CSP-554 Big Data Technologies
PROJECT REPORT
YELP DATA ANALYSIS

By:
Bhavi Shah - A20471597 - bshah44@hawk.iit.edu
Vinjal Doshi - A20473398 - vdoshi4@hawk.iit.edu
Yashwanth Praveen - A20473431 - ypaspuleti@hawk.iit.edu

Abstract

This project will help us analyze yelp data sets using technologies such as Apache Hive, Apache Pig and Apache Sqoop. We will also be presenting an analysis which could help restaurants get better ratings, or the related suggestion to make their business work. Analysis of yelp dataset will help us in getting the insights about how the business works in the food industry, which factor of a restaurant matters more to a customer and which aspects of the restaurant can be ignored, to make the investors or the owners to focus more on the important aspects. This also helps in finding which can be listed as the most trusted yelp reviewers (Like the users who are pretty active etc). Also, we will identify in which city which restaurant business is making more profit.

Introduction

This dataset is a subset of Yelp's businesses, reviews, and user information. It was initially assembled for the Yelp Dataset Challenge which is an opportunity for students to direct research or examine Yelp's information and offer their revelations. In the dataset you'll track down data about organizations across 11 metropolitan regions in four nations.

In this report we did lots of EDA, Text Mining and Analysis which is a great platform for Data Analysis.

For a business we did the following analysis:

- Top Ten most common Words reviews of the business
- Sentiment Analysis - Postive and Not So Postive Words of reviews
- Relationship among words
- Relationship of words with an **important** word in the review such as steak, crab, food

The business we are analysing are:

- **Mon Ami Gabi** , a **Las Vegas** Restaurant , the most popular and highly rated restaurants
- **Chipotle Business in Yonge Street Toronto** you guessed it right, I like **Chipotle!**

How Sentiment Analysis can help your business ?

For a business, the Sentiment Analysis is very important. If the business owners can just see the Top Ten negative reviews, they can easily find out which aspect of the business they need to improve.

How Topic Modelling can help understand your business and city ?
Topic modelling helps to pick specific topics from the huge volume of text. Topic Modelling on the Three popular restaurants and also on Phoenix City helps us to understand that complaints regarding restaurants and business are around Service.

About the data set

The business data set includes following columns:

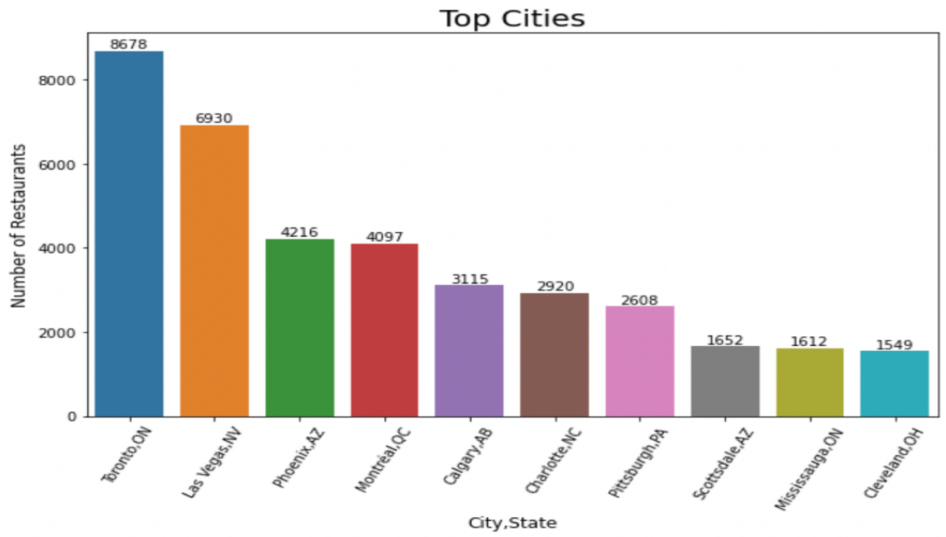
1. Address
2. Business id
3. Categories
4. City
5. is_open
6. Latitude
7. Longitude
8. Name
9. Neighborhood
10. Postal Code
11. Reviews
12. Stars
13. State

Here is a glimpse of the data set:

