Yelp Restaurant Recommendation

Team 2: Xiran Li, Sam Ding, Yun Xing

Executive Summary

Business Objective	Enhance user experience through personalized recommendations and collaborate with restaurants on Yelp
Data & Tools	**Data**: The Yelp datasets (**5GB** combined) Link here **Tools**: GCP and PySpark
EDA	Overview of the datasets + dive into details & interesting findings
ML Models	**Recommender System**: two separate algorithms to recommend top users for restaurants and top restaurants for users **NLP: sentiment analysis** for tip text, helping restaurant owners swiftly detect positive or negative feedback
Future Goal	Give users best recommendations that attained to their needs and help business make informed decisions

Agenda

1. Introduction
2. Data Intro
3. Sentiment Analysis
4. Recommender
5. Discussion
6. Next Steps

Business Problem

Yelp, a popular online platform for discovering and reviewing local businesses, relies heavily on user-generated content such as reviews and tips to provide valuable insights to its users. Yelp aims to further enhance user experience and engagement on its platform

We want to provide **personalized recommendations** for users and also **collaborate with restaurants**, as they are also an important stakeholder

Data Introduction

What is Our Data?

A subset of businesses on Yelp across the US, as well as reviews, tips, check-ins, and users associated with these businesses.

We chose **Business**, **Review**, **User**, and **Tip** datasets.

Data Size

Combined size of 5GB for all datasets. More than 150k business, along with 2M users, 7M reviews, and 900k tips.

Programs & Tools

We mainly used **Google Cloud Platform (GCP)** and **PySpark** for our data analysis and modeling.

