yelp

Recommendation System

Big Data – Fall Quarter 2019

Carrie Lu | Daniela Matinho | Hanna Kerr | Yuling Gu

Agenda

- Yelp company
- Business Problem
- Data: Exploratory analysis
- Data: Sentiment analysis
- Modeling & Model tuning
- Findings
- Future Work

Yelp

Real people, Real reviews

Yelp is a business directory service and crowd-sourced review forum, 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.

Business Problem

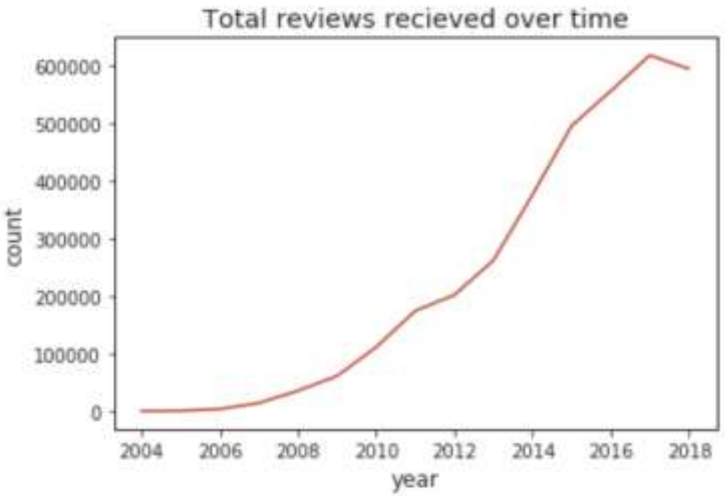
The goal of our project is to build a recommendation engine to recommend restaurants to Yelp users.

Data

- Yelp Data (8GB in Json format) – Kaggle
- Datasets used – Business, Reviews and User
- Store in RCC (HDFS) and then we move it to Google Cloud (1 Master node & 2 worker nodes)
- Data cleaning, exploratory & modeling – Pyspark

