COMP-3006 Project Proposal
Brian Miller (DU ID: 873601817)
02/24/2022

Project Proposal Reminder:
I will be performing a sentiment analysis on the two main donut competitors in Portland, OR: Blue Star Donuts vs Voodoo Donuts. I lived in Portland for the last two years, and most people I have met have a strong opinion about which is better. What distinguishes Blue Star from Voodoo? Types of positive reviews, types of negative reviews, donut attributes, something else?

Data: Yelp open source academic dataset

Update: Initial sentiment analysis steps

This week was a light week, I had other responsibilities with other classes.

Now that I have a handle on the data, I began digging into the sentiment analysis.
Here are the steps taken so far:

1. Subset data to only English reviews
2. Remove 'stop words' from the dataset
3. Generated word clouds of the most frequent words from each business

Link to new script (written just for this progress report):
https://github.com/BrianMillerS/DU_COMP3006/blob/main/final_project/sentiment_analysis.py

Step 1: Ensure that all of the reviews are in english. I'll use python's langdetect module that has a '99% over precision for 53 languages'.

```
Language counts for Blue Star:
en    6070
fr       4
ja       4
es       2
fi       1
Name: language, dtype: int64
Language counts for Voodoo:
en    11255
fr        9
ja        5
ro        3
af        2
```

Thankfully the vast majority of the reviews were in english, but not all. For reference, here are four reviews that are in Japanese.

```
                                                                text  \
date
2014-07-18 04:57:32  シナモンシュガー、スパイス、ラズベリーケーキを持ち帰り。\nできたてだから美味しいのかと思っ..
.
2016-09-03 12:11:21  東京の代官山などに店舗があるポートランドのドーナツ屋さん。\n\nポートランドは美味しいドー..
.
2019-11-04 11:43:41  日本から撤退してしまった、ブルースタードーナツ。\n大袈裟に言っているわけではなく、私が世界..
.
2017-10-03 21:18:17  ここのアップルフリッターが好きで、ホーソーン通りで見つけて入ったけど、残念、なかったーその代.
..

                     language
date
2014-07-18 04:57:32       ja
2016-09-03 12:11:21       ja
2019-11-04 11:43:41       ja
2017-10-03 21:18:17       ja
```

Step 2: Remove all 'stop words' from the dataset. Stop words are words that will not be important for us to look at. These are common words in the english language that should be dropped.

From looking online one common source of stop words is from NLTK which looks to be an academic open source toolset for natural language processing.

Their dataset has 179 english stopwords. These words were removed from our datasets.

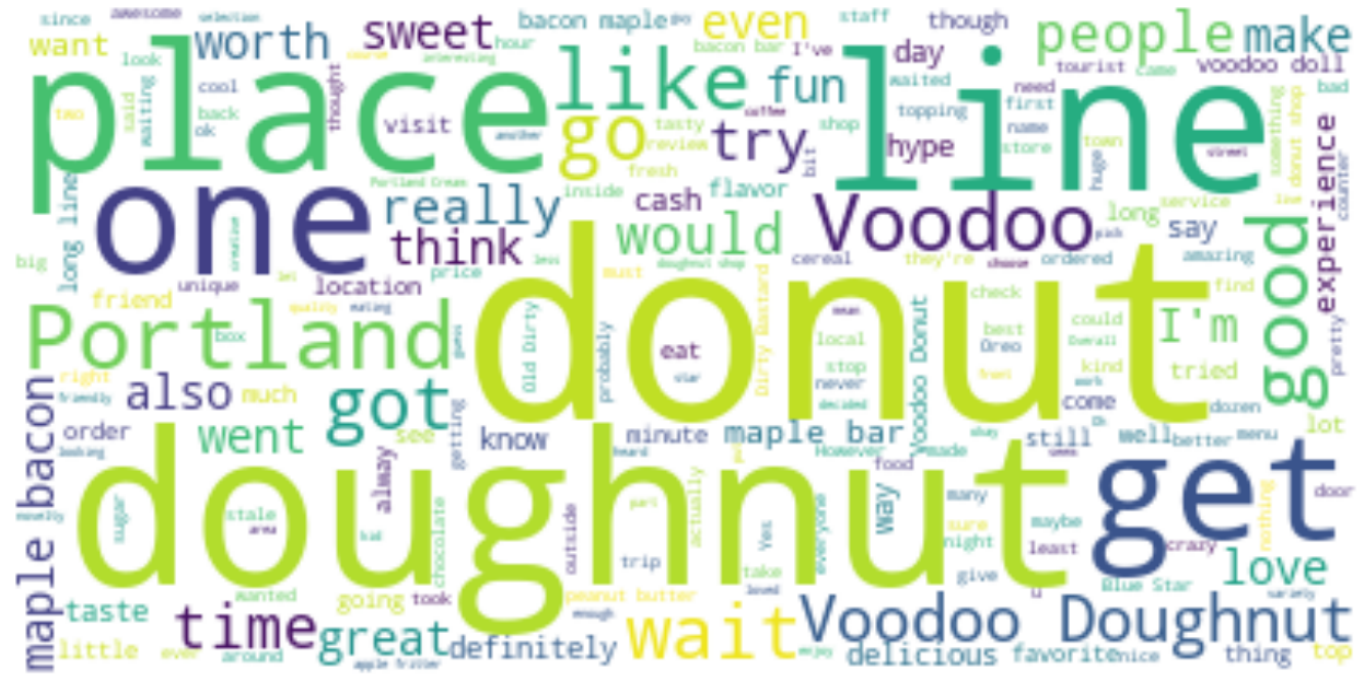
```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ou
"you'd", 'your', 'yours', 'yourself', 'yourselves',
'her', 'hers', 'herself', 'it', "it's", 'its', 'itse
'themselves', 'what', 'which', 'who', 'whom', 'this',
'is', 'are', 'was', 'were', 'be', 'been', 'being', '
'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if',
'of', 'at', 'by', 'for', 'with', 'about', 'against',
```

Step 3: Word Clouds

Next I was curious what words were the most frequent for each business. Rather than making a standard list or data frame with results I decided to make word clouds with are more fun to look at.

Word Cloud: Blue Star Donuts

Word Cloud: Voodoo Donuts



Right off the bat we can see some interesting words pop out at us: line, good, like, flavor. These are adjectives that will be useful for us to look at as this analysis gets more complex.

For my next project update I will be implementing a ‘bag of words’ approach and will be generating lists of common bi and tri word phrases for each company. I would also like to see which adjectives are around the word donut to see if we can gather some insight on whether or not one company vs the other has some donut characteristics that would be good to know.