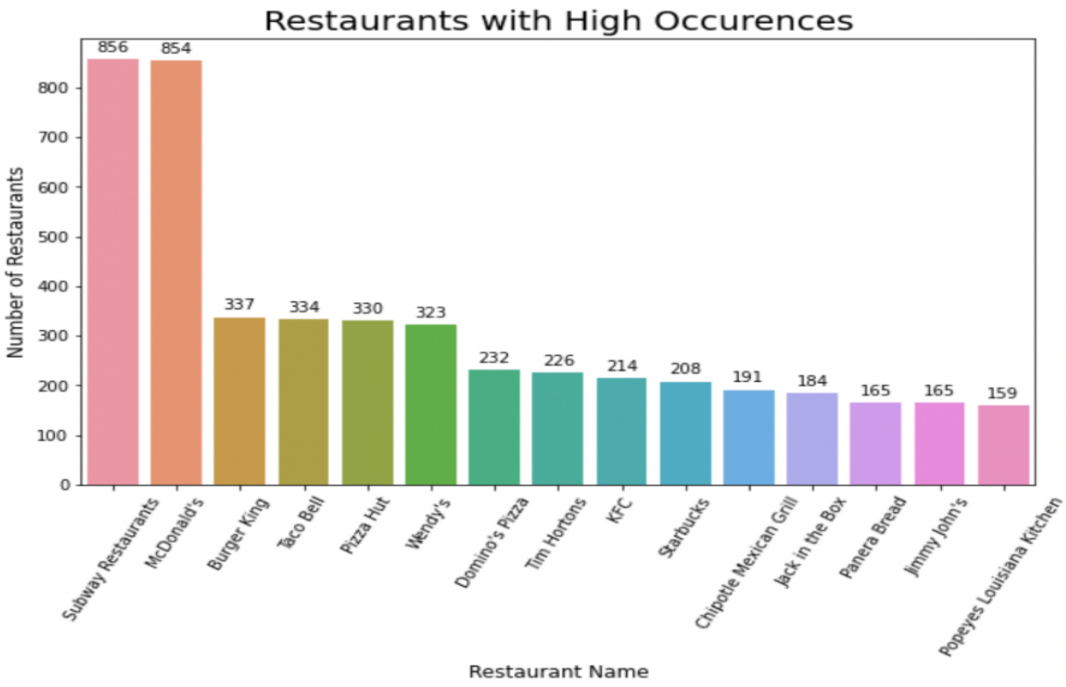
	name	neighborhood	address	city	state	postal_code	latitude	longitude	stars	review_count	is_open
1	FYWN1wneV18bWNgQjJ2GNg	"Dental by Design"		"4855 E Warner Rd, Ste B9"	Ahwatukee	AZ	85044	33.3306902	-11
2	He-G7vWjzVUysIKrfNbPUQ	"Stephen Szabo Salon"		"3101 Washington Rd"	McMurray	PA	15317	40.2916853	-80
3	KQPW8lFf1y5BT2MxiSZ3QA	"Western Motor Vehicle"		"6025 N 27th Ave, Ste 1"	Phoenix	AZ	85017	33.5249025	-11
4	8DShNS-LuFqpEWlp0HxijA	"Sports Authority"		"5000 Arizona Mills Cr, Ste 435"	Tempe	AZ	85282	33.3831468	-11
5	PfOCPjBrlQAnz__NXj9h_w	"Brick House Tavern + Tap"		"581 Howe Ave"	Cuyahoga Falls	OH	44221	41.1195346	-81
6	o9eMRCWt5PkpLDE0gOPtcQ	"Messina"		"Richterstr. 11"	Stuttgart	BW	70567	48.7272	9.1