Our Data Pipeline

1

Data Import

- Google Cloud Storage (GCS)
- Dataproc Cluster
- BigQuery

2

Data Wrangling

- Remove NA values
- Splitting categories
- Repartition to 40

3

Modeling / Analysis

- PySpark.ml
- SparkNLP
- BigQuery

4

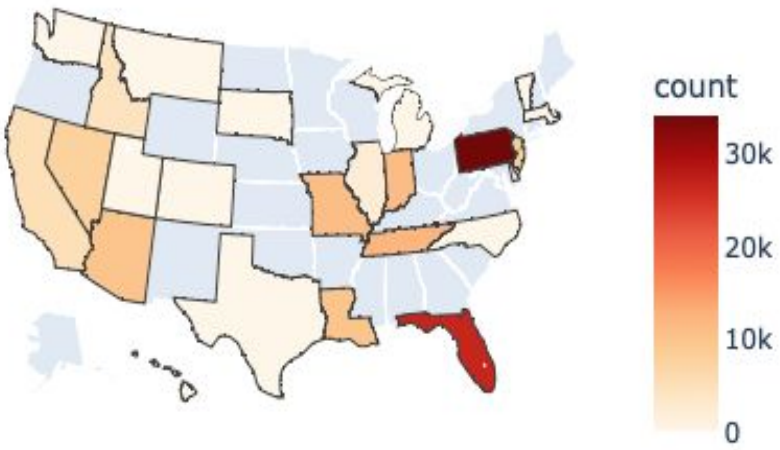
Plot Output

- Matplotlib
- Plotly

Businesses overview

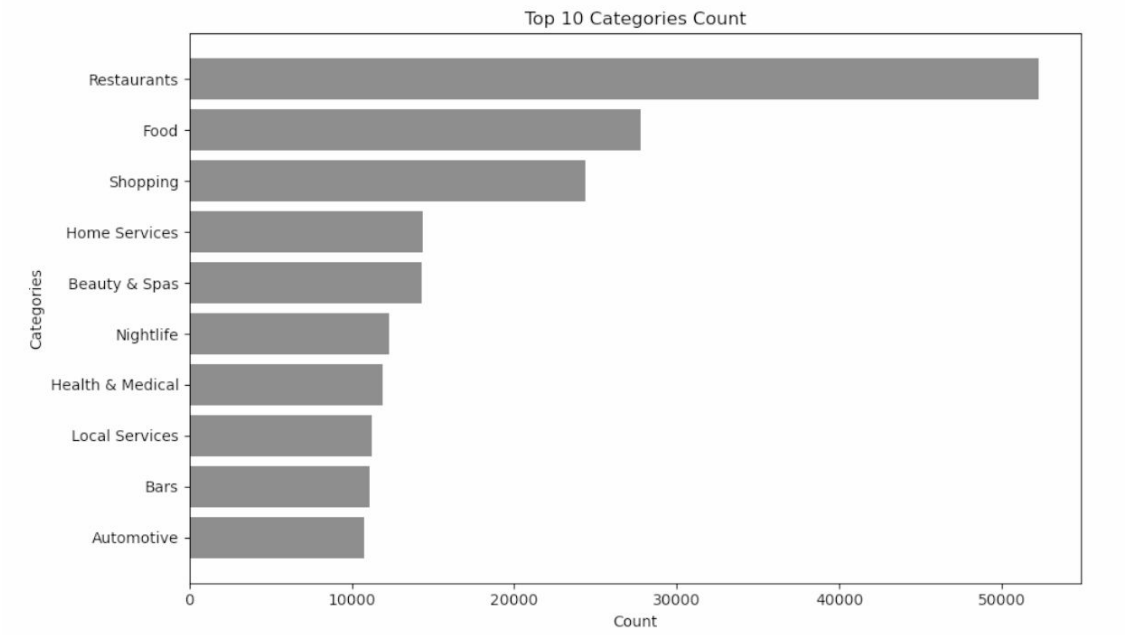
- **150k businesses** spanning 1300+ categories
- A business can be tagged with multiple different categories
- Majority of businesses are found in Pennsylvania and Florida

