**Analysis on which chain serves the best chicken sandwich based on statistical methods and computational methods**

**Introduction/Motivation**

The restaurants we are going to analyze are those who feature or serve chicken sandwiches. There are a number of restaurants who serve chicken sandwiches. However, the popularity of their chicken sandwiches varies a lot. Since in this project we have already got the data of customers’ reviews. The best idea to analyze if a chicken sandwich is popular or not as well as why it is popular is to analyze from the aspect of customer feedbacks.

Our first specific goal is to find out the chain of restaurants who serves the best chicken sandwiches as well as find out why the chicken sandwiches from this chain are the best. Our second specific goal is to find out useful information from customers’ reviews for each chain of restaurant who serves chicken sandwiches, to see what is the advantage and weakness of their recipe of chicken sandwiches, management mode and business mode. This leads to the suggestions and recommendations for each business.

To find out the chain of restaurants who serves the best chicken sandwiches, we need quantitate ratings (stars) of each chain. Statistical methods can be used for analysis and inference.

To find out the advantage and weakness of each chain’s management mode, we need qualitive ratings (text) of each chain. NLTK in Python (TD-IDF with the help of sentiment analysis and concordance analysis) can facilitate extract sentiment words as well as key words from texts.

**Background Information**

The data is downloaded from Canvas STAT628 Module 3. It is a subset of millions of reviews provided by Yelp, which is an Internet platform for users to write reviews of businesses. The data consists of five .json files (a.k.a JavaScript Object Notation). Each file consists of high-dimension data. The data focuses on reviews of restaurants which are located in Madison (U.S.), Cleveland (U.S.), Pittsburgh (U.S.) and Urbana-Champaign (U.S.).

We first come up with a list of chains who features or serves chicken sandwiches. (Popeyes, KFC, Wendy’s, Chick\_Fil\_A, Carl’s\_Jr, McDonald’s, Burger\_King)

We then open business.json with Python to find out the corresponding business\_id for these chains. In review.json, we only keep the ratings (stars) and reviews (text) for these chains.

To search for the chain who serves the best chicken sandwiches, (1)we first need to test if the average rating of chicken sandwiches for each chain is the same as the average overall rating for each chain. In review.json, we keep the data whose review (text) contains “chicken sandwich” to get ratings of chicken sandwiches for each chain. (2) According to user\_id, we then merge user.json and the data we filter from review.json to get users’ ratings (star) for chicken sandwiches and average ratings (average\_stars). This can help us get the standardized ratings for chicken sandwiches for each chain.

To find out useful information from customers’ reviews for each chain of restaurant who serves chicken sandwiches, to see what is the advantage and weakness of their recipe of chicken sandwiches, management mode and business mode, we focus on review.json which is already filtered by business\_id ( it should be among the list of chains) and text which contains “chicken sandwich”.

**Exploratory Data Analysis**

1. Hypothesis testing

We test whether chicken sandwich ratings (stars) and overall restaurant ratings (stars) are significantly different for each chain. We use t-test. The p-value is. At the significant level =0.05, we reject the null hypothesis that chicken sandwich ratings (stars) and overall restaurant ratings (stars) are the same. We can conclude that for each chain, chicken sandwich ratings (stars) is significantly different from overall restaurant ratings (stars). Thus, we need to filter chicken sandwich reviews from all reviews for each chain when we do analysis.

(user chicken sandwich rating vs user average rating)

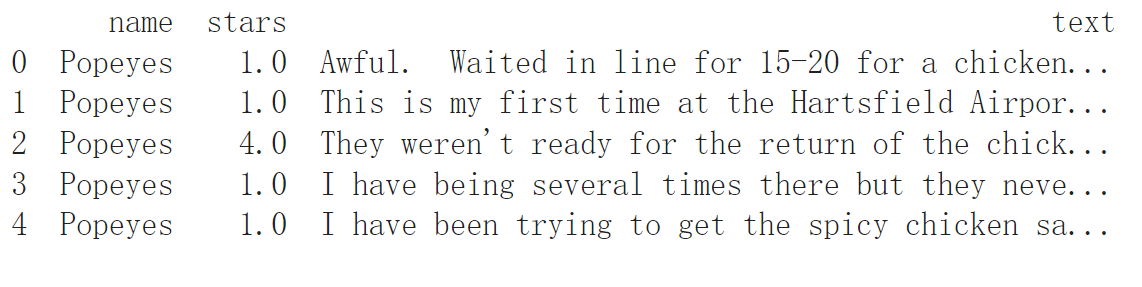
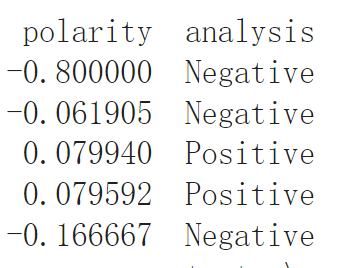
1. Sentiment analysis (word cloud)

We use module NLTK, pandas, matplotllib.pyplot, re in Python to do sentiment analysis on customers' reviews on chicken sandwiches. Sentiment analysis is to let computer simulate human reading text reviews and response positive or negative impression on the reviews. This generates the pos\_and\_neg\_reviews.csv, which shows the sentiment of each review based on the text message of each review. It divides chicken reviews into three categories (Positive, Negative, and Neutral) so as to facilitate our TD-IDF analysis.

The steps are as follows. First, we clean the dataframe so as to keep only chicken sandwich reviews as well as useful columns. (name, user\_id, text, stars) Second, we remove punctation from all chicken sandwich reviews (texts). We also remove stop words from them. Then we assign each word in chicken sandwich reviews with tags according to their part of speech. For adjectives, the tag is J. For verbs, the tag is V. For nouns, the tag is N. For adverb, the tag is R. After tagging the words in chicken sandwich reviews, we lemmatize the words so that they are in similar formats for analysis. For instance, we make words with different tense the same. Last but not least, according to the lemmatized words in each chicken sandwich review, and based on the language base in NLTK, we get the subjectivity of each review and calculate the polarity of each review. Each review is assigned with a polarity score which ranges from -1 to +1. The closer the polarity score is to -1, the more negative the review is. The closer the polarity score is to +1, the more positive the review is. We set the criteria as follows to label the reviews.

|  |  |
| --- | --- |
| Polarity score | The label we set |
| [-1, -0.025) | Negative |
| [-0.025, 0.025] | Neutral |
| (0.025, 1] | Positive |

Some results of sentiment analysis are as follows.



As a result, there are totally 1223 reviews which are tagged as “Positive”, 1011 reviews which are tagged as “Negative”, and 190 reviews which are tagged as “Neutral”.