliminating both frequently used words and seldom used words. However, when using this technique, many subjective words like awesome, amazing, love, delicious, etc, were eliminated from the useful review set, and were not eliminated from the not-useful set where they do not occur in as large a proportion. I initially came to the conclusion that useful reviews were more objective because the models I generated did not have these subjective words in them. This report was originally going to be titled "Objective Reviews More Useful". However after some checking, I realized my mistake. The opposite may actually be true. Finally, I settled on increasing the minimum word length to include in the modeling at 5 letters. This removed words like food, good, place, and nice, which were very common in both sets and appeared in almost every resulting topic.

The key parameter to choose in topic modeling is k, where k is the number of topics to model. Choosing k topics can be attempted objectively, using model fit parameters such as perplexity, or more subjectively by simply examining the terms in the topics the model produces to see if they make thematic sense. I wrote a script that computed perplexity, or the degree in which data fits the model, for k = 10,20,30,40,60, 80, and 100. The results suggested k = 60 would minimize the perplexity. However, an increase in k resulted in a significant increase in compute time for the topic model. A more subjective review of the results did not reveal any significant topical insight to be gained with higher k. A previous yelp challenge winner had used k = 20 in creating topic models from review text. See here. So I ultimately did the same.

The topic model algorithm I chose to use was the Latent Dirichlet Allocation (LDA) model. The LDA model produces topics for which each document is a mixture of those topics. It is described here. I also explored using the correlated topics model (CTM) function in the R topicmodels package. The CTM took about twice as long to compute on my laptop and had larger perplexity.

Results

The topics and the ten most relevant words for each are presented in the tables below. The topics seem to follow what can be expected from the distribution of business categories explored above. They are mostly food and restaurant related. The topics seem mostly positive, as there are a lot of "great" and "place" in both sets. Topics 17 and 14 are the highest probability fit for the most useful reviews. They are the leading topic for 158 and 138 reviews, respectively, out of 2000. Topic 17 seems to be lunch related. Topic 14 has to do with good service.

Table 1: Top Ten Words in Each Topic: Useful Reviews (continued below)

X	Topic.1	Topic.2	Topic.3	Topic.4	Topic.5
1	burger	better	place	charlotte	really
2	sushi	store	really	great	cheese
3	fries	indian	order	service	great
4	place	great	sweet	fresh	place
5	burgers	place	since	place	always
6	great	really	little	really	service
7	ordered	thing	restaurant	little	delicious
8	rolls	taste	great	bagel	bread
9	order	location	charlotte	bagels	order
10	lunch	though	people	location	restaurant

Table 2: Table continues below

Topic.6	Topic.7	Topic.8	Topic.9	Topic.10
sauce	store	service	place	steak
brisket	selection	minutes	night	cream
chicken	place	asked	great	delicious
place	great	experience	really	dessert
cheese	prices	order	people	ordered
flavor	always	waitress	music	restaurant
charlotte	items	never	service	sweet
pulled	stores	another	definitely	chocolate
really	pretty	table	charlotte	cheese
barbecue	around	first	pretty	place

Table 3: Table continues below

Topic.11	Topic.12	Topic.13	Topic.14	Topic.15
place	place	place	place	pizza
great	great	coffee	great	crust
charlotte	little	brunch	staff	really
people	really	great	friendly	service
little	always	pretty	people	order
parking	outside	lunch	always	hotel
wings	first	buffet	around	ordered
always	people	really	service	first
service	times	restaurant	really	pretty
really	front	little	every	water

Topic.16	Topic.17	Topic.18	Topic.19	Topic.20
restaurant	chicken	great	salad	place
place	lunch	service	ordered	great
great	sandwich	place	husband	tacos
ordered	place	style	place	salsa
order	salad	always	table	mexican
service	delicious	really	service	really
dinner	little	people	really	little
shrimp	great	korean	steak	charlotte
delicious	cheese	asian	restaurant	chips
dining	ordered	dishes	server	pretty

Table 5: Top Ten Words in Each Topic: Not Useful Reviews (continued below)

X	Topic.1	Topic.2	Topic.3	Topic.4	Topic.5
1	place	always	parking	service	place
2	great	place	great	customer	great
3	uptown	sushi	place	never	service
4	charlotte	chinese	people	charlotte	hotel
5	always	fresh	really	business	breakfast
6	pretty	store	inside	manager	night
7	stars	really	space	airport	really
8	movie	shopping	little	called	charlotte
9	really	service	outside	around	pretty
10	people	definitely	located	first	atmosphere

Table 6: Table continues below

Topic.6	Topic.7	Topic.8	Topic.9	Topic.10
place	burger	great	great	cream
onion	fries	staff	place	bring
great	order	friendly	service	staff
rings	always	always	steak	place
favorite	lunch	around	little	really
really	people	people	barbecue	around
something	service	service	taste	chocolate
portions	great	first	restaurant	always
sweet	place	anyone	sunday	level
chicken	really	times	brunch	selection

