Yelp Data Analysis

Big Data Platform: Final Project
Group 3

Rolamjaya Hotmartua
Yichin Tzou
Zoey Chen

Executive Summary

- Yelp is one of the biggest online platforms that allows users to search for and review businesses such as restaurants, bars, hotels, and other local services in the US.
- LDA provides restaurants with better knowledge of the content of the reviews thus make improvements in the future
- Graph analysis identified 10 potential Influencers in Restaurant category.
- ALS model will provide a recommender system for restaurants and users in IL

Agenda

Business Problem

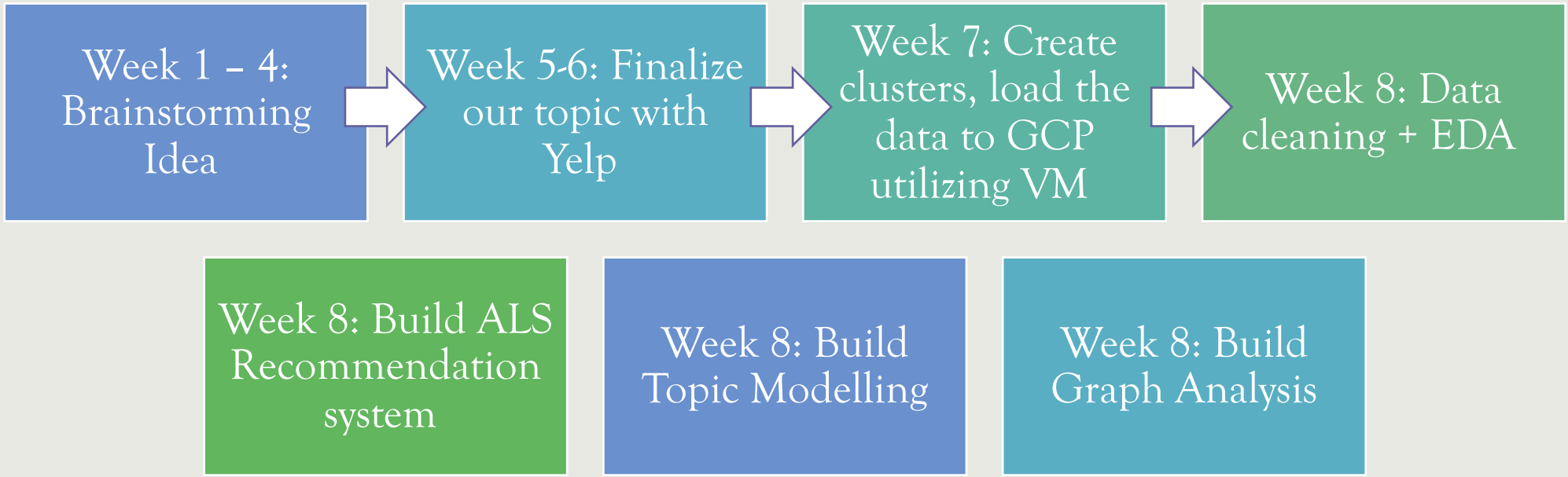
Data Preparation and Exploration

Model Building and Selection

Results and Findings

Conclusion and Improvement

Project execution timeline/ schedule



Business Problem

What/ How can restaurants gain insight from reviews for future improvement?

How can Yelp suggest users that can be a Key Opinion Leader for improving the restaurants brand image?

How can Yelp recommend a restaurant to the users and users to the restaurants for targeted promotion?

Data Profile

- Data Sources: Yelp website (Public Data)
- Data Format: JSON
- Data Size: 8 GB

We store our dataset in GCP, these data include:

- Business – information with business attributes
- Check-in - users check-in date
- Reviews – users text reviews, ratings
- Tips – users tips
- Users – users' information, also contain "friends"