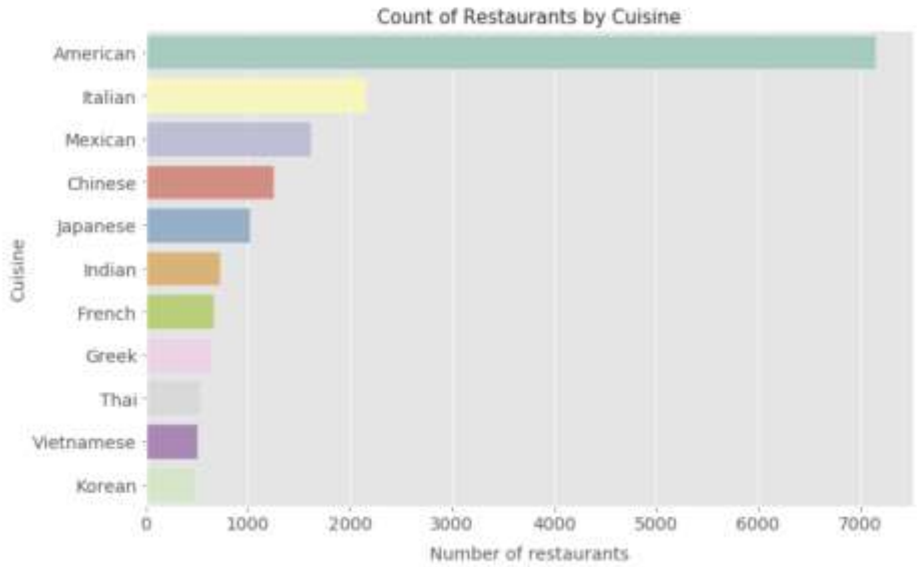
Data Cleaning

- Restaurants data
- Data in the US

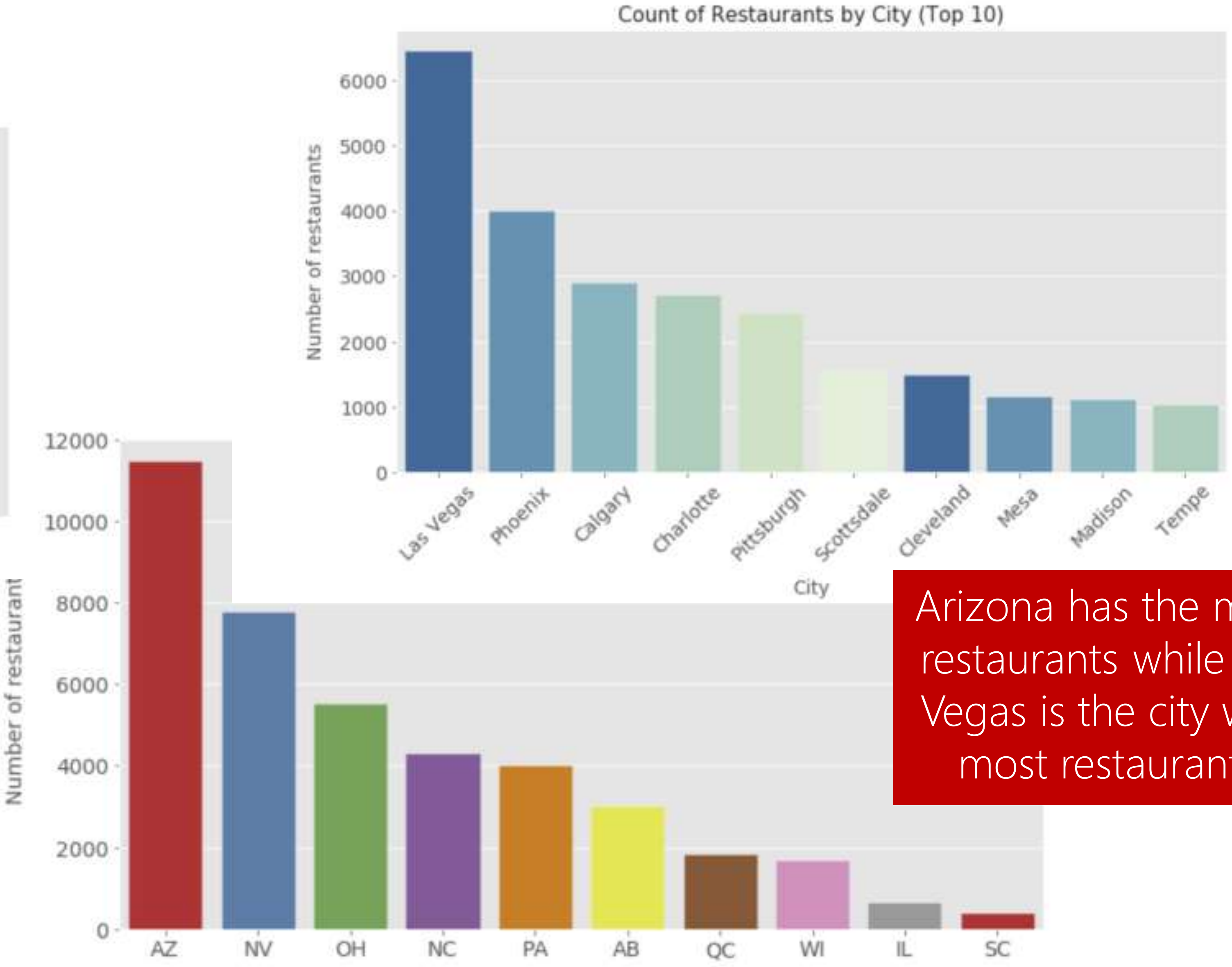


11840 open restaurants

Exploratory Analysis - Business



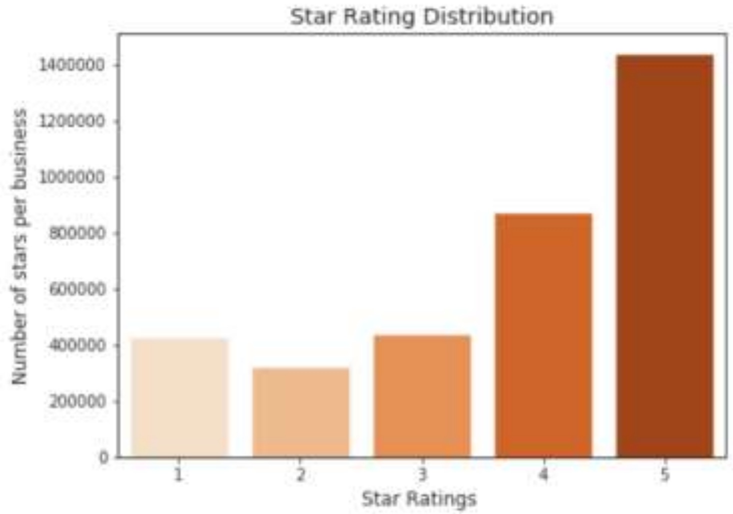
Most restaurants in the dataset are American cuisine followed by Italian and Mexican food

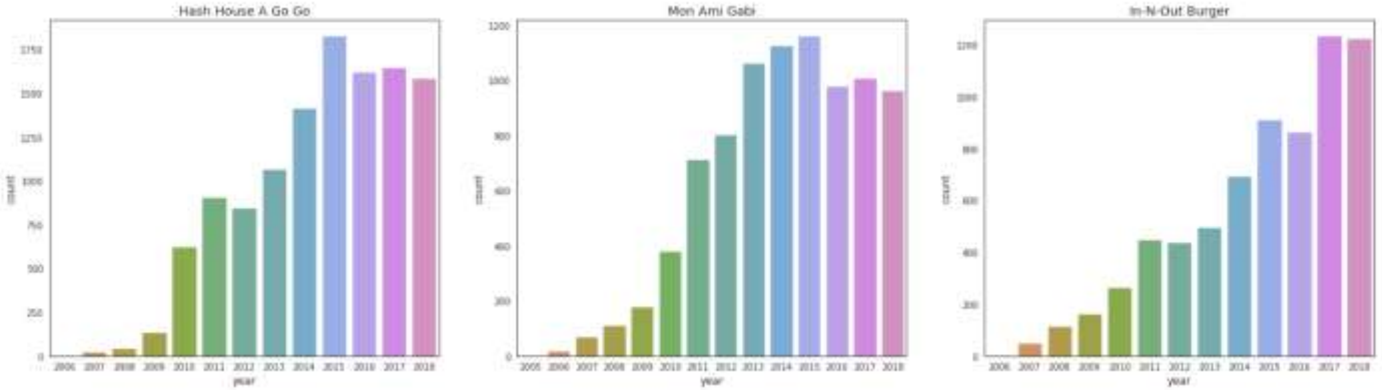


Arizona has the most restaurants while Las Vegas is the city with most restaurants

