Yelp Review Predictions

Enhancing User Experience Using Big Data Methodologies

MSCA 31013—Summer 2020

Julian Kleindiek, Jerry Sha, Vamika Venkatesan, Roy Xie

August 27th, 2020

Agenda

- Business Problem & Data
- Exploratory Data Analysis
- Data Infrastructure
- Modeling
- Next Steps

(AKA ELEVEN MADISON PARK)

SEPTEMBER 12TH, 2019

Pre-Dinner Cocktails and Appetizers

Guyere Gougeres
Prawn and Lemongrass Lollipops

Dinner Service

Amuse Bouche
Pancetta Cup with Herb Salad and Egg

First Course
Beets with Goat Cheese Mousse, Vinaigrette and Caraway Tuille

Second Course
Crab Salad with Celery-Apple Jelly and Grainy Mustard Vinaigrette

Third Course
Clam Bake, Chorizo Madeleines and Broth

Fourth Course
Ricotta Gnocchi with Butternut Squash Puree, Brown Butter, Sage and Parmesan Foam

Fifth Course
Lavender Glazed Duck Breast with Duck Confit Fennel and Peaches

Sixth Course
Beef Tenderloin with Braised Oxtail, Bone Marrow Crust and Sauce Bordelaise

Coffee

Dessert
Chocolate Palette with Peanuts and Popcorn Ice Cream

Migardise

Business Problem & Data Overview

How can we enhance the Yelp user experience using big data solutions?

POTENTIAL SOLUTIONS

Content-Based Recommendation Engine

Graph Theory-Based User Recommendation

Collaborative Filtering

Text-Based Star Suggestion

Our raw data is structured as multiple tables.

Reviews

```
root
 |-- review_id: string (nullable = true)
 |-- user_id: string (nullable = true)
 |-- business_id: string (nullable = true)
 |-- stars: double (nullable = true)
 |-- date: string (nullable = true)
 |-- text: string (nullable = true)
 |-- useful: integer (nullable = true)
 |-- funny: integer (nullable = true)
 |-- cool: integer (nullable = true)
```

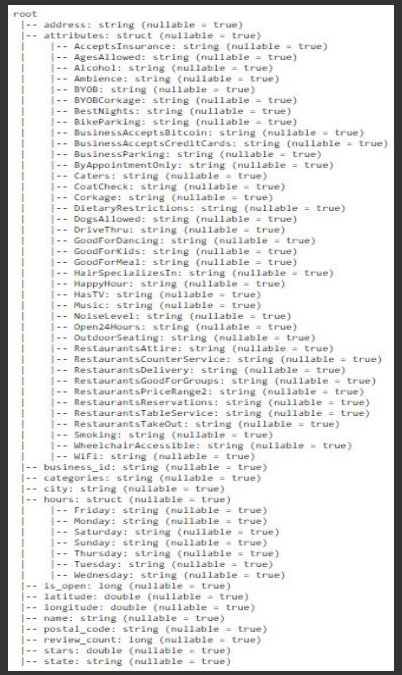
Size: 5.89 GB
Rows: 80,211,220

Tips

```
root
 |-- business_id: string (nullable = true)
 |-- compliment_count: string (nullable = true)
 |-- date: string (nullable = true)
 |-- text: string (nullable = true)
 |-- user_id: string (nullable = true)
```

Size: 251 MB
Rows: 1,363,162

Businesses



Size: 145 MB
Rows: 209,393

Users

```
root
 |-- user_id: string (nullable = true)
 |-- name: string (nullable = true)
 |-- review_count: integer (nullable = true)
 |-- yelping_since: string (nullable = true)
 |-- useful: integer (nullable = true)
 |-- funny: integer (nullable = true)
 |-- cool: integer (nullable = true)
 |-- elite: string (nullable = true)
 |-- friends: string (nullable = true)
 |-- fans: integer (nullable = true)
 |-- average_stars: float (nullable = true)
 |-- compliment_hot: integer (nullable = true)
 |-- compliment_more: integer (nullable = true)
 |-- compliment_profile: integer (nullable = true)
 |-- compliment_cute: integer (nullable = true)
 |-- compliment_list: integer (nullable = true)
 |-- compliment_note: integer (nullable = true)
 |-- compliment_plain: integer (nullable = true)
 |-- compliment_cool: integer (nullable = true)
 |-- compliment_funny: integer (nullable = true)
 |-- compliment_writer: integer (nullable = true)
 |-- compliment_photos: integer (nullable = true)
```

Size: 3.04 GB
Rows: 3,937,406

Exploratory Data Analysis

We explored our dataset for notable trends.