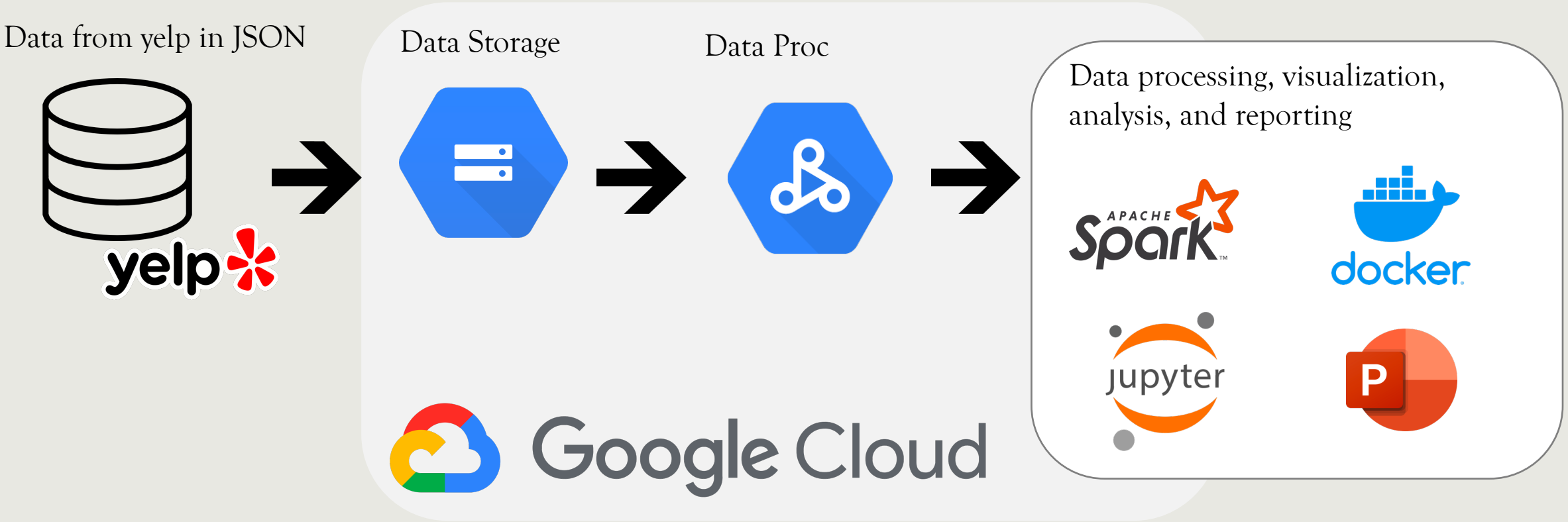
GOOGLE CLOUD PLATFORM

DOCKER

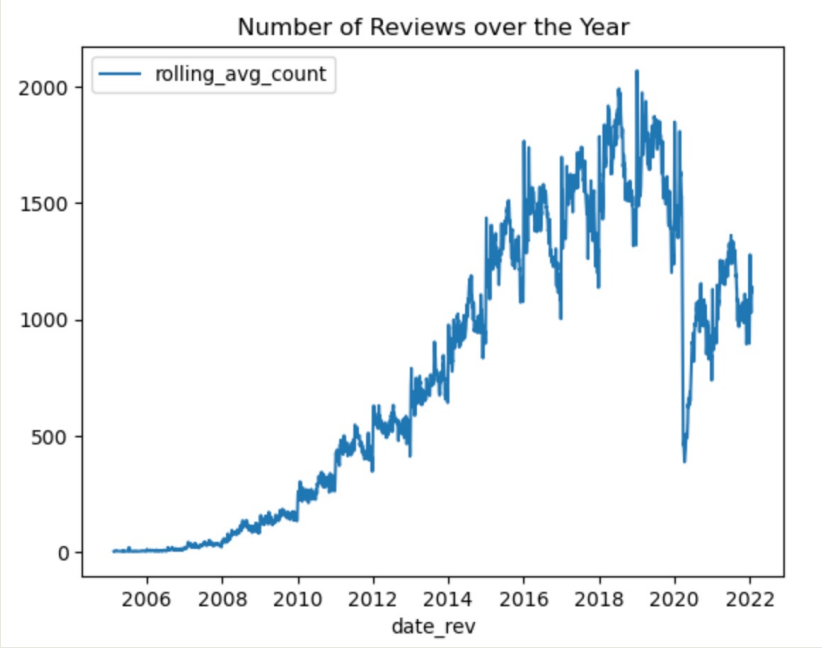
1. ALS Recommender System
2. Topic modeling
3. Graph Analysis

Name	Size	Type
Dataset_User_Agreement.pdf	78.5 KB	application/pdf
test/	–	Folder
wget-log	417 B	application/octet-stream
yelp_academic_dataset_business....	113.4 MB	application/json
yelp_academic_dataset_checkin.j...	273.7 MB	application/json
yelp_academic_dataset_review.js...	5 GB	application/json
yelp_academic_dataset_tip.json	172.2 MB	application/json
yelp_academic_dataset_user.csv/	–	Folder
yelp_academic_dataset_user.json	3.1 GB	application/json