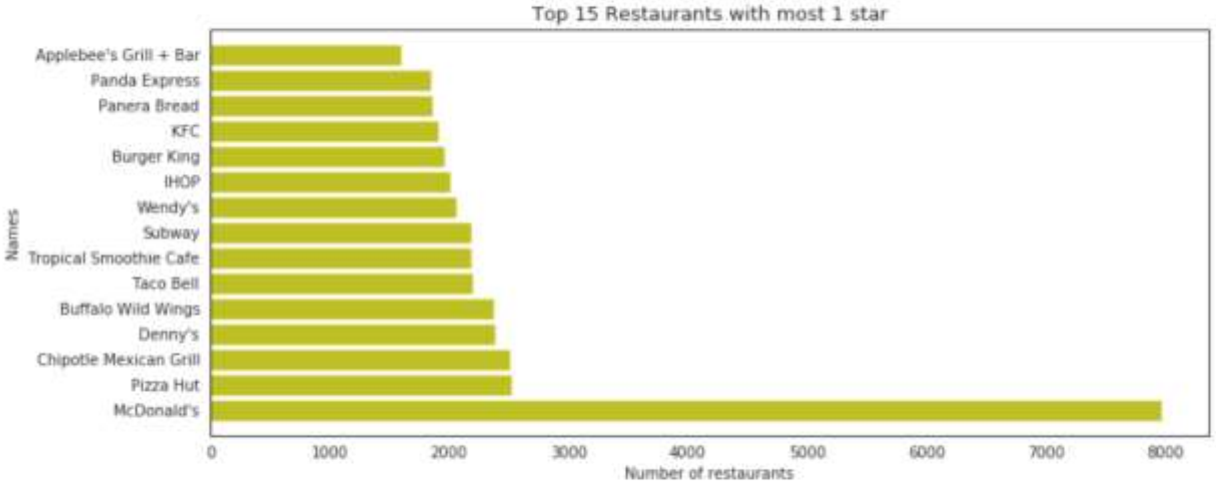
Exploratory Analysis - Reviews

Most reviews on the app have 5 stars ratings





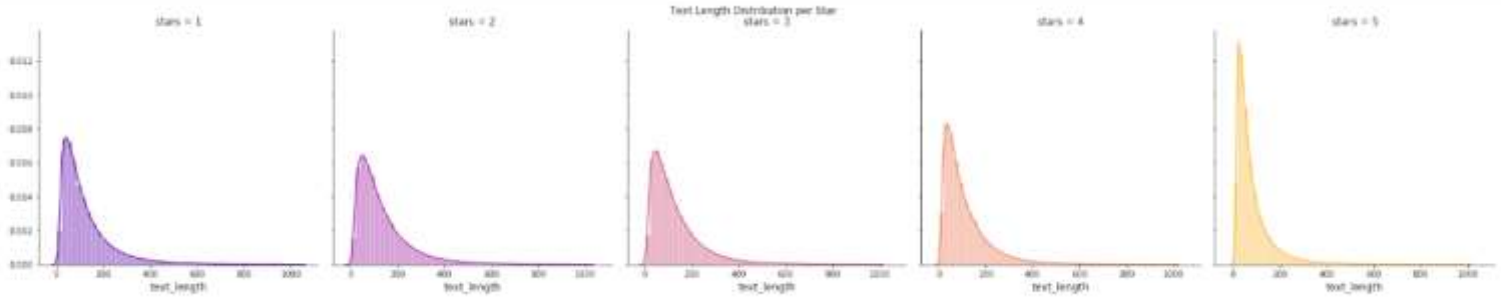
Top 3 restaurants in terms of reviews received over time



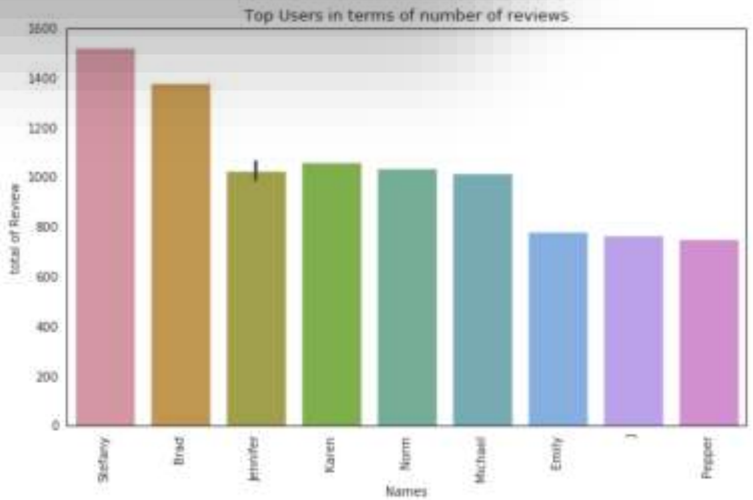
McDonald's is the least appreciated restaurant in terms of stars

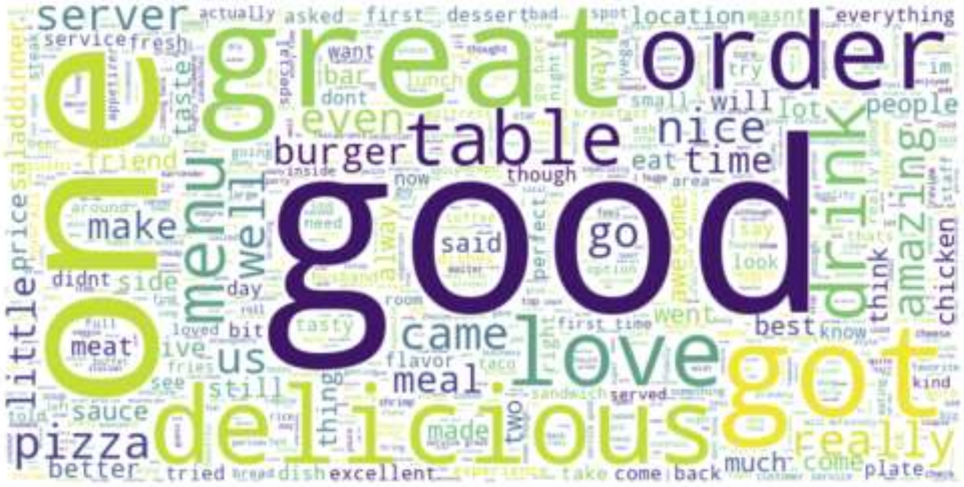
Exploratory Analysis – Reviews & User

5 stars restaurants - people tend to write shorter reviews

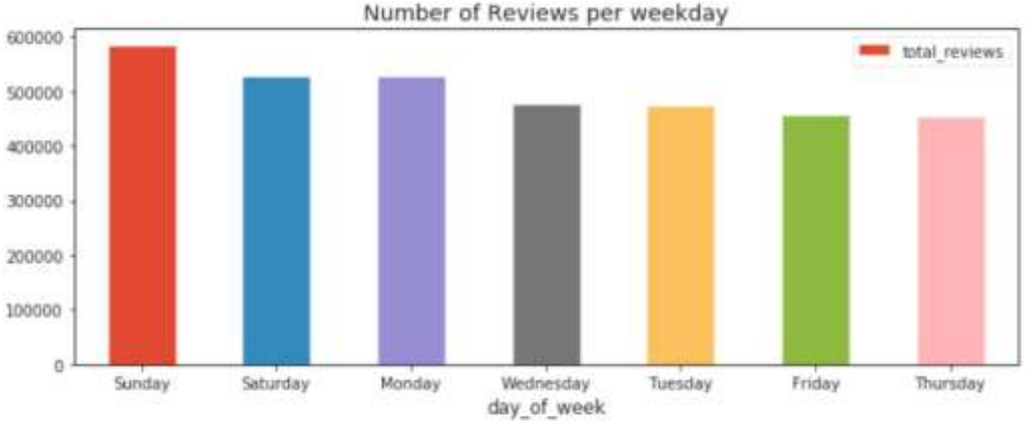


Stefany has posted the largest number of reviews





Sunday - day of the week with more reviews

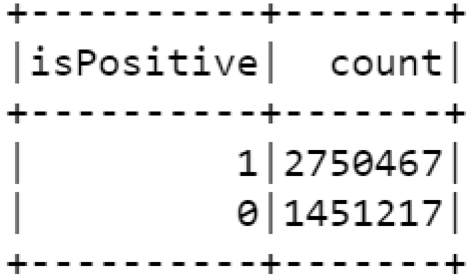


Sentiment Analysis

1

Cleaning steps:

- Change everything to lower case
- Remove punctuation
- Stop words
- Tokenize sentences into words



```
+----------+-------+
|isPositive|  count|
+----------+-------+
|         1|2750467|
|         0|1451217|
+----------+-------+
```

2

Defining values:

- Y value: > = 4 stars - Positive review (1)
- Y value: < 4 stars - Negative review (0)

65% of the reviews' restaurants received are positive

HashingTF VS CountVectorizer

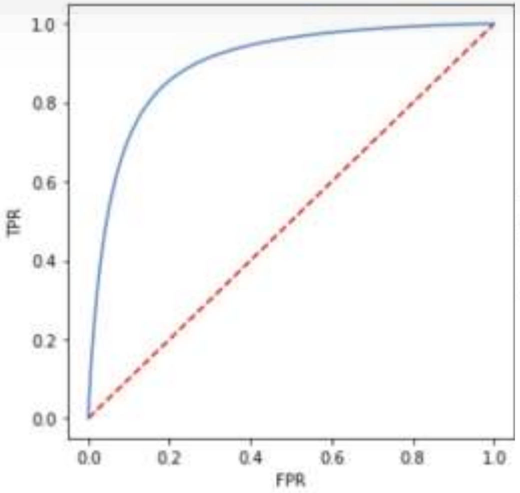
- HashingTF is a transformer that takes sets of terms (bag of words) and converts those sets into fixed-length feature vectors. A raw feature is mapped into an index (term) by applying a hash function. HashingTF is irreversible, meaning we can't restore original input from a hash vector. This requires only a single scan and no additional memory.
- CountVectorizer provides a simple way to both tokenize a collection of text documents and build a vocabulary of known words, but also encode new documents using that vocabulary. CountVectorizer with model (index) can be used to restore unordered input, meaning models can be harder to interpret. This requires additional scan over the data to build a model and additional memory to store vocabulary (index).

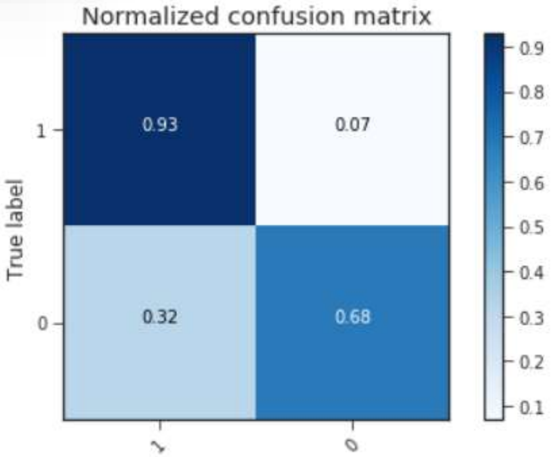
Sentiment Analysis

HashingTF

Logistic Regression Classification

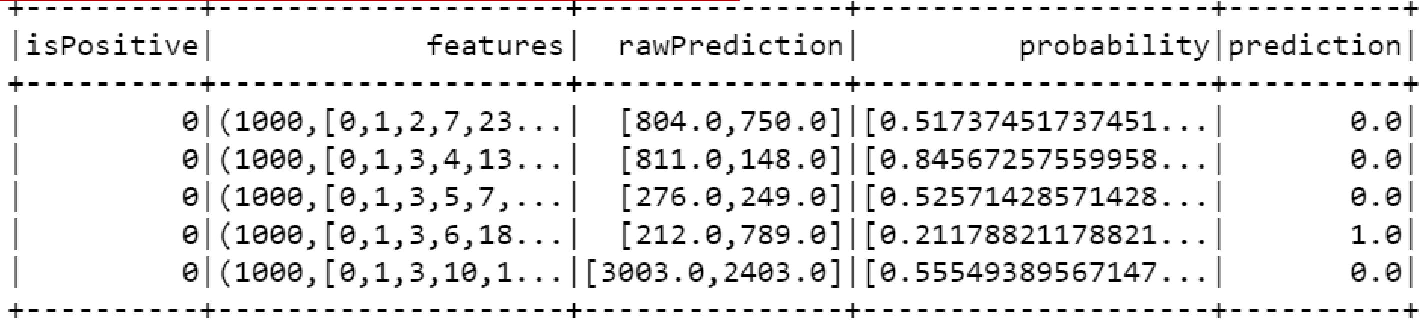
Accuracy: 84.6% | F1 score: 84.2%





Decision Tree Classification

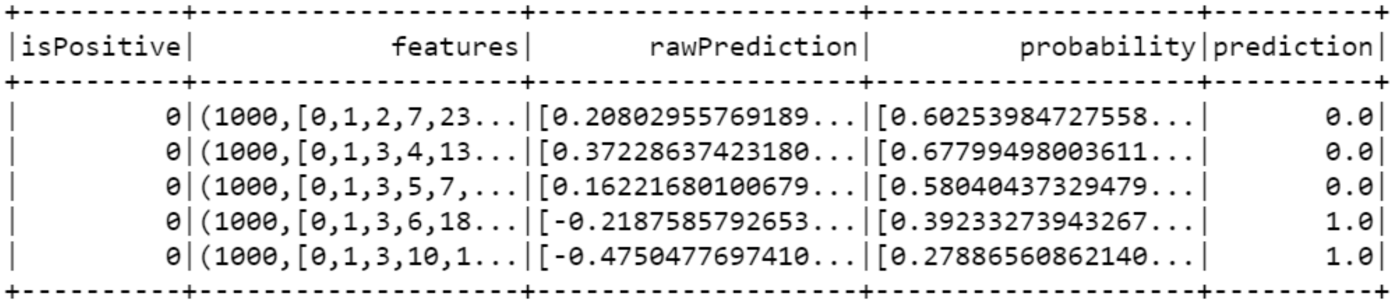
Accuracy: 73.5% | F1 score: 73.8%



```
+----------+--------------------+-------------+--------------------+----------+
|isPositive|            features| rawPrediction|         probability|prediction|
+----------+--------------------+-------------+--------------------+----------+
|         0|(1000,[0,1,2,7,23...|  [804.0,750.0]|[0.51737451737451...|       0.0|
|         0|(1000,[0,1,3,4,13...|  [811.0,148.0]|[0.84567257559958...|       0.0|
|         0|(1000,[0,1,3,5,7,...|  [276.0,249.0]|[0.52571428571428...|       0.0|
|         0|(1000,[0,1,3,6,18...|  [212.0,789.0]|[0.21178821178821...|       1.0|
|         0|(1000,[0,1,3,10,1...|[3003.0,2403.0]|[0.55549389567147...|       0.0|
+----------+--------------------+-------------+--------------------+----------+
```

