Venue User Insight via WordCloud Generation

Thien Nguyen

July 10th, 2020

1. Introduction

1.1 Background

Review sites, such as Yelp, Google Maps, and Foursquare, have been indispensable to the food and service industry as well as consumers. Businesses, like restaurants, did not have to rely exclusively on paid advertisement or secretive word-of-mouth and could instead rely on users to give assessments of the quality of their service and product. Conversely, users could rely on each other to provide a network of evaluations about a business, allowing them to make a sound judgment about whether or not to pay for the services or goods offered. These sites often rely on two basic systems for appraising a venue—a numerical rating system and detailed, user-written reviews.

1.2 Problem

Despite being around for a while, these sites do not offer users any other way to quickly assess a business. A star rating between one and five only provides useful information for scores at the extreme; it is hard to discern whether a restaurant rated at 4.2 stars should be preferred over a restaurant with 4.4 stars, given that both have a similar number of ratings. On the other hand, reviews can be useful, but require customers to scrolls through endless pages, inevitably drawing their eye to the (hopefully) modicum of negative reviews among a plethora of positive ones, which may cause them to reconsider altogether! This wide discrepancy between informational extremes has somehow become the norm.

1.3 Solution

Providing a medium choice that offers just enough information without bogging users down is the intention of this investigation. Such a solution would still rely exclusively on unedited user feedback, but it would summarize this feedback in a non-bias manner that still provides more information than a number score. The proposed solution is to generate word clouds based on user reviews to give users an intuitive idea of what other customers are discussing in relation to the business at hand. This would allow users to get a general synopsis of comments as well as providing talking points for them to springboard from, if they decide to dig deeper through the tips for more recommendations

2. Data Acquisition

2.1 Data Source

This project relied on venue tips data from Foursquare and selected venues based on an *explore* call for local restaurants in the Houston area. This call was used to set up a list of restaurants to test the function on

3. Methodology

3.1 API Calls

Foursquare API is very straightforward to set up, allowing for both user and userless access. A personal (free) account could request 500 premium calls a day, which would allow for information including user data, venue pictures, and venue tips. The free account was created to acquire this data.

Initial planning involved creating a dataset of a hundred restaurants in downtown Houston, complete with name, Foursquare ID, and restaurant food genre. A script was then created to compile the entirety of each restaurants tips as a continuous string and load it into the dataset under a corresponding column *tips*.

3.2 Natural Language Processing

Each restaurants' tips would be cleaned and processed in the following manner:

1. Punctuation removal to retain alphanumeric characters
2. Case standardization (lower-case)
3. Removal of stop words
4. Lemmatization

This processed string would be used to generate the word cloud for each venue, thus providing a summary of all tips for the venue.

4. Results

4.1 Complications

Initial results were disappointing as the free develop account limited even premium calls to only two tips per venue. Since word clouds relied on an abundance of text, two tips is far too scarce for any meaningful word cloud generation.

4.2 Workaround

To demonstrate a workable model, tips from two top Houston area restaurants were manually scraped from their Foursquare page and utilized to create the following word clouds:

Word Cloud for Brothers Tacos

Word Cloud for Tout Suite

5. Discussion

Word Clouds tend to provide a bit more information about a location than a numeric score, but not by much. An initial glance can give users a plethora of information, including general perception, food recommendations, and location details. Yet, all this information is shallow and lacks detail, especially when considering context.

Adjectives (E.g. *great)* provide remarkable data as to what customers think about a venue, but nouns are much more obscure without context (E.g. *place*). Certain words, like *atmosphere*, may carry inherent positive connotations concerning restaurants, but they are limited.

Lastly, it is regrettable the complete script could not be performed without an enterprise account with Foursquare. However, the script provided will allow any developer with such account to create such a word cloud.

6. Conclusion

Word Clouds are not meant to replace numeric scores or written reviews. Rather, they offer users a fun, visual peek about what other consumers are saying about a venue. While a number score loses relevance as a venue garners hundreds of reviews, word clouds can remain topical. For many, this may simply be enough rather than having to become bogged down in reading extensive reviews. For others that want to read reviews, they provide a good jumping point in seeing what people have to say about certain menu items or restaurant conditions.

7. Future Considerations

- Use bigrams to provide more context on nouns
- Allow specific word clouds for individual categories like service, parking, etc.
- Implement sentiment analysis to further filter out ambiguous words
- Circumvent free account restrictions with other webscraping approaches