Business Insights using Yelp Review

Ashwini Vishwanath Hosmani,

Meghana Vijaykumar Rai

Siddhant Gawsane

Feifan Wu

Xin Li

Contents

Introduction

Yelp is an American multinational corporation headquartered in San Francisco, California. It develops, hosts and markets Yelp.com and the Yelp mobile app, which publish crowd-sourced reviews about local businesses, as well as the online reservation service Yelp Reservations. The company also trains small businesses in how to respond to reviews, hosts social events for reviewers, and provides data about businesses, including health inspection scores. Yelp publishes customer reviews of local businesses, and it also provides services such as Yelp Reservations and the online food delivery service known as Eat24. Additionally, it provides data about businesses, helps train small businesses, and hosts social events.

Our project uses a series of datasets which includes information from cities in select countries, and we will be using this data to analyze various cultural and seasonal trends. First off, according to Yelp, the dataset encompasses the following areas and includes the following information:

• 4.1M reviews and 947K tips by 1M users for 144K businesses • 1.1M business attributes, e.g., hours, parking availability, ambience. • Aggregated check-ins over time for each of the 125K businesses • 200,000 pictures from the included businesses	Countries • U.K • Germany • Canada • U.S.

Yelp provides, in a Json file, multiple datasets which include general information on businesses, reviews received, information on users, check-ins logged, tips given, and user-uploaded images. For our project, we are working only on the businesses and reviews datasets.

Problem Statement

Using Yelp's businesses and reviews datasets, we attempted to answer the following questions:

- What are the restaurant's' biggest problems which brings down the ratings the most, based on customer reviews?

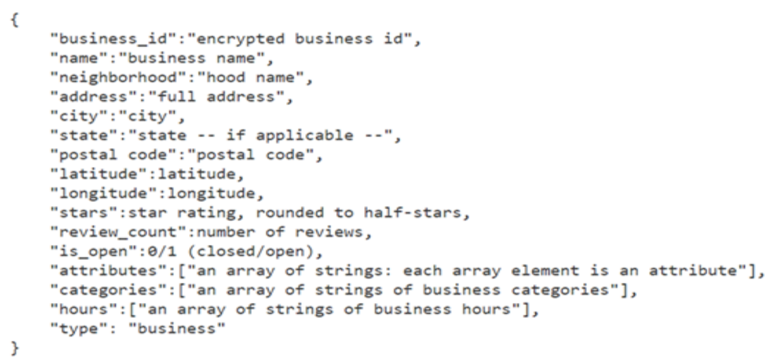
About Dataset

Yelp provides the datasets as Json files. Data source for our project comes from: https://www.yelp.com/dataset_challenge.

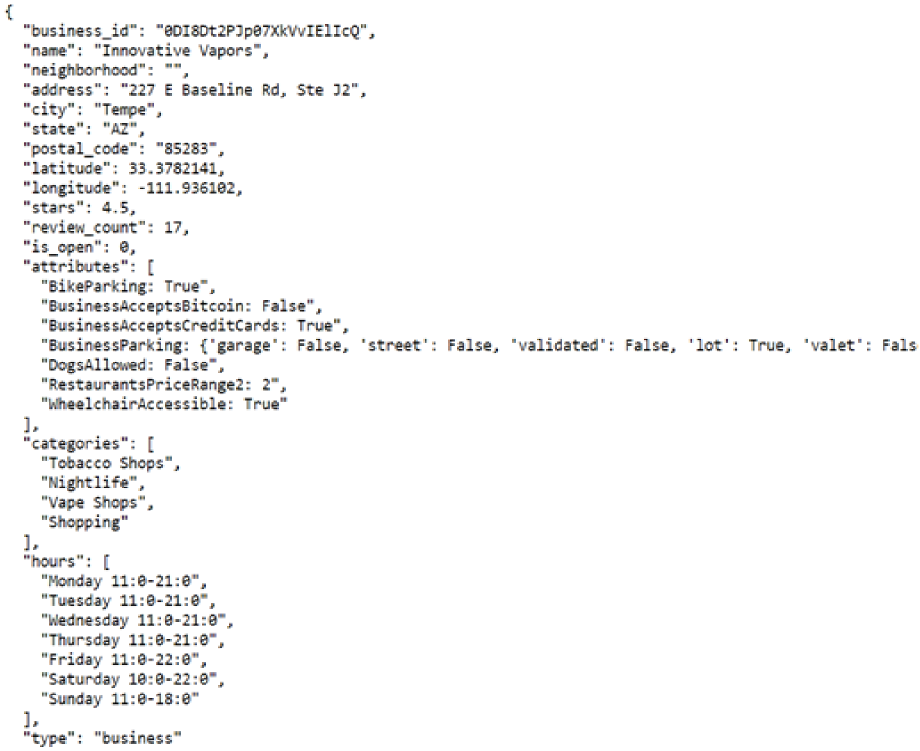
The following are the dataset we used for the project including the attribute information contained in each dataset.