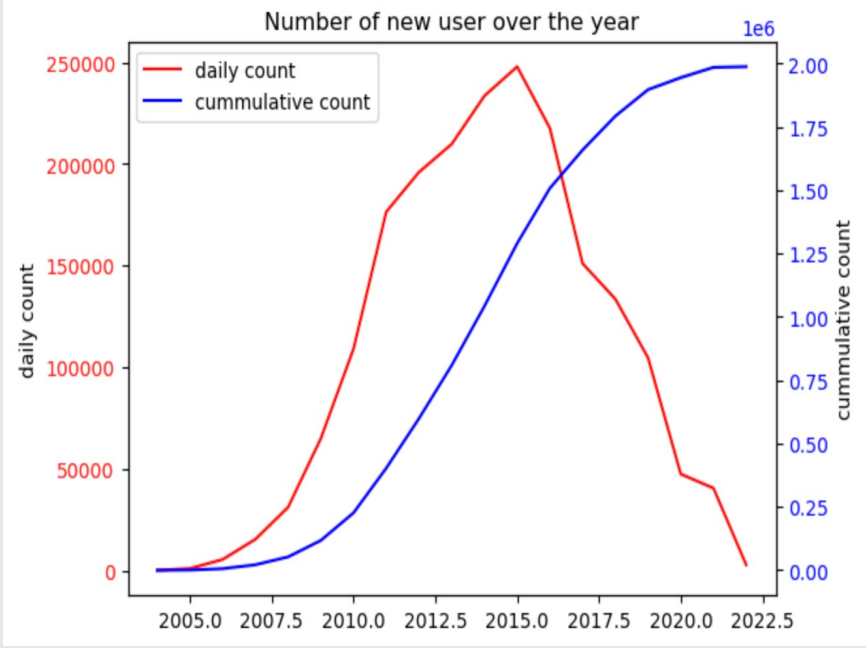
Data Preparation& Pipeline



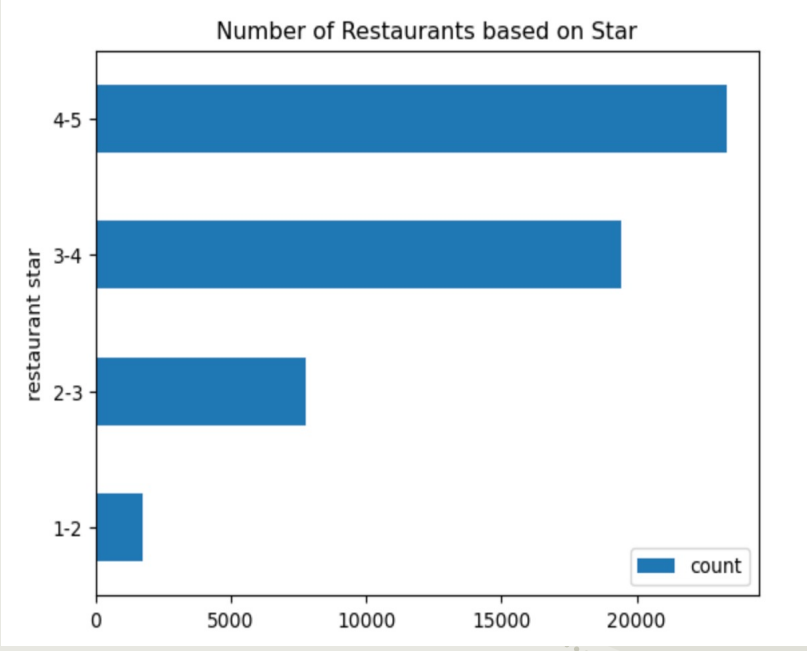
Data Exploration



- The number of review is increasing over the year except when COVID strikes.
- There are a periodical trend in each year where the number of review is increasing in the beginning of the year and then decreasing.