Choropleth of Geographic Distribution of Business

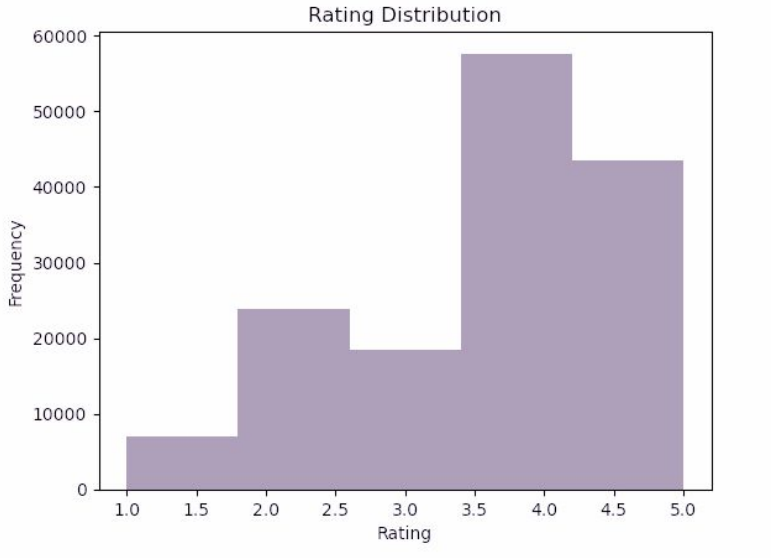


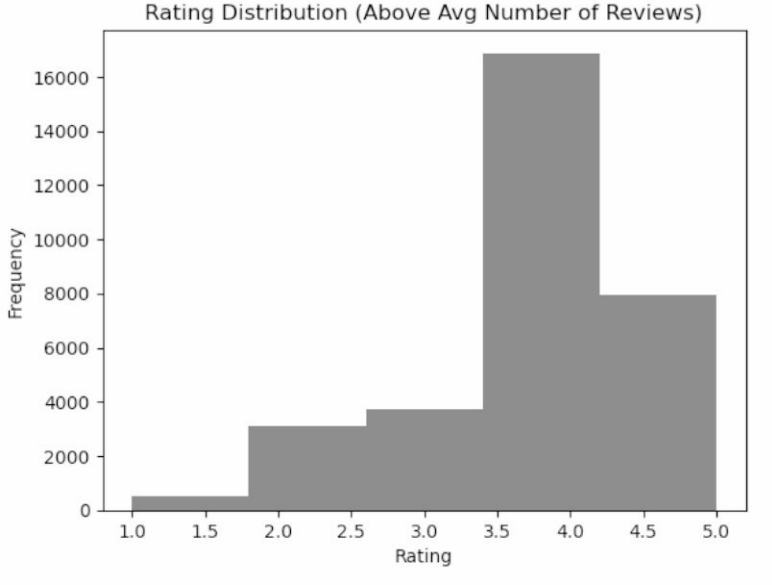
Restaurants take up majority of businesses

- Sometimes different category tags may represent similar purposes. (e.g. Restaurants vs Food, Nightlife vs Bars)
- Majority of business in this data are food/restaurants.



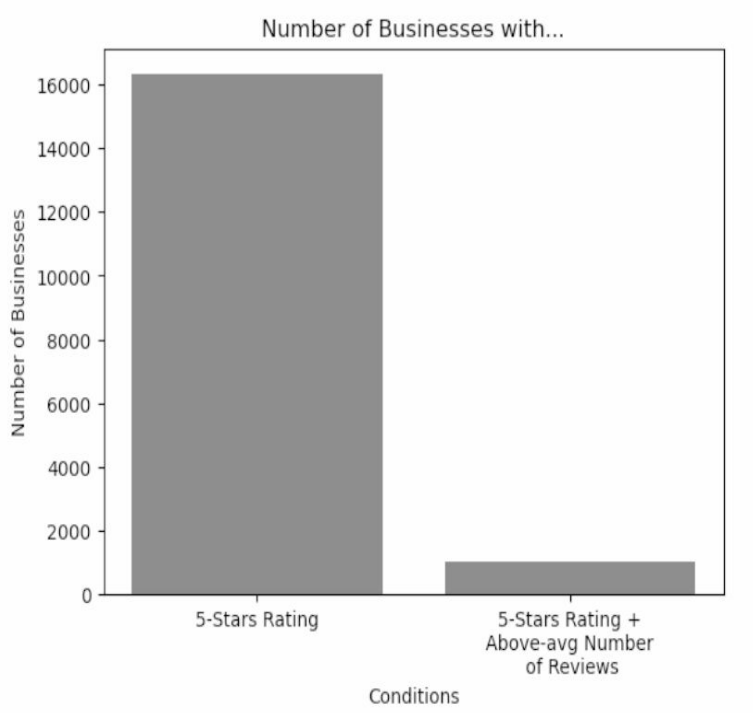
Rating distribution





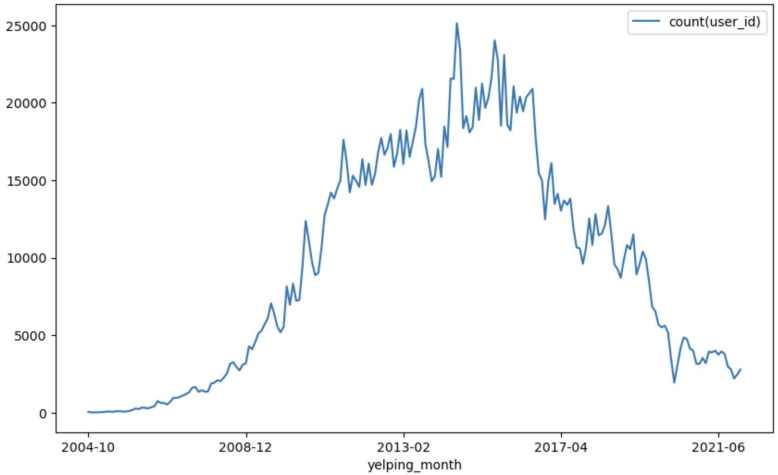
Those with five-stars...