Business Dataset

The business dataset provides information about different businesses from different countries. For each business establishment, there is a unique Business ID, name complete address with latitude and longitude of the city it is situated in, type of business, and review counts.



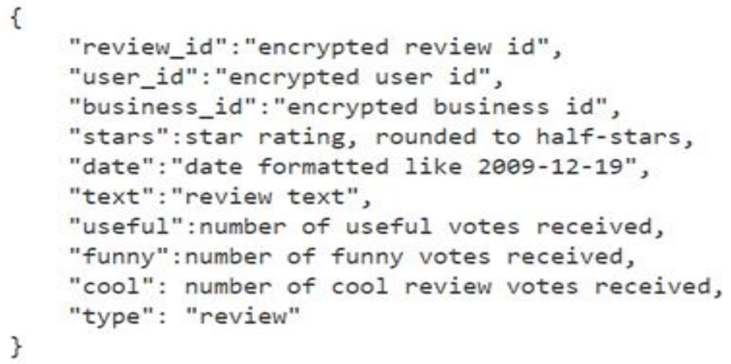
```
{
    "business_id":"encrypted business id",
    "name":"business name",
    "neighborhood":"hood name",
    "address":"full address",
    "city":"city",
    "state":"state -- if applicable --",
    "postal code":"postal code",
    "latitude":latitude,
    "longitude":longitude,
    "stars":star rating, rounded to half-stars,
    "review_count":number of reviews,
    "is_open":0/1 (closed/open),
    "attributes":["an array of strings: each array element is an attribute"],
    "categories":["an array of strings of business categories"],
    "hours":["an array of strings of business hours"],
    "type": "business"
}
```



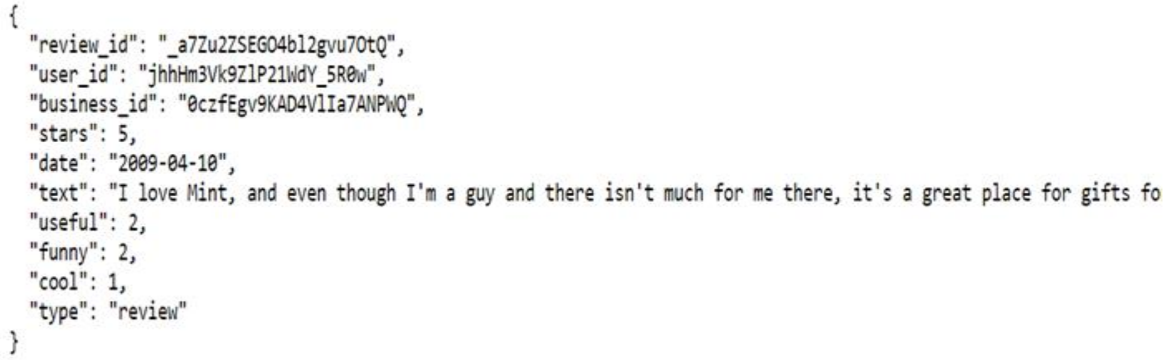
```
{
  "business_id": "0DI8Dt2PJp07XkVvIElIcQ",
  "name": "Innovative Vapors",
  "neighborhood": "",
  "address": "227 E Baseline Rd, Ste J2",
  "city": "Tempe",
  "state": "AZ",
  "postal_code": "85283",
  "latitude": 33.3782141,
  "longitude": -111.936102,
  "stars": 4.5,
  "review_count": 17,
  "is_open": 0,
  "attributes": [
    "BikeParking: True",
    "BusinessAcceptsBitcoin: False",
    "BusinessAcceptsCreditCards: True",
    "BusinessParking: {'garage': False, 'street': False, 'validated': False, 'lot': True, 'valet': Fals
    "DogsAllowed: False",
    "RestaurantsPriceRange2: 2",
    "WheelchairAccessible: True"
  ],
  "categories": [
    "Tobacco Shops",
    "Nightlife",
    "Vape Shops",
    "Shopping"
  ],
  "hours": [
    "Monday 11:0-21:0",
    "Tuesday 11:0-21:0",
    "Wednesday 11:0-21:0",
    "Thursday 11:0-21:0",
    "Friday 11:0-22:0",
    "Saturday 10:0-22:0",
    "Sunday 11:0-18:0"
  ],
  "type": "business"
```