Table 7: Table continues below

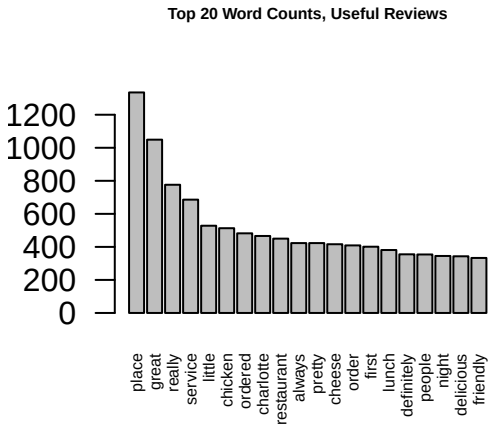
Topic.11	Topic.12	Topic.13	Topic.14	Topic.15
experience	great	ordered	place	service
really	sushi	delicious	salad	sandwich
great	little	restaurant	great	beers
pretty	pizza	great	really	lunch
items	place	spicy	family	great
place	service	chicken	always	fries
service	better	really	happy	minutes
selection	average	salad	equipment	waitress
servers	stars	amazing	patient	definitely
location	right	service	staff	place

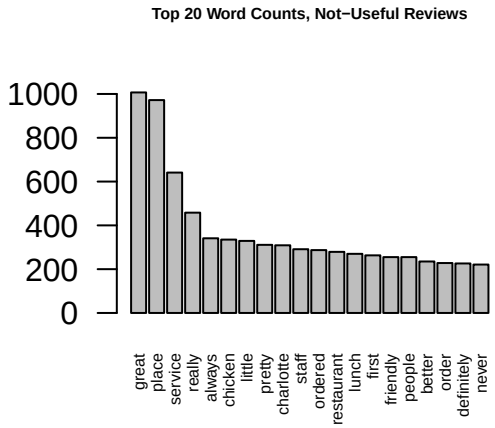
Topic.16	Topic.17	Topic.18	Topic.19	Topic.20
wings	great	chicken	minutes	coffee
place	place	sauce	ordered	place
sandwich	pizza	fried	order	burger
great	service	cheese	place	service
usually	awesome	salad	table	better
pretty	little	lunch	never	bread
starbucks	definitely	ordered	something	really
always	tacos	shrimp	another	turkey
really	night	little	restaurant	breakfast
seating	fresh	delicious	first	delicious

Topics 17 (195) and 5 (181) are the highest probability fit for the not-useful reviews. Topic 5 of the not-useful documents is interesting because it seems to be negative, with words like never, manager, customer, service. Though writing about lunch and good service might improve your odds of writing a useful review, I was not able to find significant discriminating factors between useful review topics and not-useful review topics.

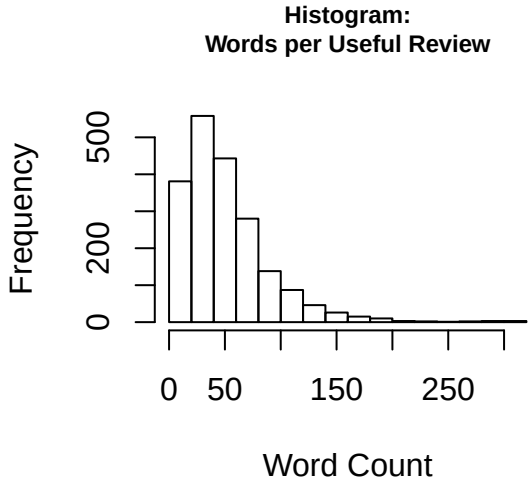
Discussion

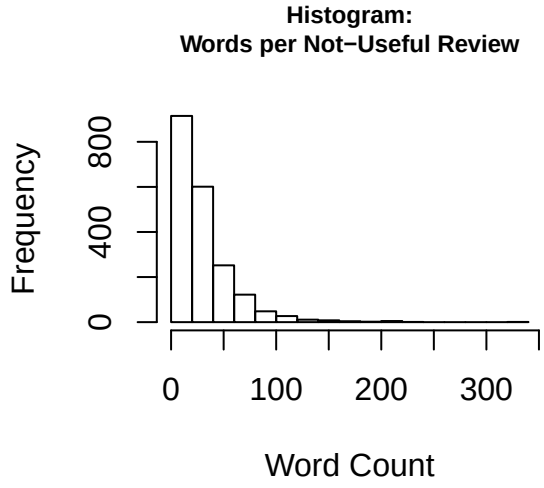
The results of the topic modeling itself did not produce the insight I had hoped into why some reviews are deemed useful while others are not. However, the exploration of the data, the creation of the document term matrices, and examination of the word counts and frequency counts do reveal some clues.





The top 20 most frequent words five letters and longer are about the same in each review set. One difference is "cheese" appears in 12th place in the useful review set. Another is that the terms "staff" and "never" are not among the top 20 in the useful review set. But a more macroscopic examination shows top word counts are higher in the useful review set. Morover, not only are the top words used more, 13508 words five letters and longer appear in the useful document term matrix. Only 9364 words appear in the not useful matrix, a 44% bigger vocabulary in the useful review corpus from the same sample size of reviews. This disparity in word count is also seen in the summary of the count of words five letters or longer per review.





Useful reviews have a median of 43 words and average of 52 words per review. Not useful reviews have only a median of 22 and a mean of 31 words five letters or longer. The median and mean word counts are almost double in the useful reviews. Therefore we can conclude that a greater variety of longer words and longer review text correlate to useful yelp reviews in Charlotte, NC.