COMP-3006 Project Final Report
Brian Miller (DU ID: 873601817)
Date: 03/11/2022

Introduction:

For my project I performed sentiment analysis on the two main donut competitors in Portland, OR: Blue Star Donuts and Voodoo Donuts (figure 1). Before starting my masters here at DU I lived in Portland for two years, and most people I have met have a strong opinion about which is better. I used Yelp review data to explore the customers' top thoughts about each company, which donut flavors were best, and whether the wait in line for their donut affected their experience.

Before starting this project I had no prior experience with NLP, sentiment analysis, sklearn, or pandas. Learning the basics of these packages has been as rewarding as the results I have produced using them.

Gathering the Data:

I first downloaded the full Yelp academic dataset which contains millions of reviews for thousands of businesses (figure 2). The data is all stored in very large JSON files. I then used the file of listed businesses to determine how many reviews were available for each of the two companies. There were eight donut shops for Blue Star Donuts, and two shops for Voodoo Donuts that I could pull Yelp reviews from (figure 3).

Sadly the full dataset of reviews was too large to load into memory, so I had to first split the JSON file into five smaller pieces. Then I ran my extract_relevant reviews.py which sequentially loaded each of the pieces and extracted the Yelp reviews that were relevant to my analysis (figure 4). Each Yelp review has a 'business id' field, and this was how I determined which donut shops were the ones I extracted reviews from. The output at this point is two pickle files, one for each company, each containing a pandas dataframe of the Yelp reviews.

Note to the grader: I am not expecting the grader to download the full Yelp dataset and perform the steps described above. I have uploaded the two pickle files to canvas, all scripts that use the pickle files can be run on any machine (given the proper py module dependencies). All scripts can be run in your text editor of choice (no terminal usage is required).

Exploratory Data Analysis:

With the data ready for use I then wrote eda.py, so that I could get a grasp on what the data looked like. Voodoo Donuts has been in Portland since 2006, while Blue Star has only been there since 2013. This is a large part of why Voodoo has twice the number of reviews as Blue Star (figure 5). The number of reviews per quarter for both companies peaked around 2015 and has been on the decline since, with a noticeable drop off in reviews around when covid hit (figure 6). Almost half of Blue Star's reviews are 5 stars, while for Voodoo a 4 star review and a 5 star review are almost just as common (figure 7).

Interestingly, even though Blue Star has been in Portland for less time it seems to have had consistently higher reviews than Voodoo (figure 8). Finally, I wanted to explore the review sentiment. For this I used the academic open source NLTK package (natural language toolkit). NLTK uses a large lookup table of positive/negative word assignments for the entire English language. Scores can range from -1, the most negative, to 1, the most positive, with 0 being neutral. Not surprisingly the sentiment scores followed closely with the Yelp star review scores (figure 9).

Sentiment Analysis:

For the sentiment analysis I decided that it would be best if I worked in an ipython notebook. I kept my helper functions in helper_functions_sa.py, while the actual analysis was done in sentiment_analysis.ipynb. First I did some quick data filtering: drop reviews that were not in english, and dropped stop words. Stop words are common words that usually don't convey sentiment, such as: I, am, me, being.

The most basic version of any sentiment analysis is to look at word frequencies. Word clouds are more pleasant to look at than long lists, so I made one for each company (figure 10). While not exceedingly informative, we do start to see some interesting words pop out. 'Line' is in both, but more pronounced for Voodoo. Blue Star has some of its flavors presented: 'bourbon basil', 'blueberry bourbon'. The same goes for Voodoo: 'maple bacon', 'maple bar'.

The next step was to classify the reviews into positive and negative. After looking at the distribution of sentiment scores for each review (figure 11) I decided that a cutoff of 0.75 or higher would capture the extremely positive reviews. Thus moving forward all 'positive reviews' have a star rating of 4 or 5, and a sentiment score of 0.75 or higher. While all 'negative reviews' have a star rating of 1, 2, or 3, and have a negative sentiment score. The counts for each of these subsets are shown on (figure 12).