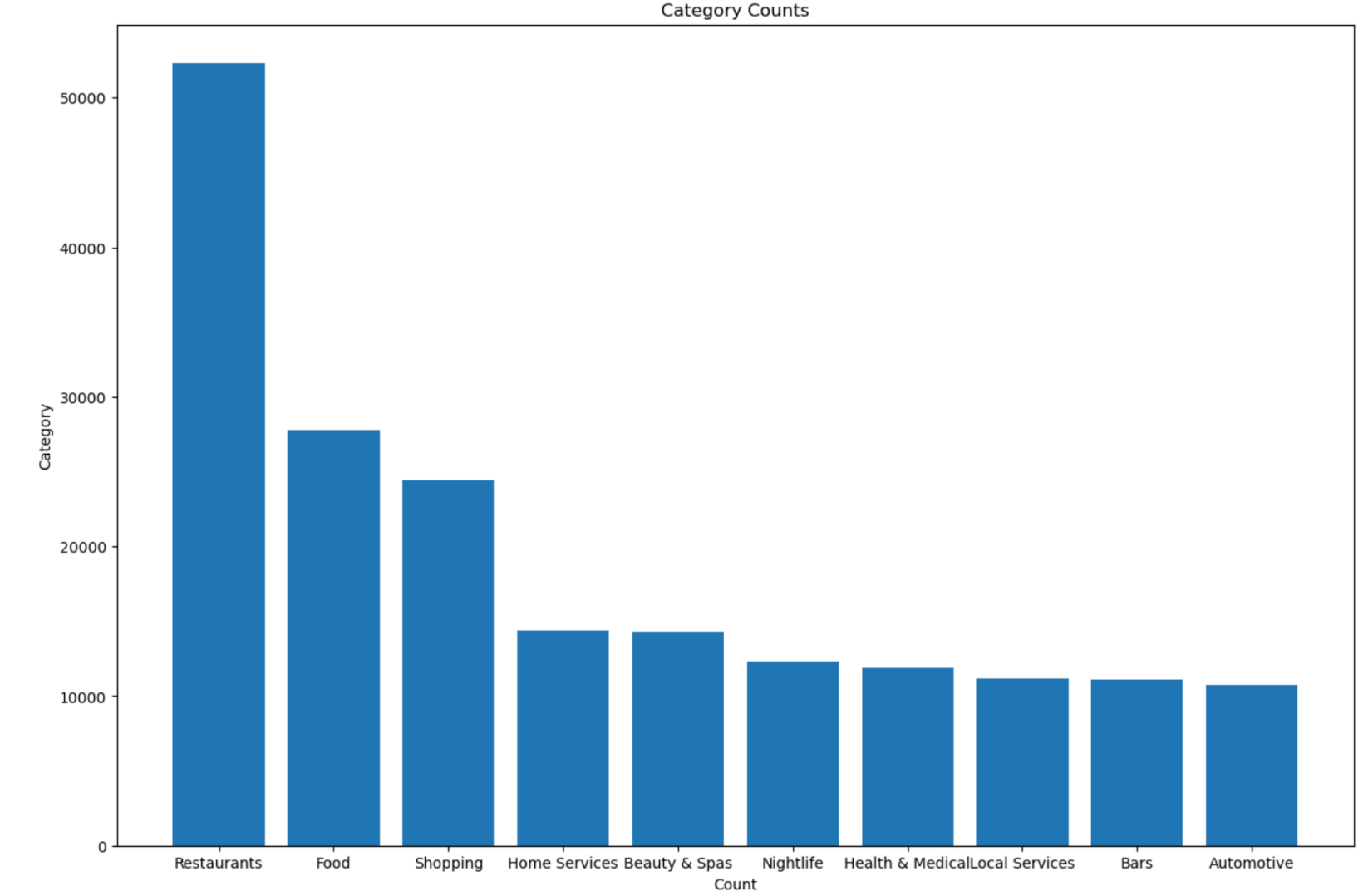


- Number of users registered in Yelp has reached its peak around 2015 and decreasing since then, giving the sign of saturation.
- This is a warning for Yelp.