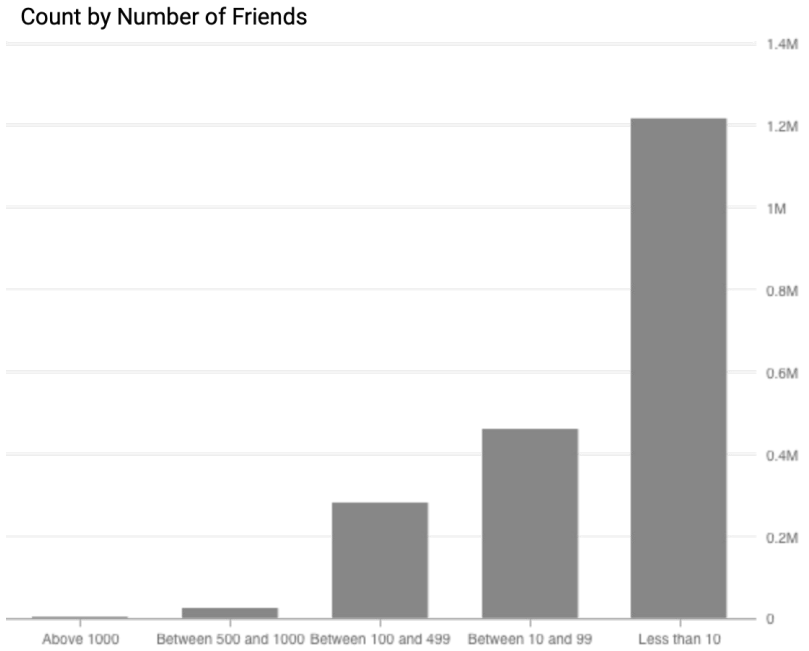
- **Majority (90%)** of the 5-star businesses have lower than average number of reviews.
- **Insights**: as business grows, it is harder to control the consistency of product or service, let alone consumer's personal preference.



User Overview

- Nearly **2M users**
- Majority of the users have less than 10 friends on Yelp





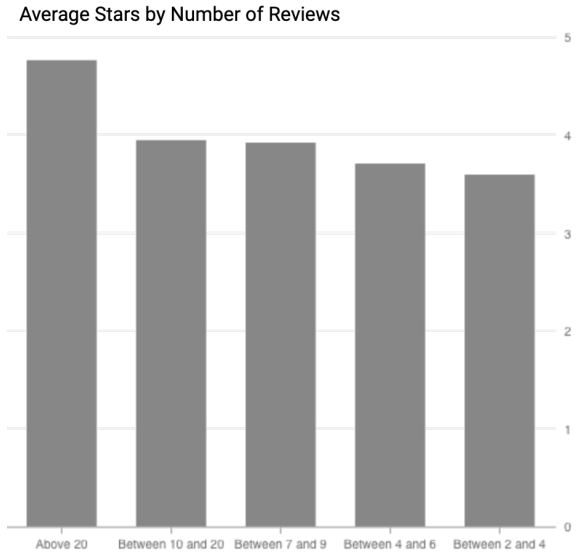
Review Overview

- Nearly **7M reviews** for 150k businesses
- Reviews consist of rating text, stars given by users, and some other tags like ‘cool’, ‘funny’, and ‘useful’
- Reviews are mainly **subjective**

WordCloud of a sample of Reviews

How do repeated customers review?