Bigrams are two adjacent words that are frequent in the text. I next constructed bigrams for each of the four subsets: Voodoo positive, Voodoo negative, Blue Star positive, Blue Star negative (figure 13). For both companies, the positive bigrams helped explain which flavors were the best (please note that 'old dirty bastard' is indeed a donut flavor at Voodoo). Looking at the negative review subsets, both companies unsurprisingly have the bigram 'customer service', but in addition to that only Voodoo has the bigrams 'wait line', 'long line', and 'tourist trap'. This goes hand in hand with the fact that Voodoo only has two locations, while Blue Star has eight.

The final sentiment analysis approach I applied was a word prediction analysis. The benefit to this type of method is that it can make word associations in sentences for words that are not directly next to each other. I implemented a basic 3-layer ‘skip-gram’ neural network model that given a target word will output related words and their probabilities. Words could form a target relation if they are at max 10 words apart, 2000 unique predicted words was the search space, and words from the corpus that are stop words or have a frequency of 1 were dropped. While the results for ‘flavor’ were not that insightful, the results for [‘line’, ‘wait’] gave relevant word predictions for Voodoo’s negative reviews, while they were inconclusive for Blue Stars’ negative reviews. This suggests that indeed, many Voodoo reviews mention waiting in line (figure 14).

<u>Conclusion:</u>

If you are in Portland and are craving a donut, this is what you need to know. Voodoo may be the best place for you if you are from out of town, or are looking for a novelty factor. Go to one of the two shops, wait in line for longer than you’d like, and order a maple bacon bar, voodoo doll, or an old dirty bastard (again, it’s a flavor). If you are looking for a quick donut pickup, go to one of the eight Blue Star shops and order the blueberry bourbon, bourbon basil, or the passion fruit.

Figure 1: Donuts!

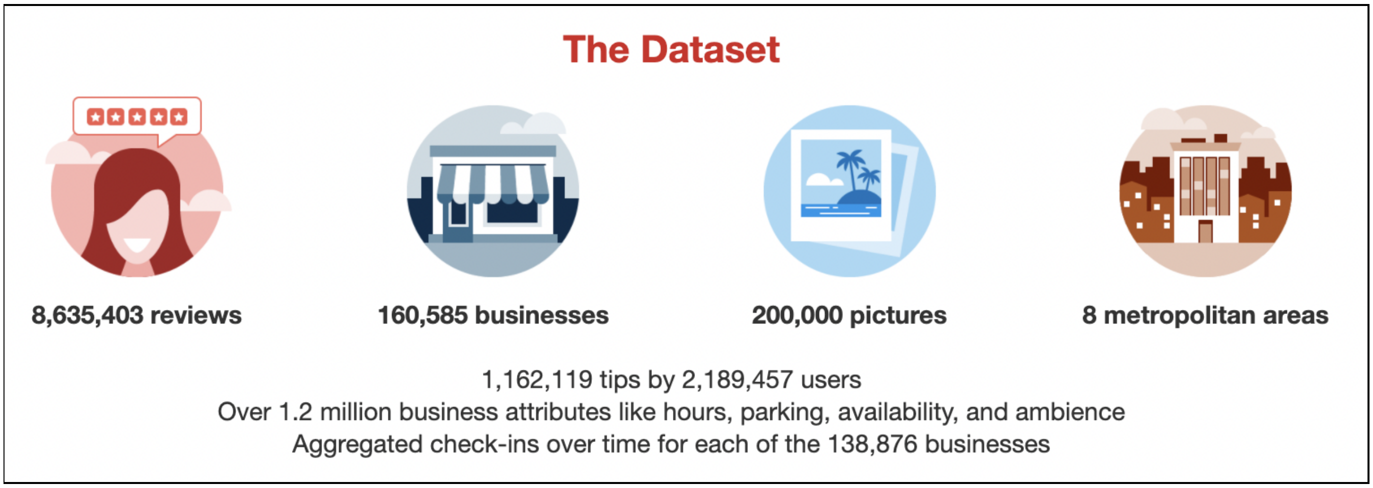

Figure 2: Overview of the Yelp Academic Dataset
Link to the full Yelp dataset: https://www.yelp.com/dataset