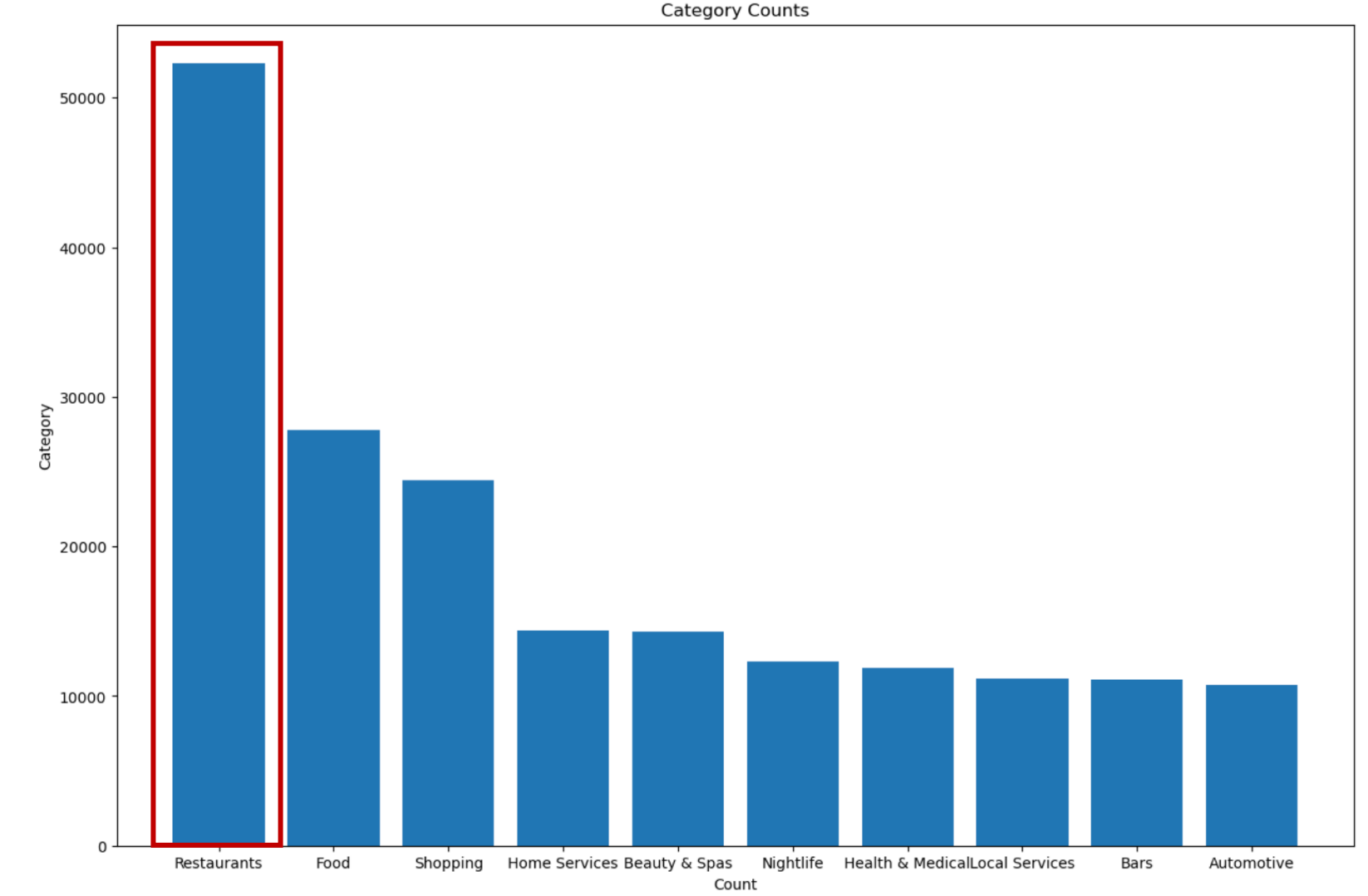


- Imbalance number of restaurants in each category

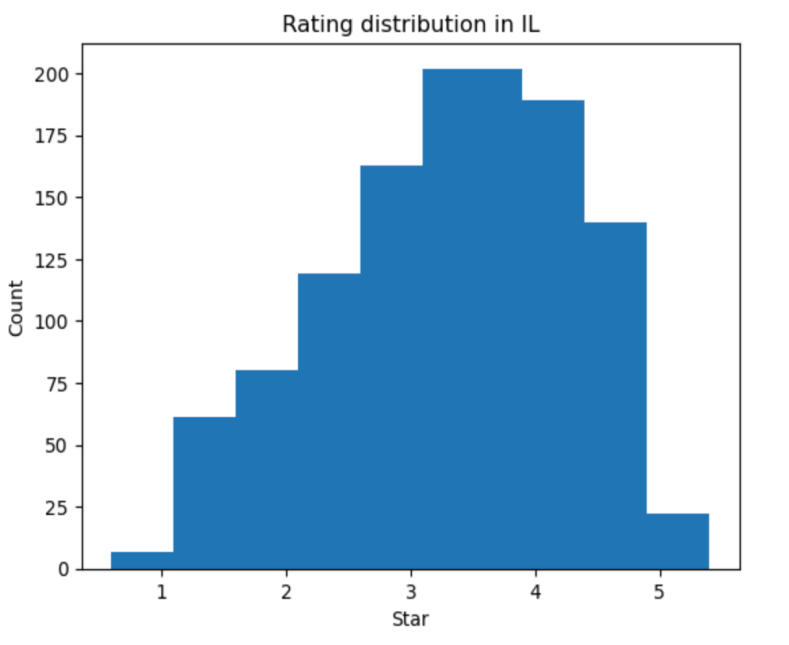
What business are we focusing on?