Review Dataset

The review dataset provides information about the different reviews posted by different users for different business establishments. Each record has a unique review ID, user ID of the Yelp user, and business ID of the business reviewed by the Yelp user and other details such as date of the review and type of reviews.



```
{
    "review_id":"encrypted review id",
    "user_id":"encrypted user id",
    "business_id":"encrypted business id",
    "stars":star rating, rounded to half-stars,
    "date":"date formatted like 2009-12-19",
    "text":"review text",
    "useful":number of useful votes received,
    "funny":number of funny votes received,
    "cool": number of cool review votes received,
    "type": "review"
}
```



```
{
  "review_id": "_a7Zu2ZSEGO4bl2gvu7OtQ",
  "user_id": "jhhHm3Vk9Z1P21WdY_5R0w",
  "business_id": "0czfEgv9KAD4VlIa7ANPWQ",
  "stars": 5,
  "date": "2009-04-10",
  "text": "I love Mint, and even though I'm a guy and there isn't much for me there, it's a great place for gifts fo
  "useful": 2,
  "funny": 2,
  "cool": 1,
  "type": "review"
}
```

Technology

Text mining/K means clustering

To collect and analyze the keywords in the businesses and review dataset, we apply text mining as one of our technologies. Text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern. By using text mining and referring the result it shows us, we are able to have a straight data and visual show on how general reviews look like and what customers think about the restaurants they have been visited. Then it is helpful for us digging into the problem that we want to discuss in our project.

Also, we utilize K-means clustering as the other technology for our project. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. In our project, we are able to tell the most frequent keywords and problems customer mentioned in their reviews and collect the most influential problems restaurants always have when using K-means clustering. The results of K-means clustering show directly the average of each problem that matters the impression of restaurants customer have been visited.

Tools

For our project, we use 3 main Python libraries:

1. NLTK
2. Sklearn
3. TextBlob

NLTK (Natural Language Toolkit) is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for English written in the Python programming language. This tool allows us import word_tokenize from nltk, import PorterStemmer from nltk. stem and import stopwords from nltk.corpus.

Sklearn, a Machine Learning library in Python is a simple and efficient tool for data mining and data analysis. Sklearn library provides us with function such as Kmeans and TfidfVectorizer thru which we can perform analysis on the data.

Textblob, which is our last tool, provides us with text translation as the reviews come from different countries.

Analysis

Our code takes business ID with the rating of less than 3 and review count of more than 500 and attempting to pin point the main reason for lesser reviews by the users checking in to the restaurant.

First, in the **function words_count**, we mainly concentrate on removing the stop words from the user reviews and finding the word count of each word. Then in the

In the **function process_text**, we tokenize and stem the user reviews to remove words of lower relevance.