name	address	city	stars	review_count
Blue Star Donuts	1155 SW Morrison St, Ste 102	Portland	4.0	4011
Blue Star Donuts	7709 SW Capitol Hwy, Ste 111C	Portland	4.5	11
Blue Star Donuts	921 NW 23rd Ave	Portland	4.0	376
Blue Star Donuts	14985 SW Barrows Rd, Ste 127	Portland	3.5	87
Blue Star Donuts	3325 SE Division St, Ste 1	Portland	4.0	802
Blue Star Donuts	3753 N Mississippi Ave	Portland	4.0	403
Voodoo Doughnut - Old Town	22 SW 3rd Ave	Portland	3.5	9185
Blue Star Donuts	7000 NE Airport Way	Portland	4.0	214
Blue Star Donuts	672 S Gaines St, Ste 2	Portland	4.0	112
Voodoo Doughnut - Davis	1501 NE Davis St	Portland	4.0	1960

Figure 3: Overview of which Blue Star Donuts and Voodoo Donuts locations were used in this analysis.

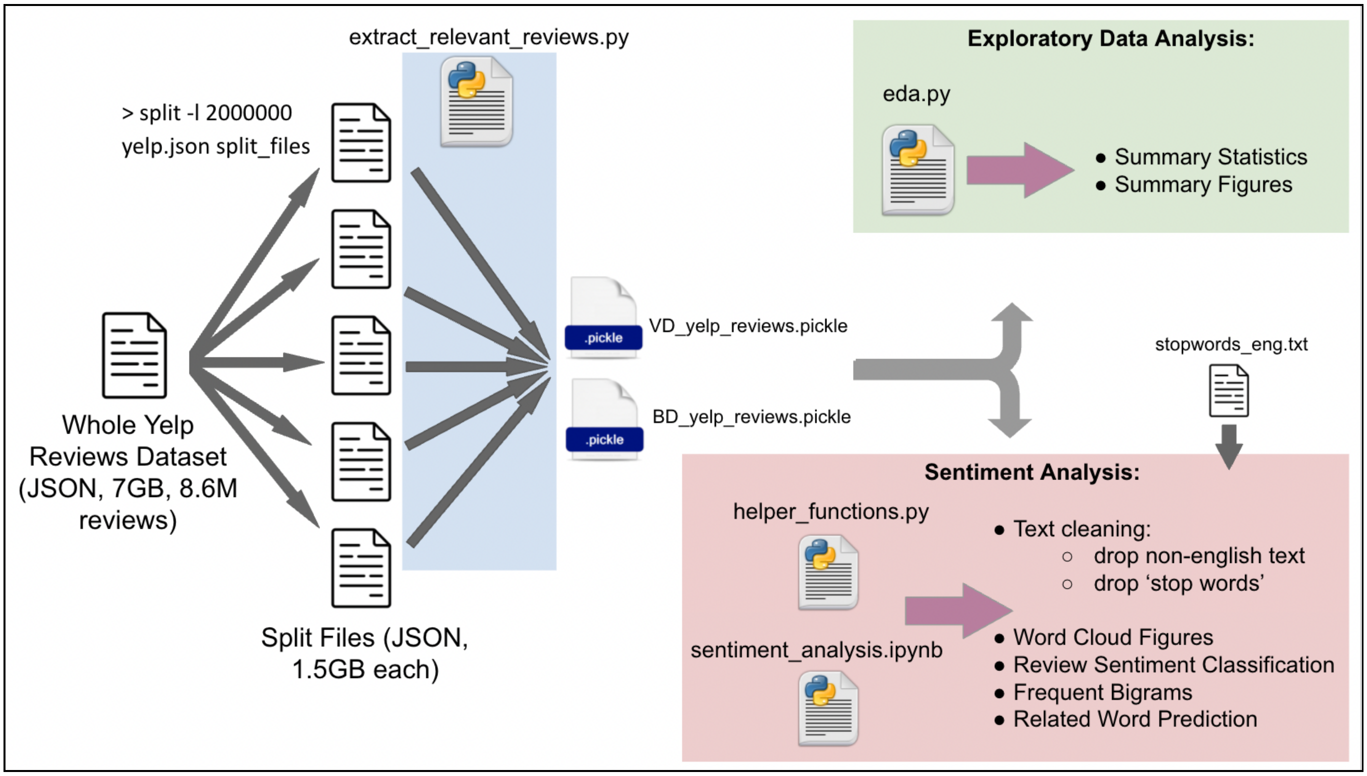

Figure 4: Analysis Workflow/ Systems Diagram

The whole Yelp dataset and the slip .json files were not uploaded to canvas as they are large, and not needed for someone else to run the eda.py and sentiment_analysis.py scripts.