Gradient Boost Classifier

Accuracy: 76.5% | F1 score: 75.1%



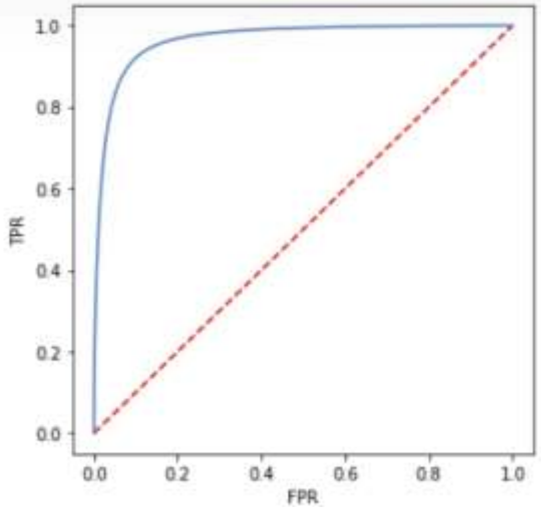
```
+----------+--------------------+--------------------+--------------------+----------+
|isPositive|            features|       rawPrediction|         probability|prediction|
+----------+--------------------+--------------------+--------------------+----------+
|         0|(1000,[0,1,2,7,23...|[0.20802955769189...|[0.60253984727558...|       0.0|
|         0|(1000,[0,1,3,4,13...|[0.37228637423180...|[0.67799498003611...|       0.0|
|         0|(1000,[0,1,3,5,7,...|[0.16221680100679...|[0.58040437329479...|       0.0|
|         0|(1000,[0,1,3,6,18...|[-0.2187585792653...|[0.39233273943267...|       1.0|
|         0|(1000,[0,1,3,10,1...|[-0.4750477697410...|[0.27886560862140...|       1.0|
+----------+--------------------+--------------------+--------------------+----------+
```