- We found that some users **give same businesses repeated reviews**, which takes up about **8%** of all reviews.
- It appears that the more reviews a customer repeatedly make to a business, the higher the average rating of the stores are



Tips Overview

- **Over 900k tips in this dataset for over 100k businesses**
- According to Yelp, tips are meant to pass along key information about a business without going into a full review. As such, there is no rating system available for Tips
- Many tips are **actually subjective**
- We thought it would be more accurate to combine information from both Reviews and Tips to determine whether a restaurant is good or bad

WordCloud of a sample of Tips

Sentiment Analysis - Business Context

Goal — Build an assessment framework for the Yelp tip

Solution — Create a sentiment analysis model using NLP techniques for **tip** text data, with training based on Yelp **review** data

Impact — Lead to quicker service improvements for restaurant owners based on prompt sentiment detection

Sentiment Analysis - Data Overview

01

Restaurant Review Data

~ 4.72 M, 44% are 5-stars

02

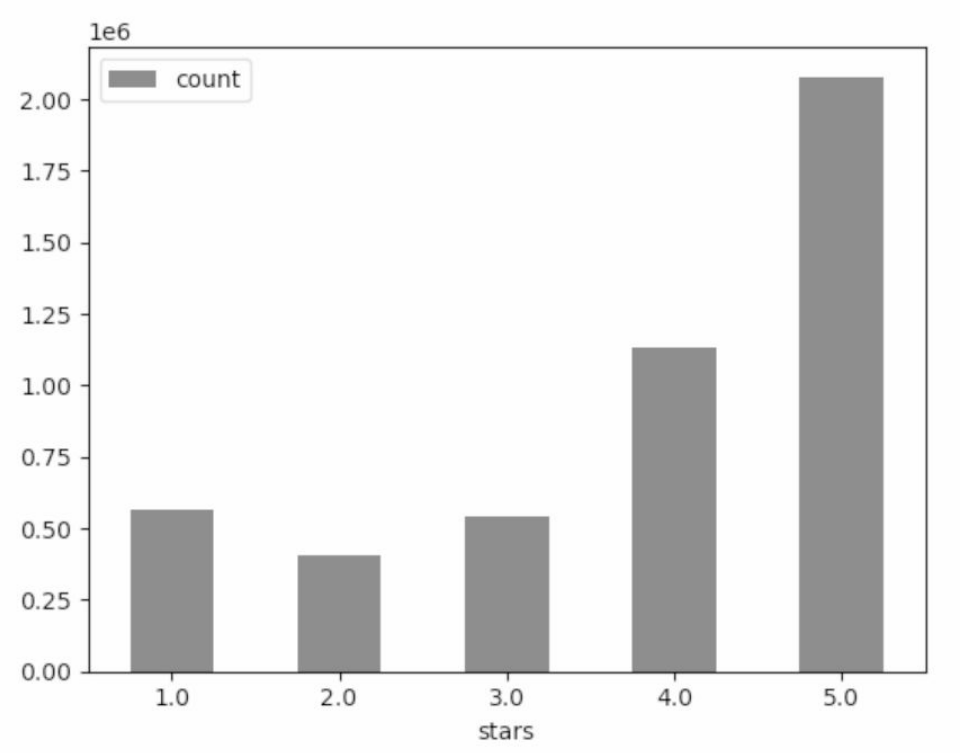
Two Assumptions

- Tendency to Express Strong Feelings
- Bias towards Strong Reactions