Basic Exploratory Data Analysis

Top cities

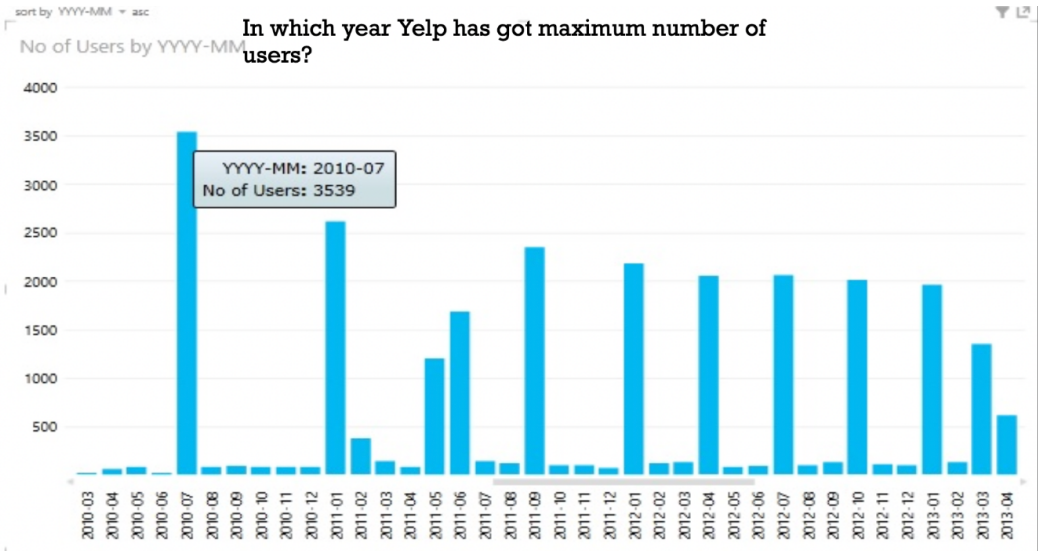


Restaurants with high occurrences

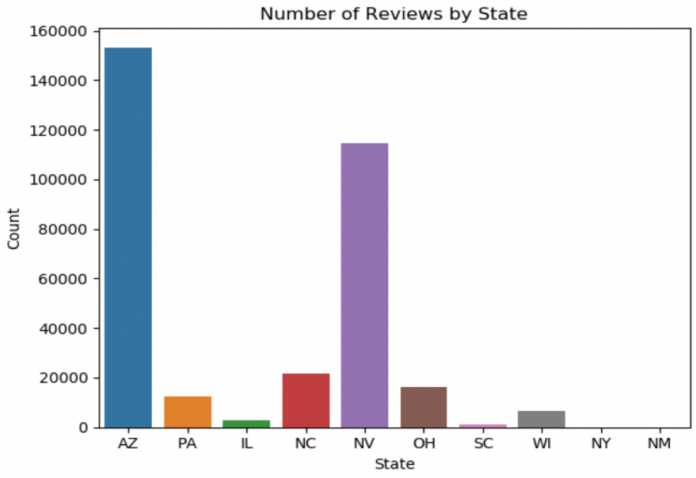


User count

In which year Yelp has got maximum number of users?

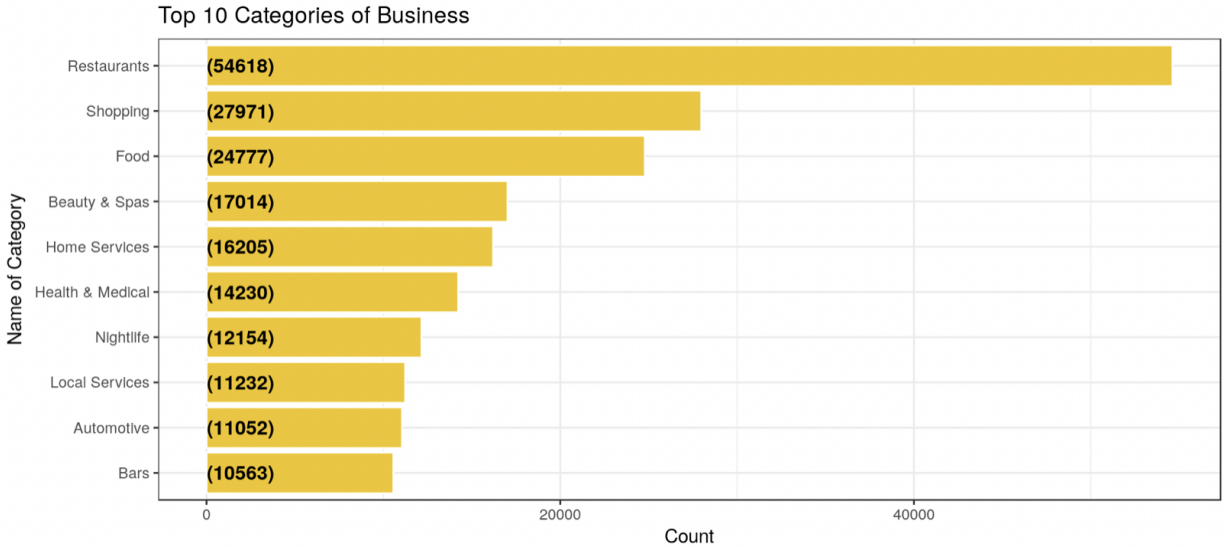


States and their review count



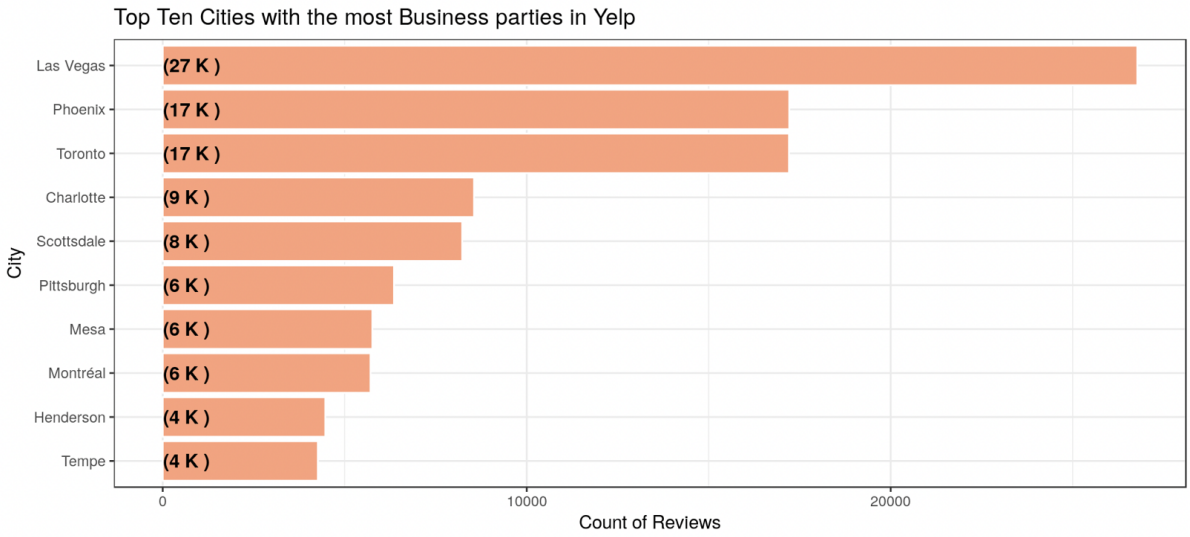
Most Popular Categories

The most popular categories of business are plotted in the bar plot.