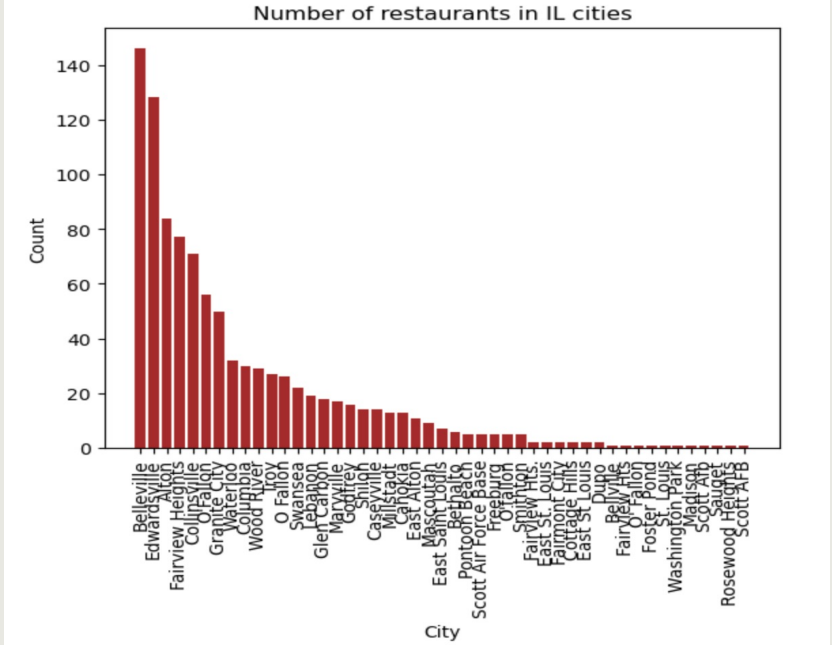
What business are we focusing on?



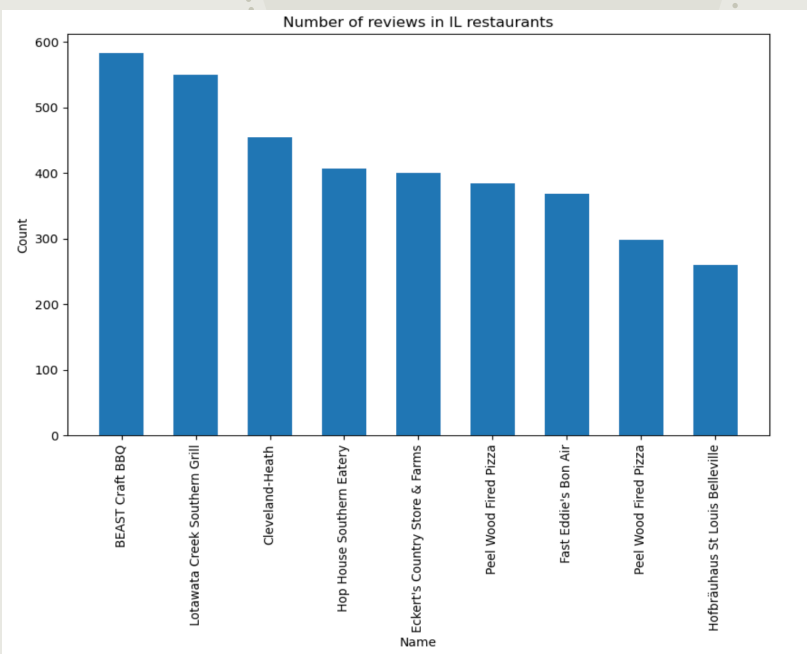
Data Exploration in restaurants in IL



- The rating distribution has the most in 3, 3.5 stars, this also gives us an insight when splitting data into positive and negative