Finally, in the function **cluster_texts,** we use tfidf to vectorize the words in the reviews and then use K-means clustering. We now run K-means clustering on the stem words and cluster the words with similar tfidf score.

Once we get the clusters, we take only the nouns from the cluster, run sentimental analysis and also count the number of times these nouns recur from in each cluster which will rightly pin point the main problems the users have pointed out in their reviews.

Business Cases

To demonstrate this, we have chosen reviews of three business cases:

1.Mon Ami Gabi

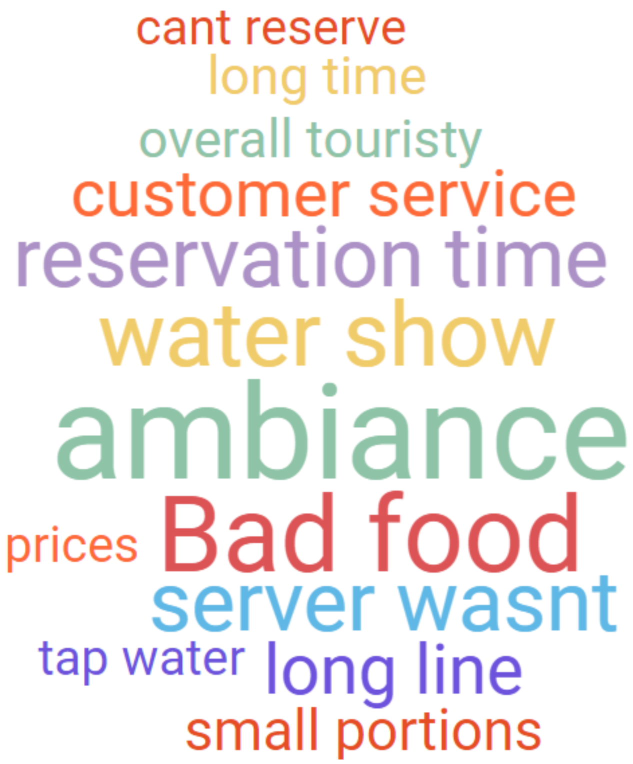
2. Monte Carlo Hotels and Casino

3.LoLo's Chicken and Waffle

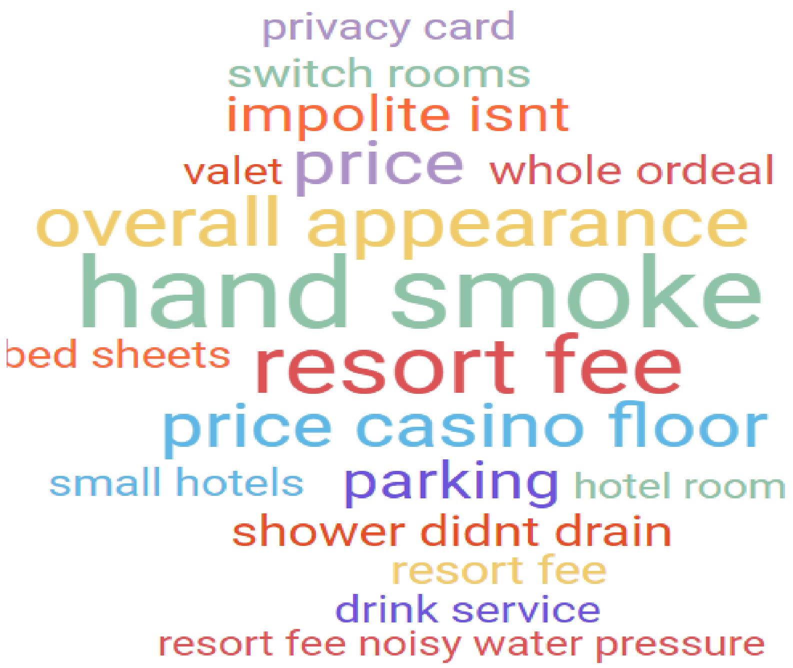
To compare these three cases and make it clear, we make a table to show total number of reviews, total number of words in reviews, hours to read and total price at a rate of $12 per hour.