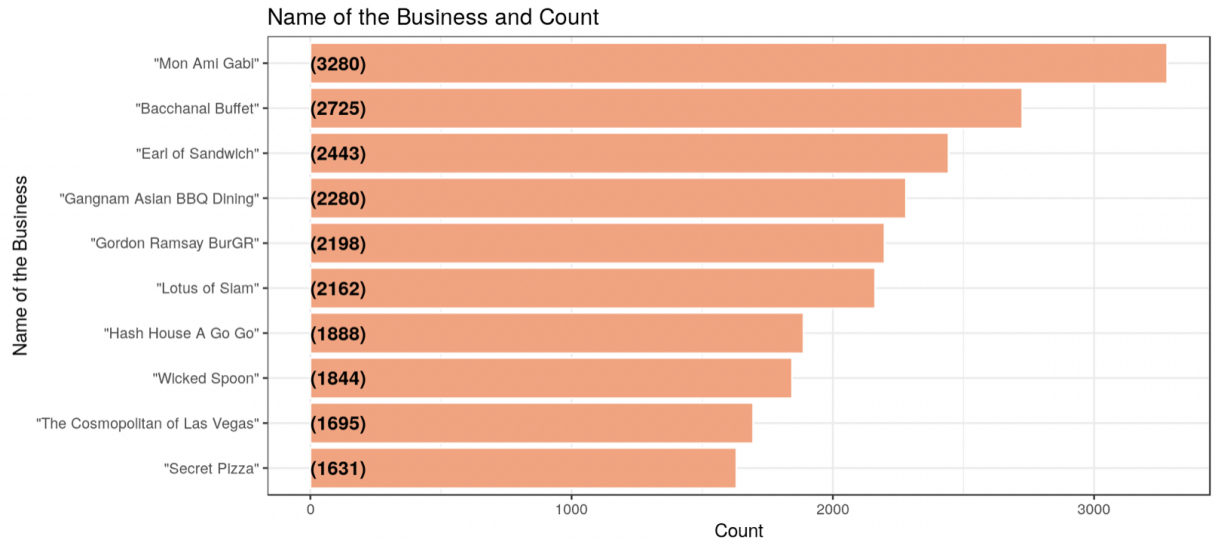
Top Ten Cities with the most Business parties mentioned in Yelp

We show the Top Ten Cities which have the most Business parties mentioned in Yelp.



Business with most Five Star Reviews from Users

The following plot shows the names of businesses with the most Five Star Reviews.**Mon Ami Gabi** and **Bacchanal Buffet** are the Two most popular restaurants from the Yelp reviews with **Five Star** ratings. We will do a deep dive for these restaurants.