Exploratory Data Analysis Results:

```
Number of reviews for Voodoo Donuts: 11280
Number of reviews for Blue Star Donuts: 6081
Number of unique restaurants for Voodoo Donuts: 2
Number of unqiue restaurants for Blue Star Donuts: 8

Columns in the pandas df:
['review_id', 'user_id', 'business_id', 'stars', 'useful', 'funny', 'cool', 'text', 'language']

Example: First two rows of Voodoo Donuts reviews:
                                  review_id                 user_id  \
date
2011-08-24 21:59:48  X4-EMXS7-XJZ9tuVJhn0-g  lzQLHoJdaMNwiNt76wrc4g
2015-10-07 21:39:23  19msT9vUbN1xI17FTsFBOw  s-xMG4eoG1YO0N795PaAiQ

                                business_id  stars  useful  funny  cool  \
date
2011-08-24 21:59:48  4CxF8c3MB7VAdY8zFb2cZQ      2       5      2     3
2015-10-07 21:39:23  4CxF8c3MB7VAdY8zFb2cZQ      4       0      0     0

                                                                  text  \
date
2011-08-24 21:59:48  I am alway leery of places the trade on schtic...
2015-10-07 21:39:23  It's a must when visitors come to town, not th...

                    language
date
2011-08-24 21:59:48       en
2015-10-07 21:39:23       en
```

Figure 5: Summary statistics output of eda.py

While a little messy I thought it would be nice to include a sample of the data.