03

Relabelled Reviews

- Stars >= 4 → Positive, 1
- Stars <= 3 → Negative, 0



Distribution of Review Ratings

Sentiment Analysis - Data Overview

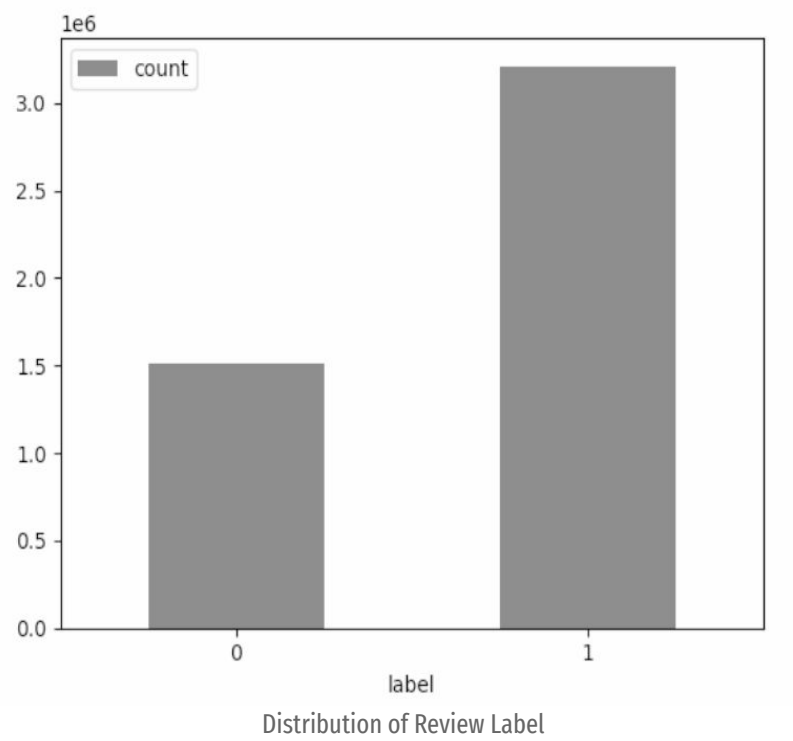
Positive Reviews

Key words: Great, Amazing, Delicious

Negative Reviews

Key words: Service, Time
(Indicative of dissatisfaction experiences)

Sentiment Analysis - Data Preprocessing



Distribution of Review Label

01

Deal with imbalanced dataset

- Label 0: 32%, Label 1: 68%
- Built a user-defined function to compute rebalancing weights for the imbalanced classes

Review	Label	Weight
Positive	1	0.74
Negative	0	1.56

Sentiment Analysis - Data Preprocessing

02

Only keep reviews written in English

Used John Snow Labs' Language Detection Annotators[1] to detect review

Language	Percentage
English	99.8%
Other (Spanish, French, German...)	0.02%

03

NLP Pipeline

- Converted raw text data into numerical features suitable for model training
- Included five stages: Tokenizer, Normalizer, Lemmatizer, Finisher, and HashingTF

1. Language Detection & Identification Pipeline - 21 Languages | detect_language_21 | Spark NLP 2.7.0

Sentiment Analysis - Classification Models

01 On sample data

LinearSVC, Random Forests, and Gradient-Boosted Trees (GBTs)