USERS

id	inDegree
None	834851
ZIOCmdFaMIF56FR-nWr_2A	5266
Oi1qbcz2m2SnwUeztGYcnQ	5059
8DEyKVyplnOcSKx39vatbg	4957
yLW8OrR8Ns4X1oXJmkKYgg	4475
hizGc5W1tBHPghM5YKCAtg	4460
djxnI8Ux8ZYQJhiOQkrRhA	4438
YttDgOC9AlM4HcAlDsbB2A	4346
qVc8ODYU5SZjKXVBgXdI7w	4186
iLjMdZi0Tm7DQxX1C1_2dg	4168

Degree for users (vertices) quantifying the number of connections/friends (edges)

BUSINESSES

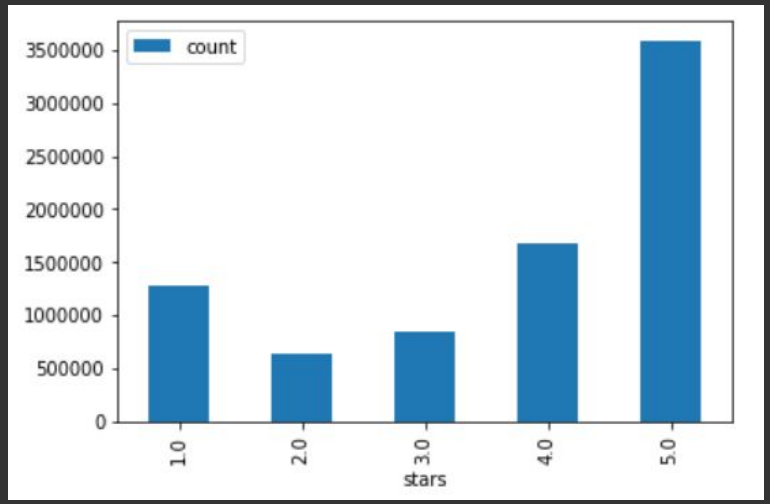
city	count
Las Vegas	31631
Toronto	20366
Phoenix	20171
Charlotte	10422
Scottsdale	9342
Calgary	8377
Pittsburgh	7630
Montréal	6979
Mesa	6577
Henderson	5272

primary_category	count
Restaurants	19456
Food	8539
Shopping	8516
Beauty & Spas	6304
Home Services	6282
Health & Medical	5249
Automotive	4657
Local Services	3825
Nightlife	2796
Active Life	2454

Top 10 cities reviewed and top 10 types of business reviewed

We explored our dataset for notable trends.

REVIEWS



Distribution of star ratings by review

TIPS

user_id	count
null	69231
mkbx55W8B8aPLgDqe...	2600
CxDOIDnH8gp9KXzpB...	1667
0tvCcnfJnSs55iB6m...	1589
6ZC-0LfOAGwaFc5XP...	1510
eZfHm0qI8A_HfvXSc...	1324
O8eDScRAg6ae019Bc...	1300
8DGFWco9VeBAxjqsu...	1179
2EuPAGalYnP7eSxPg...	1165
WJKocp9RE0KatUwh3...	1111