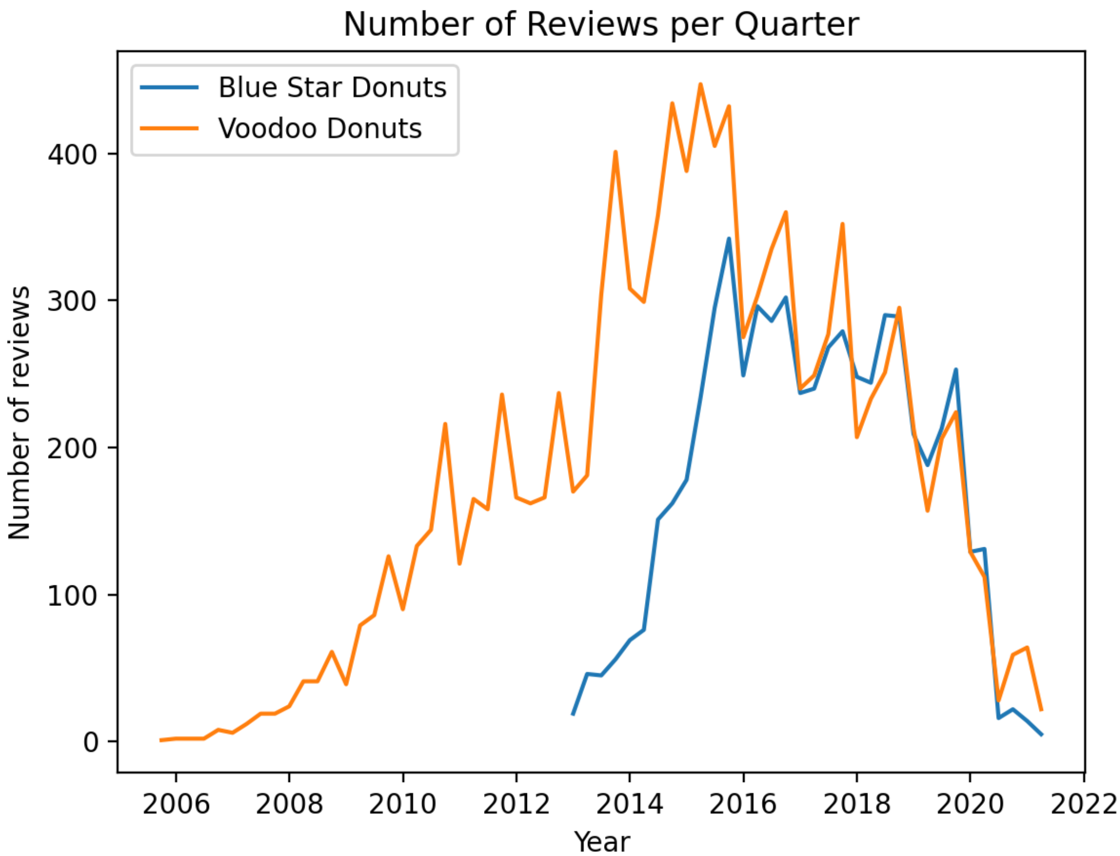

Figure 6: Number of reviews decline after 2015-2016 peak popularity. Also you can clearly see the drop in reviews due to covid around 2020.

Figure 7: The distribution for Voodoo Donuts is more uniform, while Blue Star Donuts has a large percentage of its reviews being 5s. If you look closely you can see that Voodoo Donuts actually has more 4 star reviews than 5 star reviews, not good.