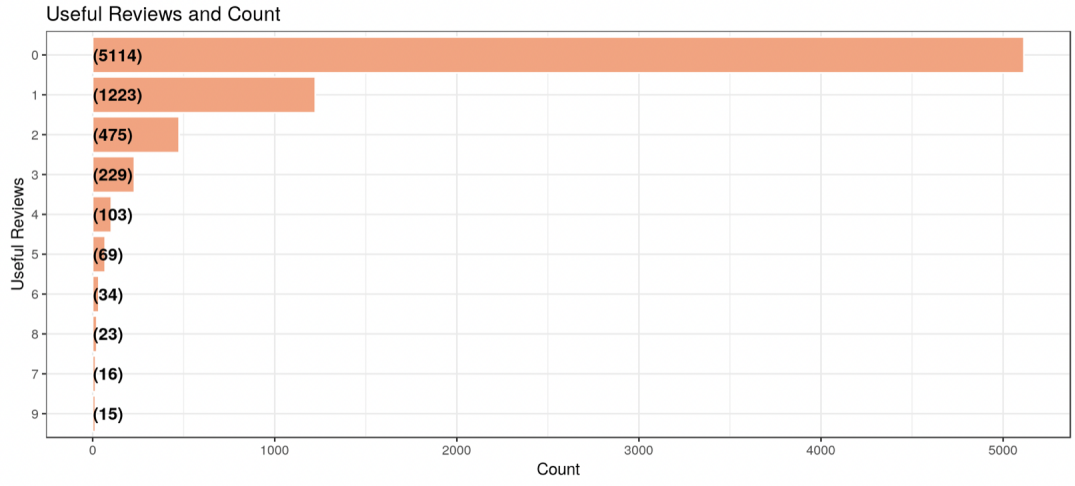
A. **“Mon Ami Gabi”**

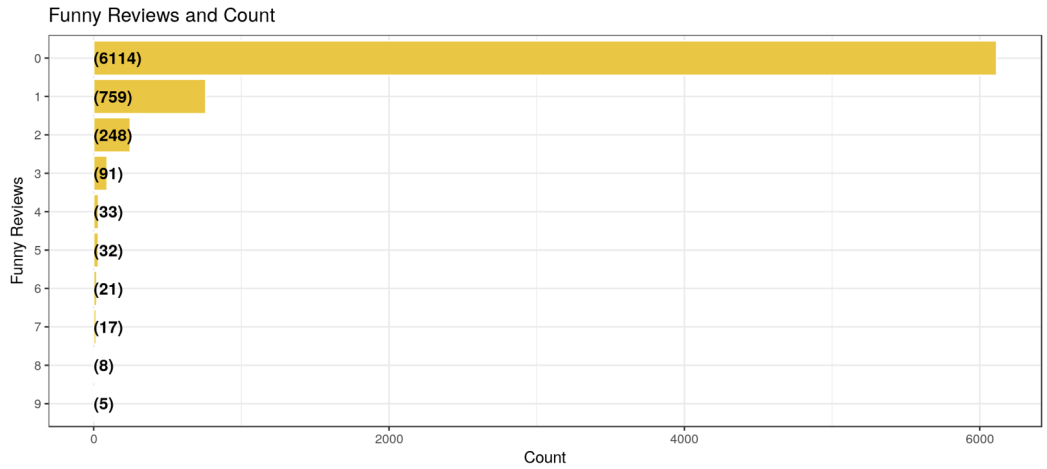
The location and category of the most liked business **Mon Ami Gabi** is shown below

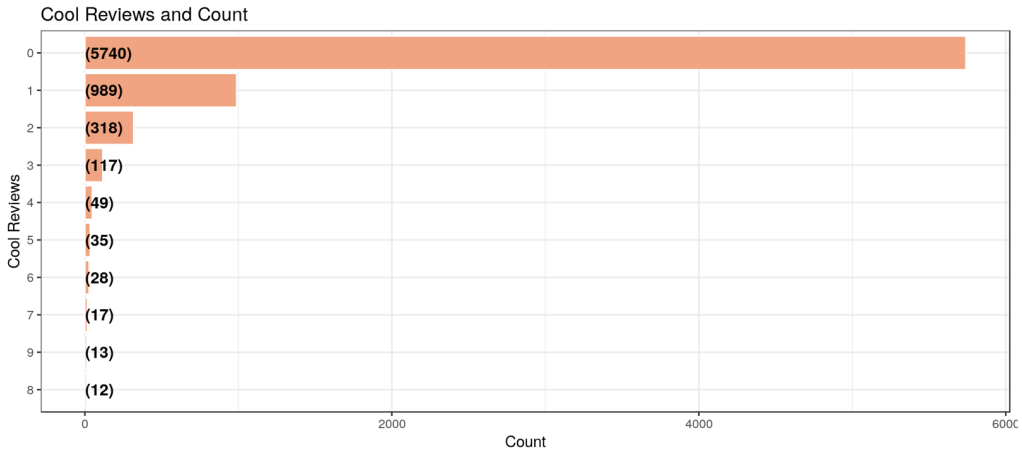
	name	neighborhood	city	state	postal_code	categories
1	"Mon Ami Gabi"	The Strip	Las Vegas	NV	89109	French;Steakhouses;Restaurants;Breakfast & Brunch

Useful, Funny and cool reviews

The following plot describes the number of **Useful, Funny and Cool** reviews.

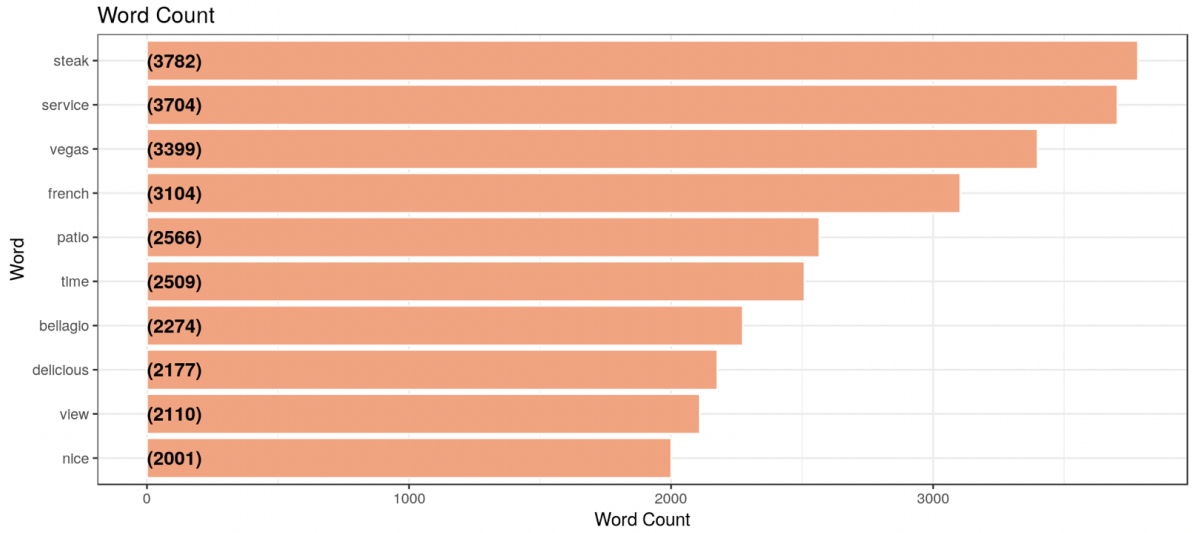






Top Ten most common Words of the business “Mon Ami Gabi”