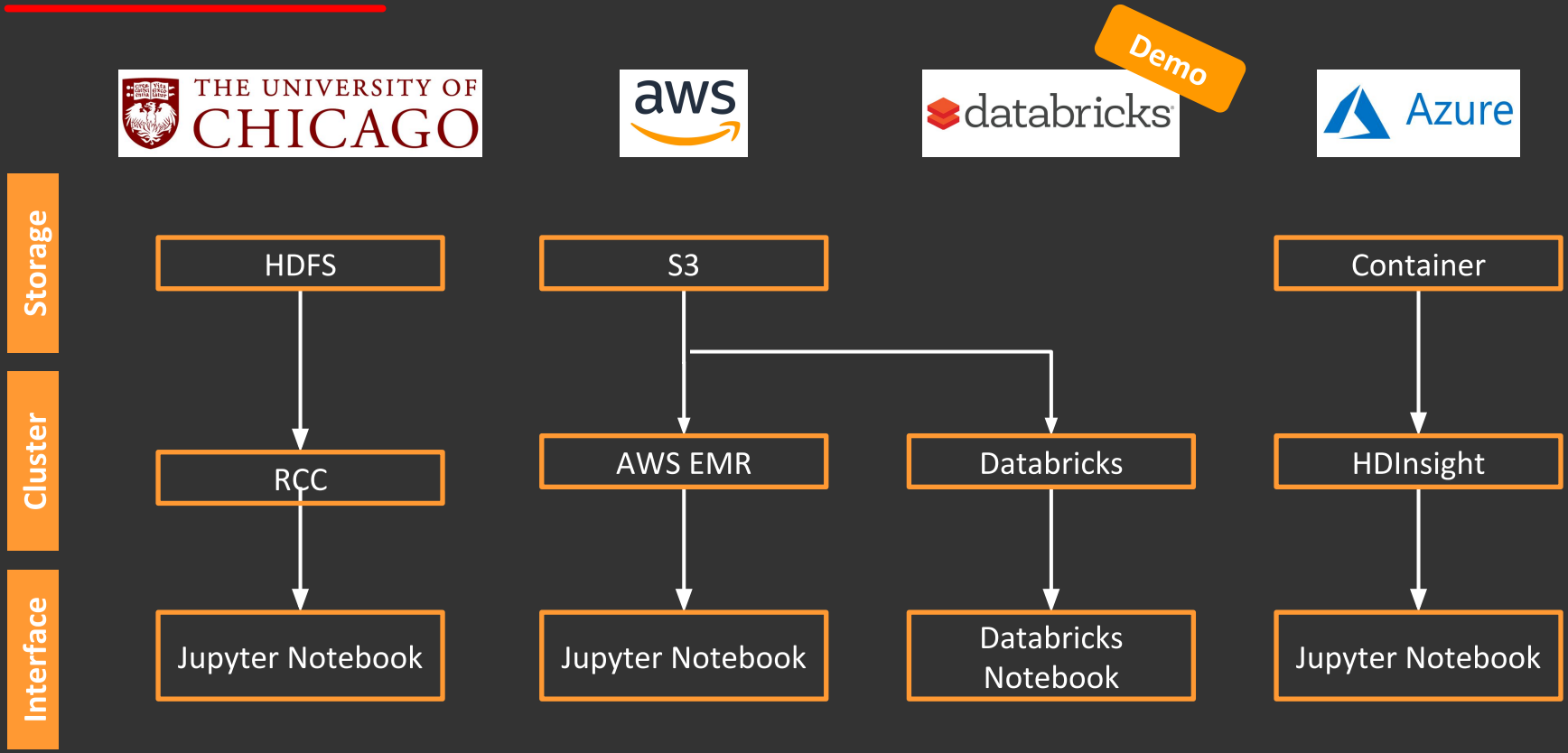
Top 10 users by number of tips given

business_id	count
FaHADZARwnY4yvlvp...	3679
JmI9nslLD7KZqRr__...	2494
DkYS3arLOhA8si5uU...	1530
5LNZ67Yw9RD6nf4_U...	1525
K7lWdNUhCbcnEvI0N...	1434
hihud--QRriCYZw1z...	1394
RESDUcs7fIiihp38-...	1386
4JNXUYY8wbaaDmk3B...	1185
yfxDa8RFOvJPQh0rN...	1154
iCQpiavjjPzJ5_3gP...	1145

Top 10 businesses by number of tips received

Data Infrastructure

We tested several leading big data frameworks…



… and found the following pros and cons to each.

	THE UNIVERSITY OF CHICAGO	aws	databricks	Azure
Pricing	• Free	• $100 credit via aws educate	• Free community edition (one 15GB cluster for a limited time)	• $100 student credit
Pros	• Free, powerful resources	• Clusters cannot be paused but cloned	• Convenient installation of packages • Notebook with additional functionality	• Great price transparency esp. before spinning up services
Cons	• Inflexibility with custom installations • Competing for resources	• Price is a surprise	• Cluster needs to be restarted incl. all installations after 2h(applies to community edition only)	• Clusters disappear once terminated

Modeling

We examined two different methods of recommendation.

	Simple Description	Pros	Cons
Collaborative Filtering	Utilize the “wisdom of crowd” to recommend businesses for each user Recommendations are based on the knowledge of users’ rating of businesses	No feature engineering effort	Cold start is needed for new users and new businesses
Content-based Filtering	Produces recommendation based on the similarity between businesses	Good for comparison between businesses	Review description is necessary

We built one model based on collaborative filtering.