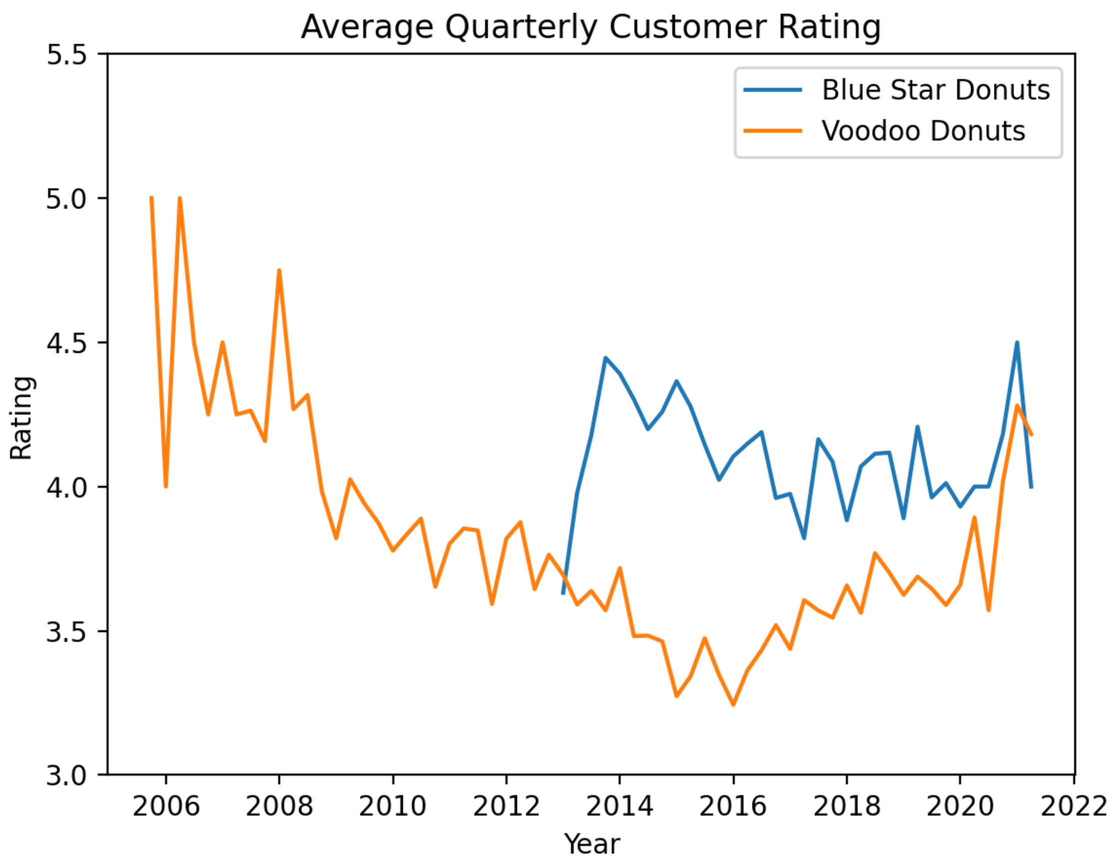

Figure 8: Voodoo Donuts have had more inconsistent review ratings, compared to Blue Star Donuts. Around 2016 the reviews for Voodoo began trending upwards again, while Blue Star has consistently outperformed Voodoo on a quarterly basis.

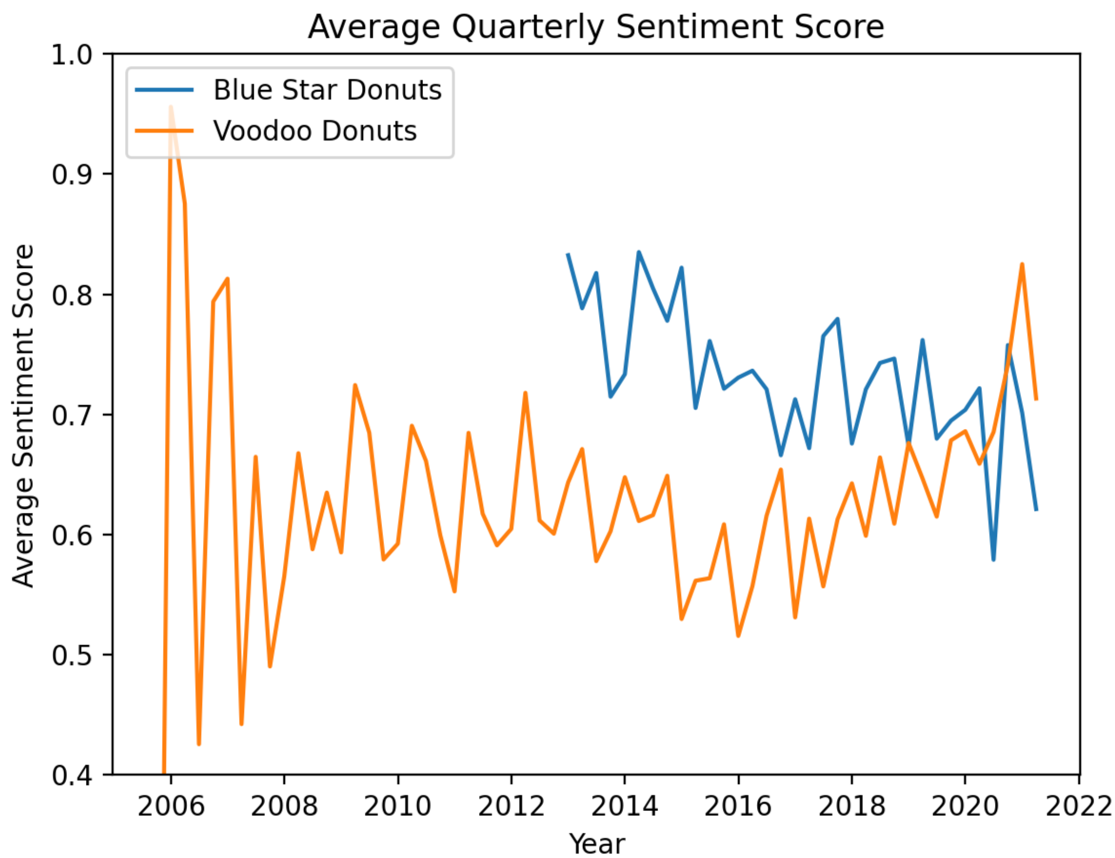

Figure 9: Similar to figure 7. The sentiment scores were generated using the NLTK (natural language toolkit) package. NLTK uses a large lookup table of positive/negative word assignments for the entire English language. It also uses intra-sentence punctuation and grammar to help determine sentiment. Link package: [NLTK]