Name of business case	Total Number of Reviews	Total number of Words in Reviews	Hours to read	Total price at a rate of $12 per hour
1.MON AMI GABI	6414	781207	65	$780
2.MONTE CARLO	2080	363732	30	$360
3.LO-LO'S CHICKEN & WAFFLES	1276	156102	13	$156

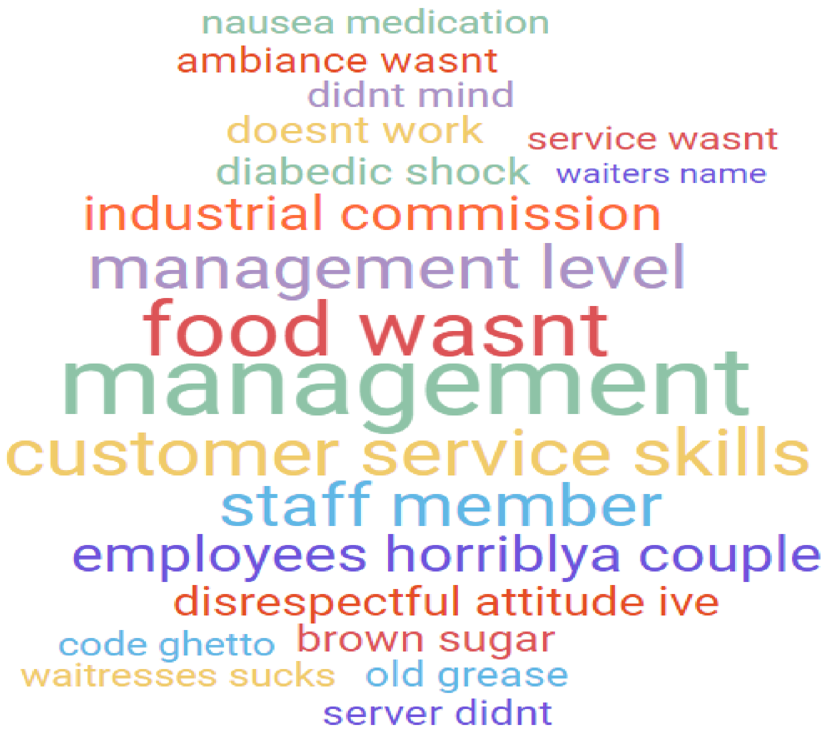
After analysis, we found commonly occurring negative comments of each case as below:



Commonly Occurring Negative Comments of MON AMI GABI



Commonly Occurring Negative Comments of MONTE CARLO



Commonly Occurring Negative Comments of LO-LO'S CHICKEN & WAFFLES

Conclusion

In the current technological advanced world, its imperative that businesses make sure they provide the best services to their customers. With the help of Kmeans clustering and Tfidf Vectorizer we could pin point the recurring problems the users faced and mentioned in their reviews. We make it easy for the businesses to swift thru millions of reviews on Yelp and quickly find the root cause for any bad reviews and fix it immediately. Because in the case of restaurant business the only kind of publicity that works is good publicity.

Future Scope

Online presence for any business establishment is a top priority. Recent studies have shown that 90% of the customers read online reviews and 88% use these reviews to take decisions.

Online presence does not involve just the content businesses puts in but the content generated by the users. It is essential to understand the impact of online reviews and also online review sites as many users use this content to take decisions.

Using our project, we can copy the results to other sites which includes:

1. Search engines such as Google+,Yahoo,
2. Sites such as Tripadvisor for travelers, UrbanSpoon for restaurants
3. Social media such as Facebook, Twitter, Instagrma
4. Third party blogs and articles

References

https://thrivehive.com/power-online-customer-reviews/

https://en.wikipedia.org/wiki/Text_mining

https://en.wikipedia.org/wiki/K-means_clustering

https://en.wikipedia.org/wiki/Yelp