BUSINESS USE CASE

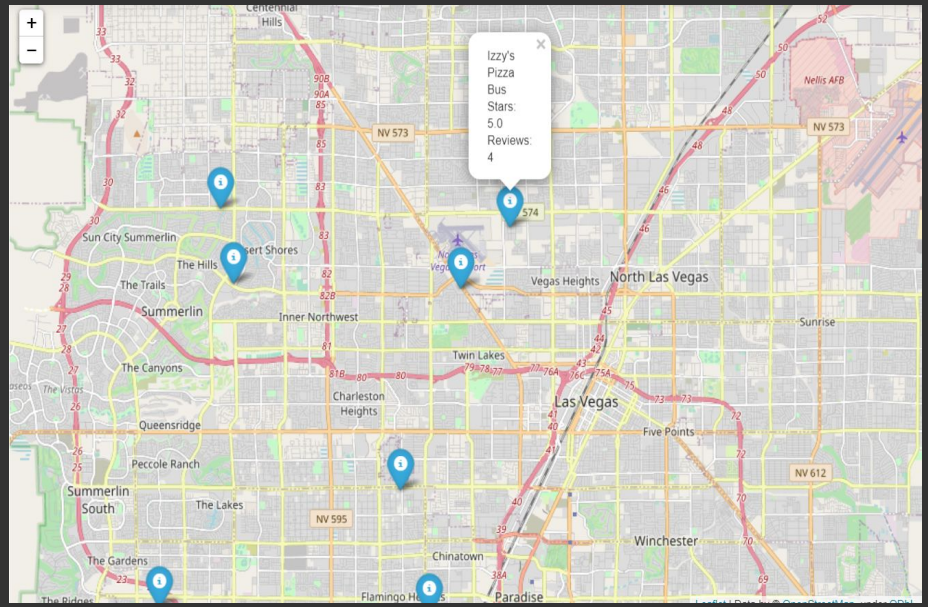
The goal is to provide 10 recommendations for each user, as well as to provide an interface for users to view specific information of businesses (business name, business rating, business location)

MODEL PIPELINE

- Create a subset of the review data for city Las Vegas
- Generate unique numeric ID for both the business data and the review data
- Using ALS from pyspark.ml.recommendation for modeling. UserID,BusinessID, and Ratings are used as inputs.
- Tuning parameters and comparing model results
- Use the best model to generate 10 recommendations for a user and show them in map

We built one model based on collaborative filtering.

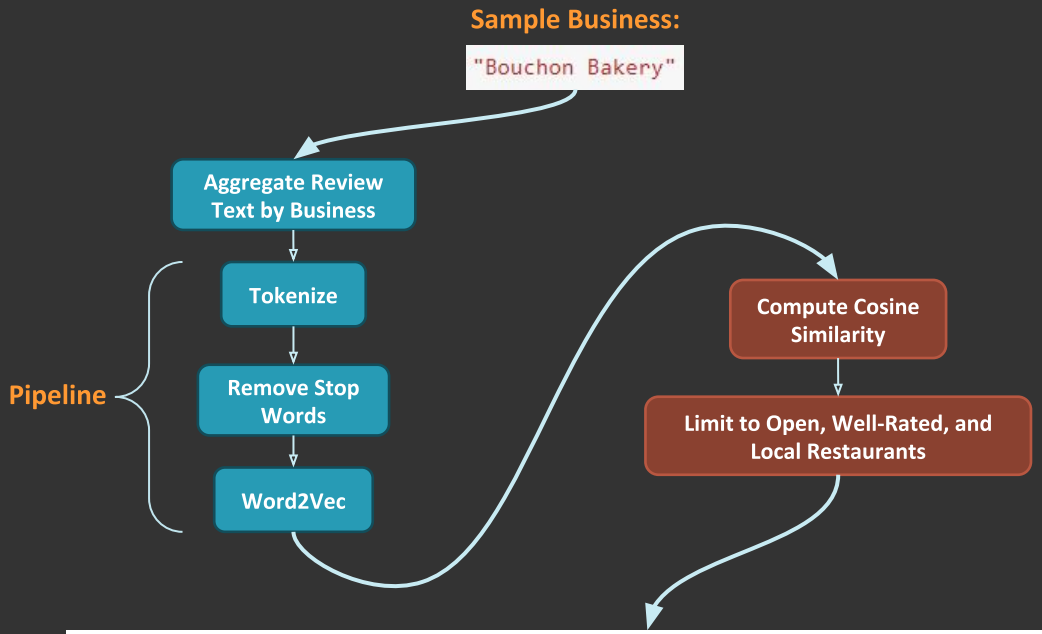
	Parameters	RMSE
Model I	rank = 10 maxIter = 10 regParam = 0.1	2.12
Model II	rank = 20 maxIter = 20 regParam = 0.3	1.85



We built another model using content based-filtering.