Sentiment Analysis Results:

Blue Star Donuts:

Voodoo Donuts:

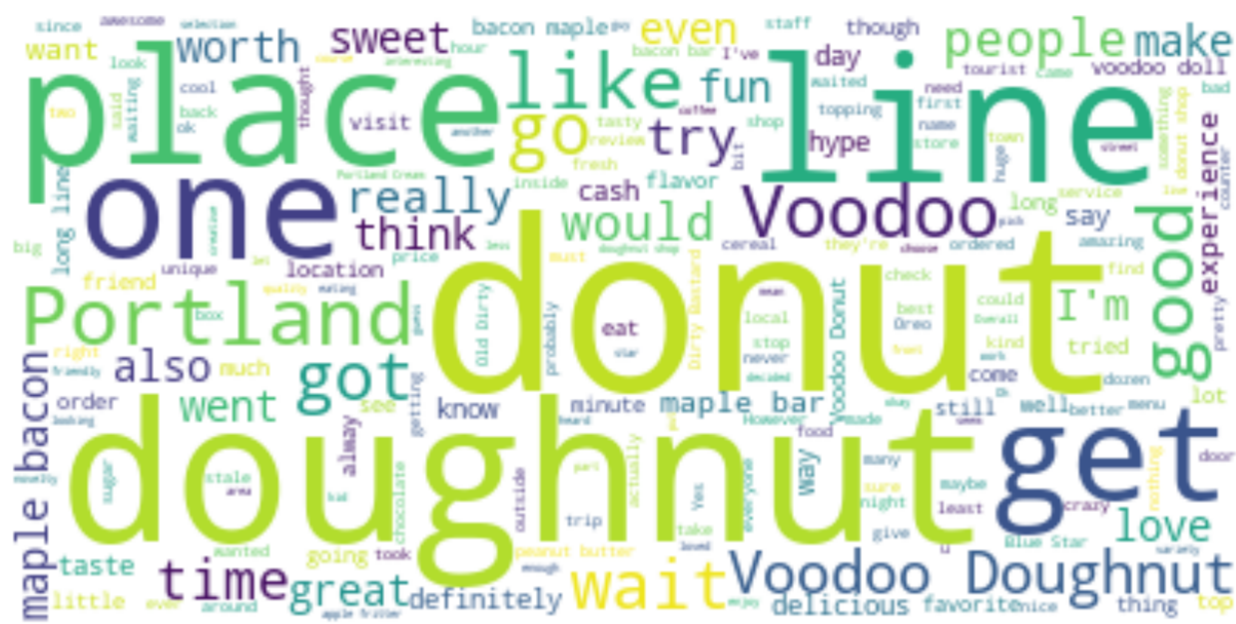

Figure 10: Word Cloud of most common words. Stop words were removed before these were generated.