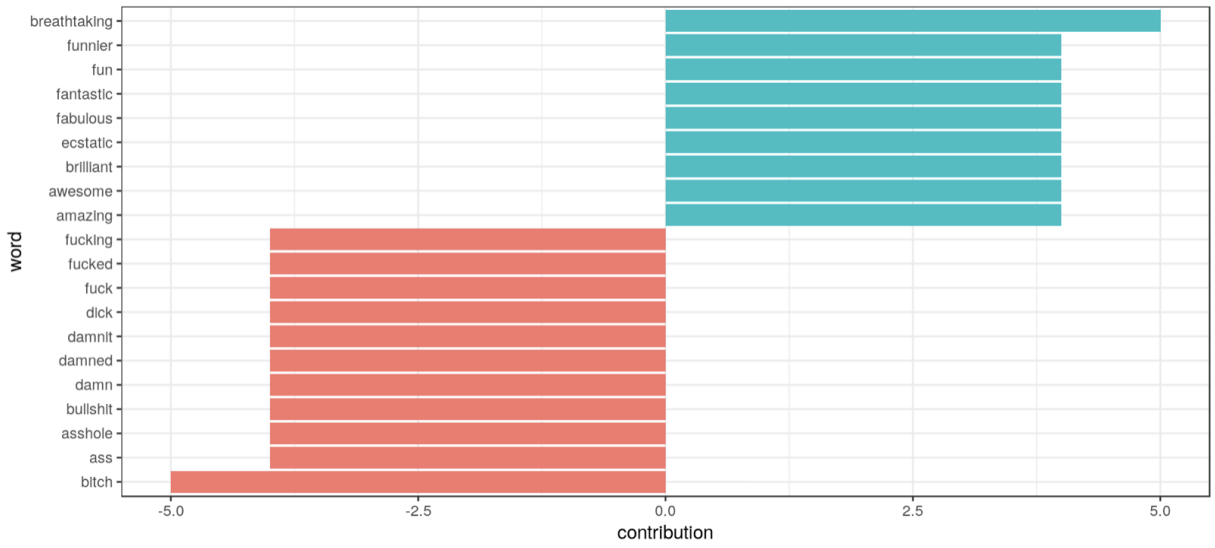
We examine the Top Ten Most Common words and show them in a bar graph. The words steak, service, vegas,french,patio,bellagio,delicious, nice are the words which have been used very frequently in the **reviews**.



Sentiment Analysis - Positive and Not So Positive Words of “Mon Ami Gabi”

We display the Positive and Not So Positive words used by reviewers for the business Mon Ami Gabi.We have used the **AFINN sentiment lexicon**, which provides numeric positivity scores for each word, and visualize it with a bar plot.

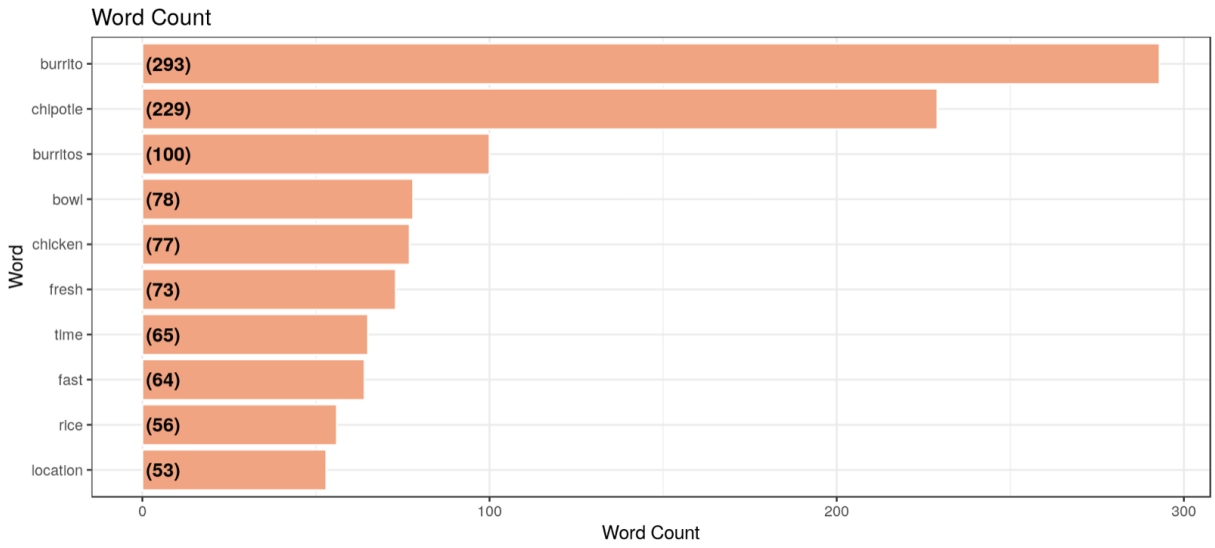
Breathtaking,funnier,fun,fantastic,fabulous,ecstatic,brilliant,awesome,amazing are some of the postive words that we have seen in the reviews of the business.