BUSINESS USE CASE

This use case builds a recommendation engine based on an aggregated view of all the reviews pertaining to a particular business. More specifically, it will recommend the top 5 most similar businesses to the selected business based on a ‘score’.



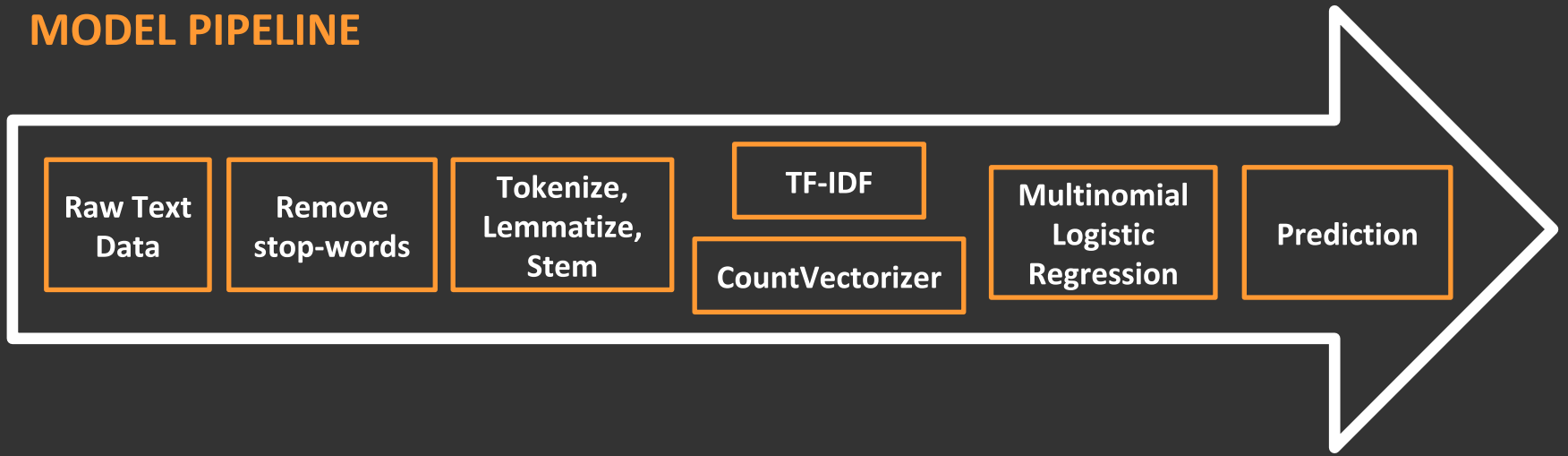
	name	address	city	cat_primary	score
0	Pinkbox Doughnuts	7531 W Lake Mead Blvd	Las Vegas	Food	0.907878
1	Drago Sisters Bakery	6870 S Rainbow Blvd, Ste 116	Las Vegas	Food	0.877925
2	District: Donuts. Sliders. Brew	3708 Las Vegas Blvd S, Fl 2, The Blvd Tower	Las Vegas	Food	0.874499
3	Nielsen's Frozen Custard	9480 S Eastern Ave, Ste 100	Las Vegas	Food	0.862000
4	Honolulu Cookie Company - Grand Canal Shoppes	3327 Las Vegas Blvd S	Las Vegas	Food	0.859122

We want to predict review score based on review text.

BUSINESS USE CASE

User review scores often don't reflect the content of the review due to psychological biases. Adding a "suggested score" based on NLP model predictions can help users give more accurate scores.

MODEL PIPELINE



Results from the score prediction model demonstrate predictive power.

	RMSE	RUNTIME
TF-IDF	1.3147	1.04 minutes
CountVectorizer	1.0982	1.42 minutes

CONCLUSION

- While current model features and framework do not achieve outstanding performance, the pipeline is able to successfully process 5GB+ of text data
- Review score prediction using textual big data is feasible

Next Steps

We can continue to explore alternative options at different steps of the project pipeline.

ADDITIONAL STORAGE OPTIONS

Cloudera, Google Cloud Platform

SUPPLEMENT REVIEW DATA FROM ALTERNATIVE SOURCES

Google Reviews, Facebook Reviews, digital news publications

EXPLORE OTHER BUSINESS USE CASES

Recommend users to follow based on graph network


Thank you!