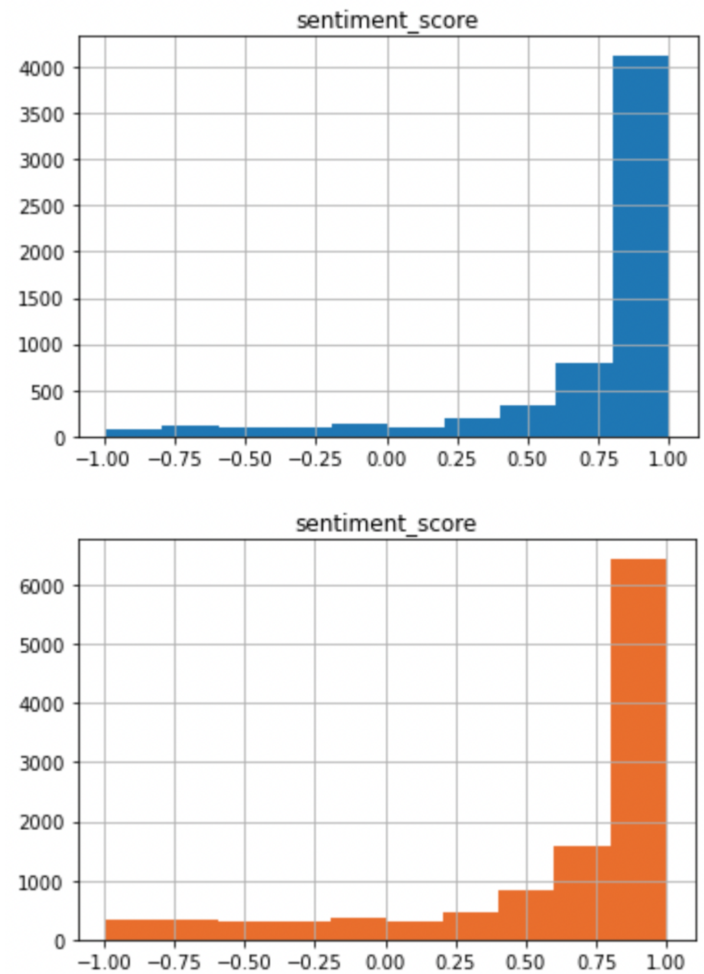

Figure 11: Distributions of sentiment scores. -1 is the most negative, 1 is the most positive, 0 is neutral. Obviously both are skewed heavily towards exprementy positive sentiment reviews. A cutoff of 0.75 was used to identify extremely positive reviews.

```
Number of Positive Voodoo Donuts Reviews: 4749
Number of Negative Voodoo Donuts Reviews: 1083
Number of Positive Blue Star Donuts Reviews: 3682
Number of Negative Blue Star Donuts Reviews: 341
```

Figure 12: Counts of reviews for each subset. Positive reviews were any review that gave a 4 or 5 star rating, and had a sentiment score between 0.75 and 1. Negative reviews were any review that gave a 1, 2, or 3, star rating, and had a sentiment score between -1 and 0.

Voodoo Donuts (Positive):

	bigram	freq
80112	maple bacon	834
9297	bacon maple	649
80115	maple bar	646
135430	voodoo doughnut	625
135422	voodoo doll	611
135431	voodoo doughnuts	510
94701	peanut butter	508
9082	bacon bar	343
135427	voodoo donuts	315
32723	dirty bastard	304

Voodoo Donuts (Negative):

	bigram	freq
3371	blue star	133
33502	tourist trap	107
20344	maple bacon	81
35514	wait line	79
35253	voodoo doughnuts	78
6714	customer service	74
19452	long line	71
35252	voodoo doughnut	71
22245	nothing special	70
20345	maple bar	64

Blue Star Donuts (Positive):

	bigram	freq
9262	blue star	2359
9307	blueberry bourbon	794
9775	bourbon basil	722
82726	star donuts	639
65432	passion fruit	479
56294	maple bacon	451
62205	old fashioned	355
95827	voodoo donuts	350
7941	best donuts	346
18792	creme brulee	257

Blue Star Donuts (Negative):

	bigram	freq
1266	blue star	122
7400	maple bacon	34
3030	donut shop	31
10961	star donuts	29
2417	customer service	26
578	apple fritter	23
8066	nothing special	23
3244	donuts good	22
1271	blueberry bourbon	21
1342	bourbon basil	19

Figure 13: The most common bigrams for each of the subset of reviews.

```
Voodoo Donuts (positve):
Target Word(s): ['flavor']
Related Words: ['flavor', 'toppings', 'grape', 'combinations', 'texture', 'topping', 'taste', 'flavors', 'combination', 'candy']

Voodoo Donuts (negative):
Target Word(s): ['line', 'wait']
Related Words: ['outside', 'stand', 'see', 'protest', 'want', 'gigggle', 'lines', 'waiting', 'shop', 'minutes']

Blue Star Donuts (positive):
Target Word(s): ['flavor']
Related Words: ['unique', 'combinations', 'subtle', 'texture', 'interesting', 'strong', 'flavor', 'overly', 'taste', 'rich']

Blue Star Donuts (negative):
Target Word(s): ['line', 'wait']
Related Words: ['rival', 'zoroastrianism', 'joked', 'commoners', 'potted', 'flaw', 'usually', 'ol', 'counted', 'connected']
```

Figure 14: Word prediction results for each of the subset of reviews.