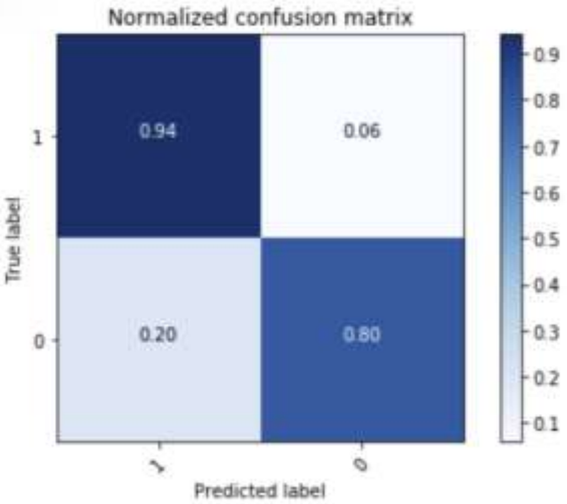
Sentiment Analysis

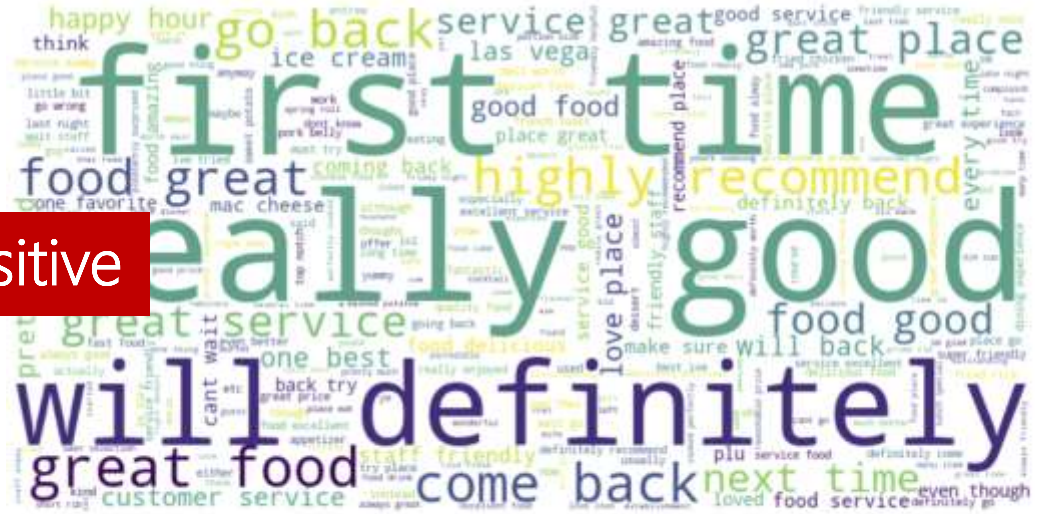
CountVectorizer

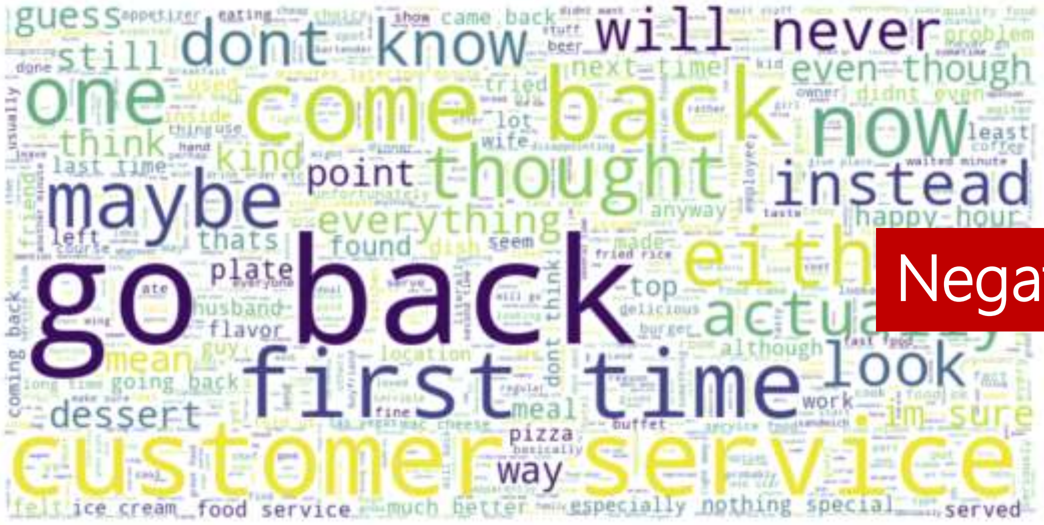
Logistic Regression Classification

Accuracy: 89.3% | F1 score: 89.2%



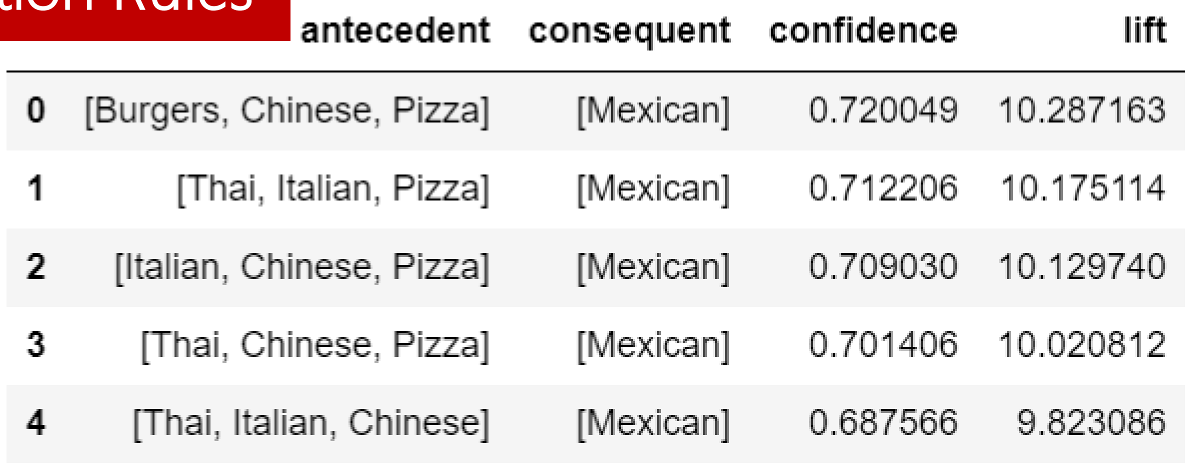



Positive


Negative

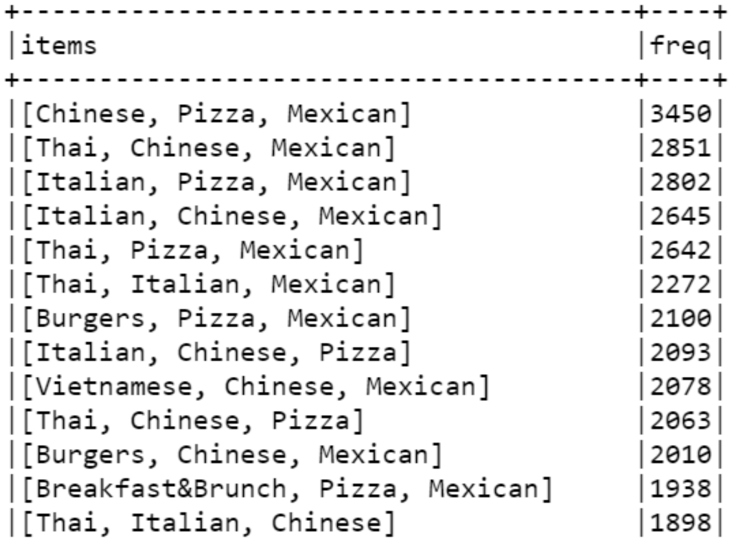
Association Mining

Association Rules



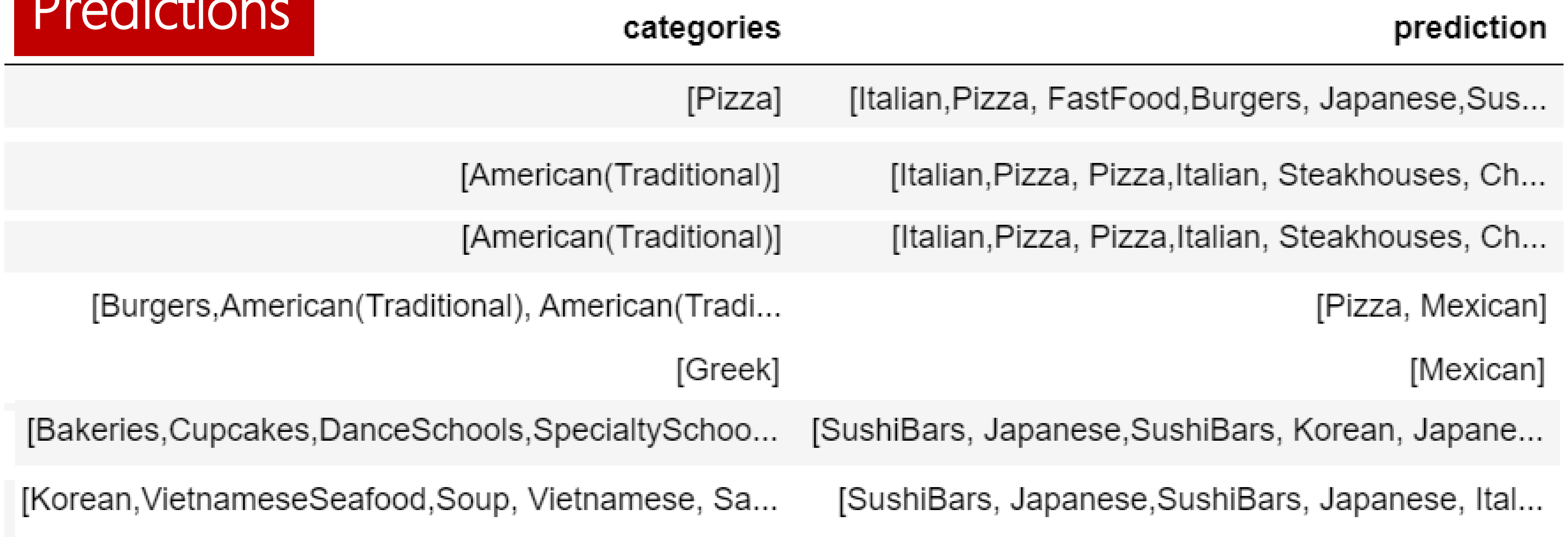
	antecedent	consequent	confidence	lift
0	[Burgers, Chinese, Pizza]	[Mexican]	0.720049	10.287163
1	[Thai, Italian, Pizza]	[Mexican]	0.712206	10.175114
2	[Italian, Chinese, Pizza]	[Mexican]	0.709030	10.129740
3	[Thai, Chinese, Pizza]	[Mexican]	0.701406	10.020812
4	[Thai, Italian, Chinese]	[Mexican]	0.687566	9.823086