B. Chipotle Business in Yonge Street Toronto

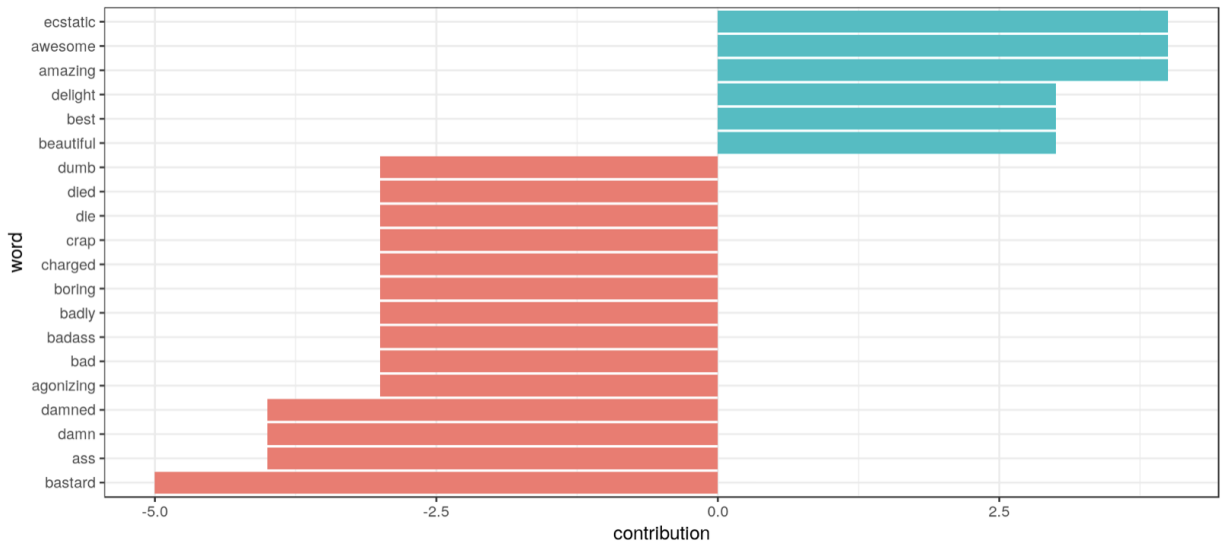
We explore in detail the Chipotle business in Yonge Street Toronto since this has recived the **highest** number of reviews among the Chipotle business.

Top Ten most common Words of the business “Chipotle Business in Yonge Street Toronto”



Sentiment Analysis - Positive and Not So Positive Words of Chipotle Business in Yonge Street Toronto

We display the Positive and Not So Positive words used by reviewers for the business Chipotle Business in Yonge Street Toronto.We have used the **AFINN sentiment lexicon**, which provides numeric positivity scores for each word, and visualize it with a bar plot.



Phoenix City Analysis

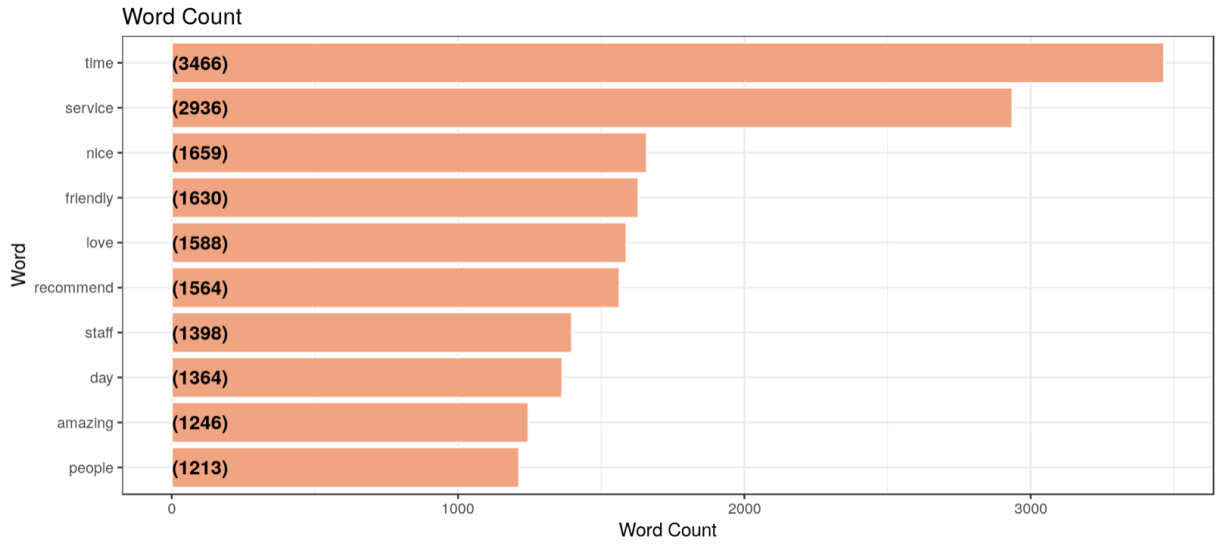
Top Ten Business in Phoenix

We list the Top Ten businesses in Toronto giving importance to the number of reviews and then to the number of stars obtained by the business.