02 On large-scale data

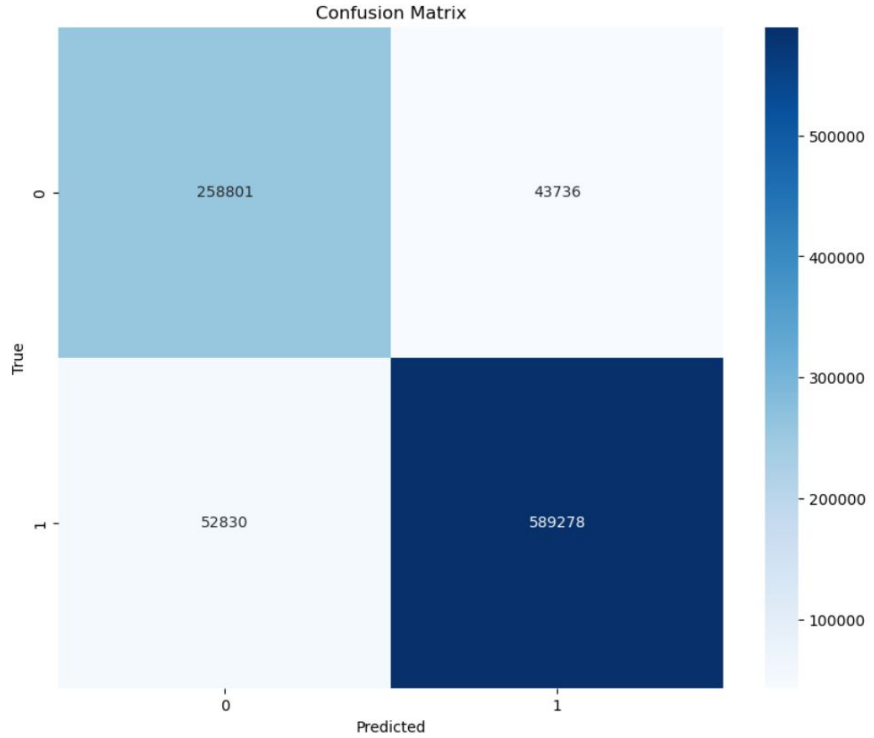
Accuracy: 0.8977, F1: 0.8981

03 Hyperparameter Tuning

- Accuracy: 0.8978, F1: 0.8982
- Best Params: regParam = 0.01, maxIter = 5, etc.

04 No Overfitting Issues

Training accuracy: 0.9034, Training F1 score: 0.9038



Confusion Matrix of the Best CV Model

Sentiment Analysis - Predictions on Tip

Restaurant Tip Data ~0.65 M

Prediction	Count
0	149229
1	471770

Tip Examples	Prediction
Starbucks substitute in boring downtown Tampa. Ugh. Never again!	0
Basically same food as rally's for $5 more	0
Dont go inside cause it stinks of stale greasy carpet. I guess just drive thru.	0
Best General Tso around. Also, try the grilled sweet pork, ask for extra sauce.	1
The pimento cheese tastes great and is a large portion for a starter - great for sharing! The fried pickles are also awesome!	1
Very good will definitely be coming back!!	1

Recommender System – Overview

Model

Collaborative Filtering
PySpark ALS (Alternating Least Squares)

Goal

Personalized recommendations based on user-item interactions

Pros

01 **Efficiency**: well-suited for large-scale datasets

02 **Matrix Factorization**: dimensionality reduction

03 **Cold Start**: handle new items or users with limited interaction.

Recommender System

Combined Data: Review + Business info for all restaurants. ~4.7 Million. Split into test/train – [0.2, 0.8]

On sample data

ALS model + Hyperparameter tuning (9 models). Overfitting occurs.
Current best model: RMSE = 0.47 (maxIter=5, regParam=0.09, rank=10)

2 algorithms on full data

a. **Recommender for users**:
recommend top 10 restaurants (name, categories, wordcloud of categories)
b. **Recommender for restaurants**:
recommend top 10 users (user ID, past visits, wordcloud of restaurant categories user have been to)

Evaluation

a. RMSE = 1.52 on full dataset
b. Compare results with current records. (if keywords in wordcloud matches)

Recommender System Results – User 1

Top 10 Recommendations:

Key words: Cafe, Burger, Sandwich, Mexican, America, Nightlife, etc.

Past Visits:

Key words: American (New, Traditional), Breakfast , Brunch, Burger, Nightlife, Fast, etc.

Recommender System Results – Restaurant 1

Name

The Pharmacy

Category

Burgers, Sandwiches, Food, Beer, Wine & Spirits, American (Traditional), Bars, Nightlife, Restaurants

Past visits of Top 10 recommended users:

Key words: Sandwich, American, Truck, Frozen, Yogurt, Ice Cream, Juice, Smoothie

Discussions

Applications

- **For yelp users:**
  personalized recommendations
- **For yelp restaurants:**
  Make marketing decisions based on top users' profile.
- **Other industries**: movie, music

Limitations

- User ratings may be biased.
- Hyperparameter tuning on sample dataset, not full dataset, because of overfitting on sample data.
- Did not consider location factor

Next Steps

1. Perform hyperparameter tuning for recommender systems using full dataset, avoid overfitting

2. Integrate NLP results with recommender system, adjust rating scores / assign weights for ratings

3. **Scheduling**: Train recommender system **quarterly** with most recent data

Thank You!
Q & A