Frequent Items



```
+----------------------------------------+----+
|items                                   |freq|
+----------------------------------------+----+
|[Chinese, Pizza, Mexican]               |3450|
|[Thai, Chinese, Mexican]                |2851|
|[Italian, Pizza, Mexican]               |2802|
|[Italian, Chinese, Mexican]             |2645|
|[Thai, Pizza, Mexican]                  |2642|
|[Thai, Italian, Mexican]                |2272|
|[Burgers, Pizza, Mexican]               |2100|
|[Italian, Chinese, Pizza]               |2093|
|[Vietnamese, Chinese, Mexican]          |2078|
|[Thai, Chinese, Pizza]                  |2063|
|[Burgers, Chinese, Mexican]             |2010|
|[Breakfast&Brunch, Pizza, Mexican]      |1938|
|[Thai, Italian, Chinese]                |1898|
```

Predictions



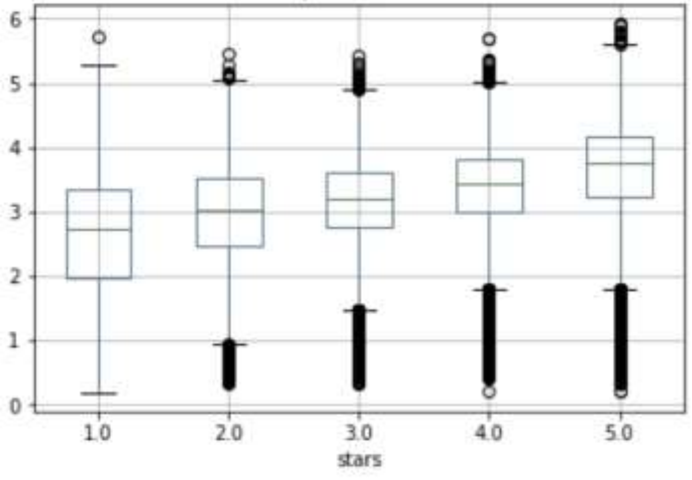
categories	prediction
[Pizza]	[Italian,Pizza, FastFood,Burgers, Japanese,Sus...
[American(Traditional)]	[Italian,Pizza, Pizza,Italian, Steakhouses, Ch...
[American(Traditional)]	[Italian,Pizza, Pizza,Italian, Steakhouses, Ch...
[Burgers,American(Traditional), American(Tradi...	[Pizza, Mexican]
[Greek]	[Mexican]
[Bakeries,Cupcakes,DanceSchools,SpecialtySchoo...	[SushiBars, Japanese,SushiBars, Korean, Japane...
[Korean,VietnameseSeafood,Soup, Vietnamese, Sa...	[SushiBars, Japanese,SushiBars, Japanese, Ital...

Alternative Least Squares

- ALS is an iterative optimization process where we for every iteration try to arrive closer to a factorized representation of the original data. The algorithm is based on “Collaborative Filtering for Implicit Feedback Datasets”.