	name	neighborhood	address	review_count	stars
1	"Phoenix Sky Harbor International Airport"		"3400 E Sky Harbor Blvd, Ste 3300"	2215	3
2	"Pizzeria Bianco"		"623 E Adams St"	2035	4
3	"Bobby Q"		"8501 N 27th Ave"	1940	4.5
4	"Lux Central"		"4400 N Central Ave"	1770	4.5
5	"Cibo"		"603 N 5th Ave"	1698	4.5
6	"La Santisima"		"1919 N 16th St"	1694	4
7	"The Arrogant Butcher"		"2 E Jefferson St, Ste 150"	1526	4
8	"Matt's Big Breakfast"		"825 N 1st St"	1520	4
9	"Lo-Lo's Chicken & Waffles"		"1220 S Central Ave"	1519	4
10	"Little Miss BBQ"		"4301 E University Dr"	1463	5

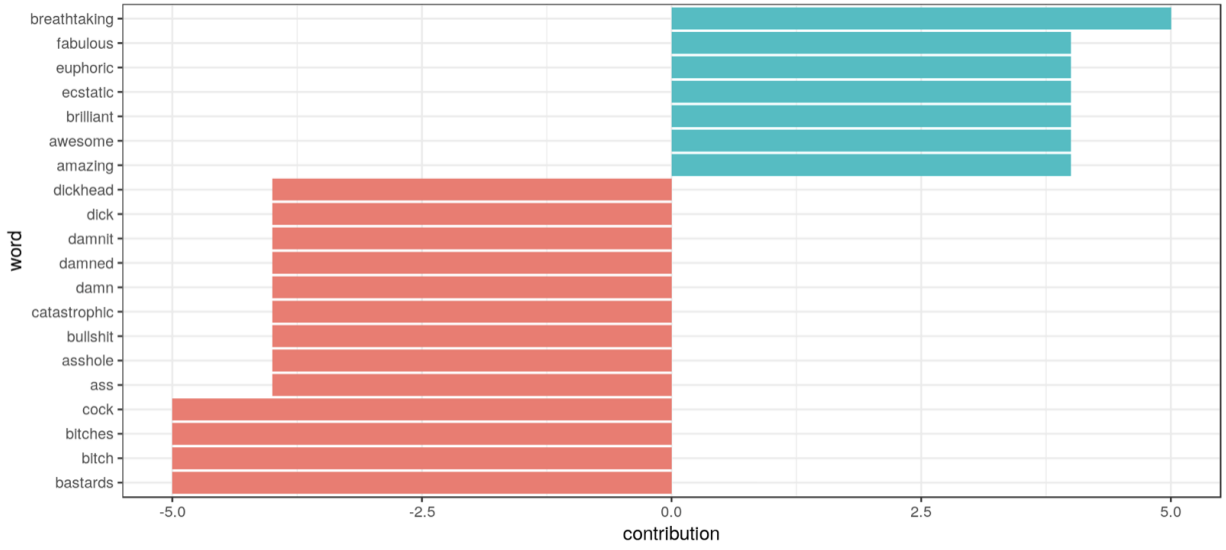
Top Ten most common Words of the business Phoenix City

We examine the Top Ten Most Common words and show them in a bar graph.



Sentiment Analysis - Postive and Not So Postive Words of Phoenix City

We display the Positive and Not So Positive words used by reviewers for Phoenix City.We have used the **AFINN sentiment lexicon**, which provides numeric positivity scores for each word, and visualize it with a bar plot.



Tools Used

We have used Hive and Pig to extract the dataset files from Json to csv format. The queries we have run in Hive and Pig. For the visualisation of the analysis of the output we have used the tableau software and tools.

Conclusion

In this project, our goal is to help consumers and existing and new merchants to use Yelp more efficiently.We have analysed two restaurant businesses: Mon Ami Gabi and Chipotle. Performed text mining and done sentimental analysis. Also for each restaurant we have figured which is the most talked about dish of the restaurant. They also can have a better overview about the restaurants on features that the restaurants are famous for or needs to improve on by looking at the keywords, which helps them to choose places that they would mostly like to dine in.

Future Scope

- Creating a predictive model with the help of our data analysis
- Making a recommendation system
- Explore the hidden patterns and use Natural Language Processing
- Fake review detection
- Information retrieval based on true reviews
- User interface
- Optimised efficiency of the model
- Exploring other database options such as MongoDB

References:

[1.]AnArticlefoundonTowardsDataScienceblog-
[https://towardsdatascience.com/tagged/yelp]
[2.]AnarticledetailingonSentimentanalysisandopinionmining-[https://www.morganclaypool.com/doi/abs/10.2200/s00416ed1v01y201204hlt016]
[3.]AnarticleaboutNaturalProcessingLanguageonHadoop-
[https://arxiv.org/pdf/1608.04434.pdf]