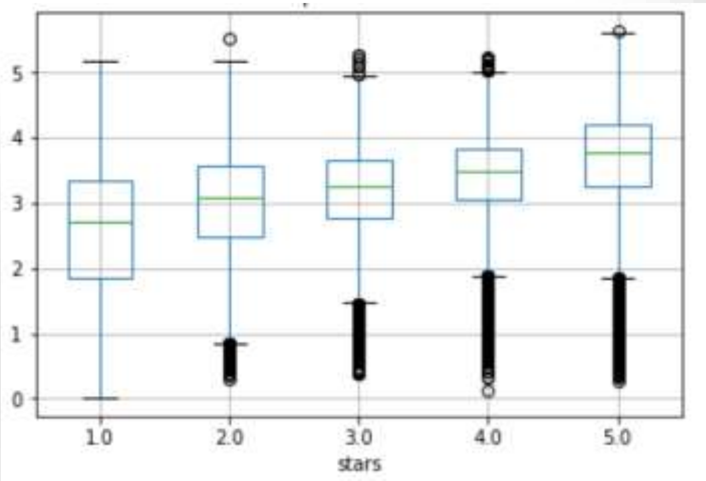
Las Vegas

RMSE: 1.374 | r2: -0.097



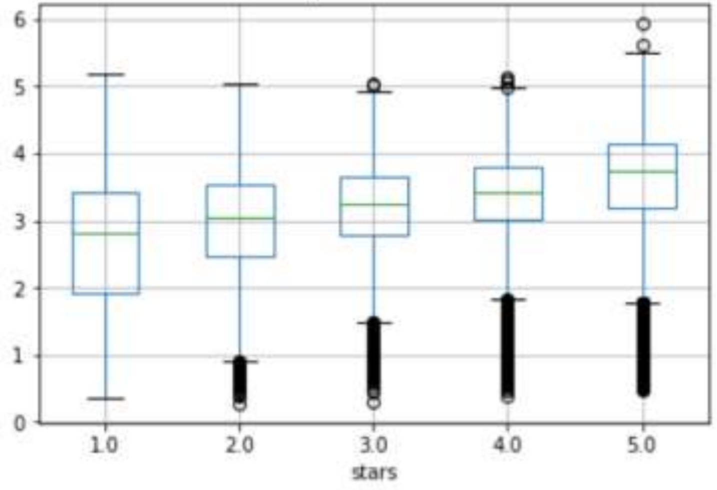
Phoenix

RMSE: 1.366 | r2: -0.102



Scottsdale

RMSE: 1.392 | r2: -0.180



Alternative Least Squares

Recommendation for Restaurants and Users

Albinas Italian American Bakery

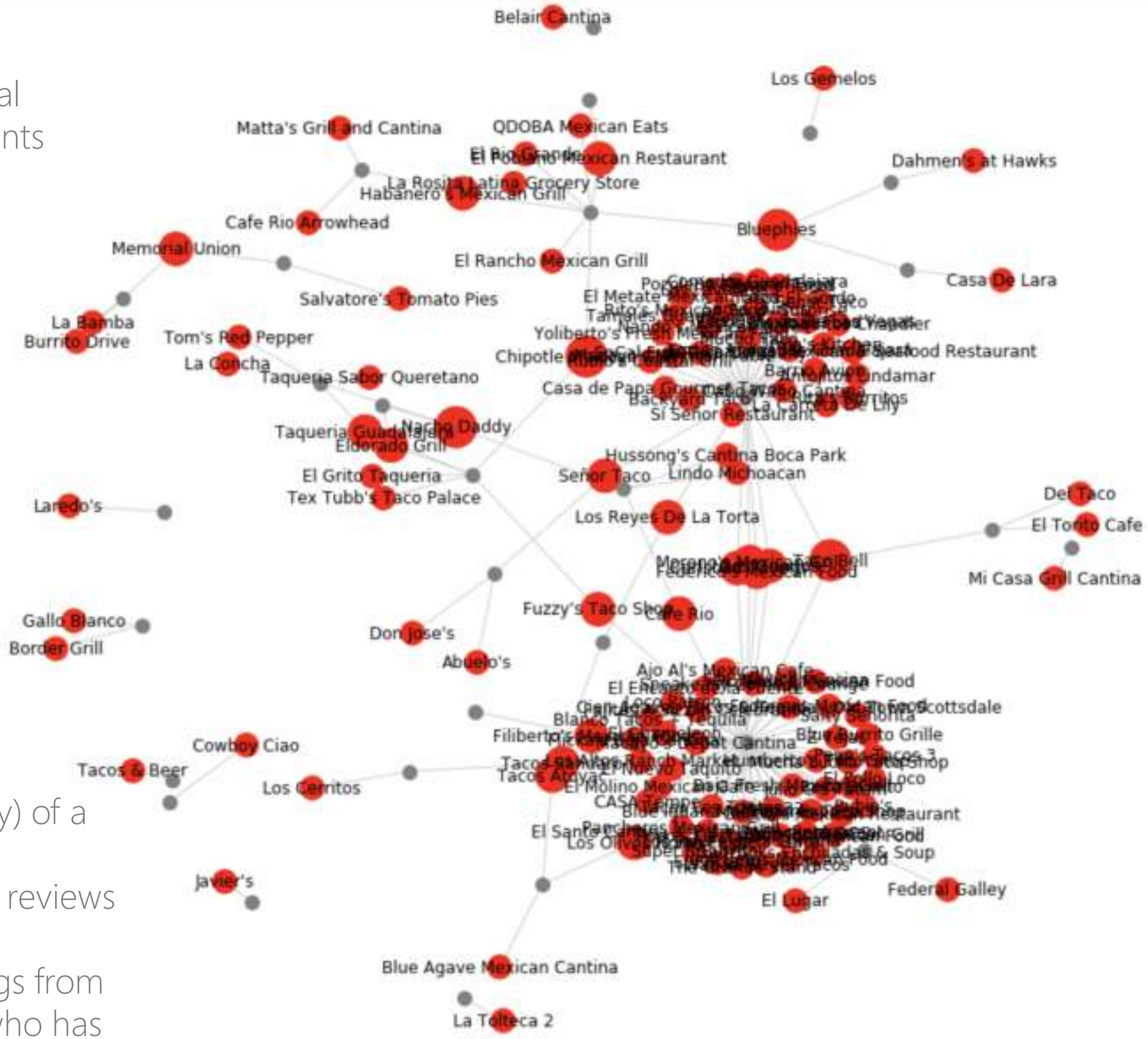
- Paul
- Kastle
- Cynthia
- Shanna
- Brooke
- Patrick
- Johan
- Aimee
- Paul
- Danyelle

Allan

- BBQ Concepts
- Fernandez Hot-Dogs
- The Steakhouse at Treasures
- Tacos LV
- Southern Kitchen
- Pretzel Time
- Potato Valley Café
- Pho Ha Noi
- China Gourmet
- Winter In July

Graphs

- This allows the user to have a visual representation of popular restaurants among your friends
- Here we can filter on a restaurant category and # stars per review



- Each node represents a friend (grey) of a user and a restaurant (red)
- The larger the red nodes the more reviews from friends
- This graphic shows 4 and 5*s ratings from Mexican restaurants for one user who has ~200 friends

Findings

- Business:
  - Sentiment Analysis can help businesses to quickly classify as good or bad review without looking at the text
  - ALS can help restaurants targeting the users that are more likely to go to the restaurant

- Users:
  - Association mining can give recommendations to users based on the preference of restaurants
  - ALS can give users the restaurants that align with their reviews
  - Graphs can help users to know what type of restaurants they will like according to their friend’s preferences

Work Limitations & Future Work

- Resources limitations - RCC and Google Cloud (we end had to limit the amount of data to be able to run the models)
- Package limitations with RCC
- Visualization limitations with PySpark

- Further cleaning restaurant categories i.e. clusters
- Use geolocation to give more precise recommendations
- Use Page Rank to rank users giving more weight to their reviews

Thank You

#enjoy vacation with Yelp