- We can see that Bellevillie has the most restaurants in IL
- Followed by Edward Villie, and Alton



- BEAST Craft BBQ has the top reviews in IL
- Followed by Lotawata Creek Southern Grill, and Cleveland-Heath

Model Building and Selection – Topic Modelling

Data joining, filtering	Building a pipeline for natural language processing(NLP)	Use stars rating of the review to split the data	Use *Latent Dirichlet Allocation*(LDA) for topic modeling
• Business • Reviews	• Remove punctuations, symbols • Remove stopwords • Tokenized • TF-IDF	• 1:Positive(rating>=4) • 0:Negative(Rating<4)	• Gain the top frequency words for positive/negative rating • Num_topics: 10

Result and Finding

topic	topicWords
0	[pretty, going, , give, get, like, much]
1	[cheese, burger, sandwich, fries, , sauce, ordered]
2	[fresh, come, order, made, want, , good]
3	[every, favorite, eat, day, , great, menu]
4	[, little, excellent, ordered, definitely, great, really]
5	[server, table, , friendly, really, staff, lunch]
6	[, us, chicken, meal, dinner, said, know]
7	[, shrimp, restaurant, minutes, recommend, order, day]
8	[pizza, sauce, , want, hot, got, get]
9	[bar, love, chicken, always, fried, great, area]

Topic for all reviews

topic	topicWords
0	[pretty, going, get, around, wait, much, go]
1	[cheese, burger, fries, sandwich, , sauce, delicious]
2	[fresh, come, want, delicious, made, order,]
3	[every, favorite, eat, day, menu, great, delicious]
4	[, little, excellent, ordered, definitely, great, night]
5	[friendly, staff, , great, really, lunch, server]
6	[, chicken, meal, dinner, us, fried, lunch]
7	[, shrimp, recommend, restaurant, best, definitely, excellent]
8	[pizza, sauce, hot, , bit, best, want]
9	[bar, love, always, chicken, fried, great, place]

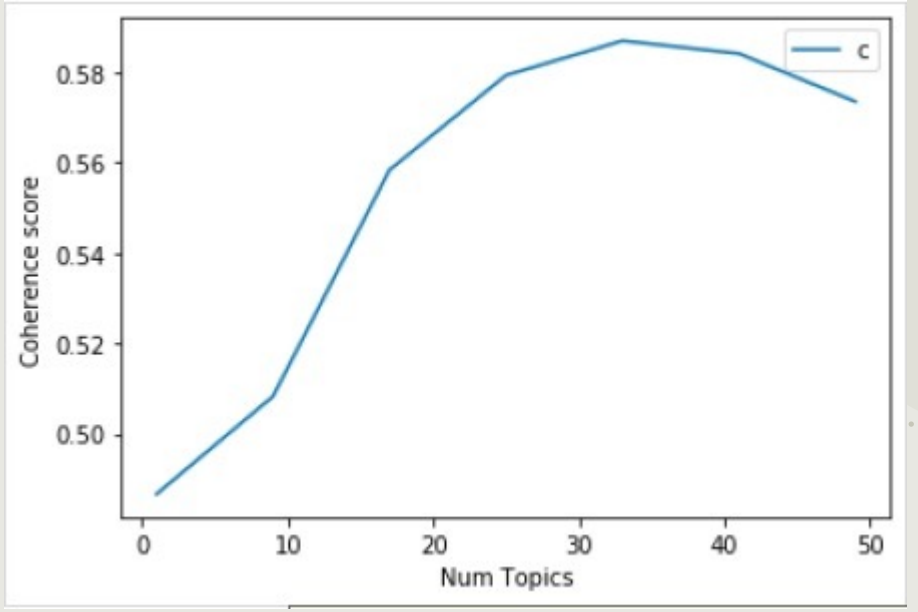
Topic for all positive reviews

topic	topicWords
0	[give, pretty, going, salad, went, like,]
1	[cheese, burger, sandwich, fries, , ordered, sauce]
2	[order, , want, come, wait, take, staff]
3	[even, eat, times, , people, every, time]
4	[, ordered, little, times, never, pizza, experience]
5	[table, server, , us, really, like, food]
6	[, us, said, asked, never, know, came]
7	[minutes, , order, asked, said, took, us]
8	[pizza, sauce, , ever, want, got, get]
9	[chicken, bar, fried, shrimp, good, nice, better]

Topic for all negative reviews

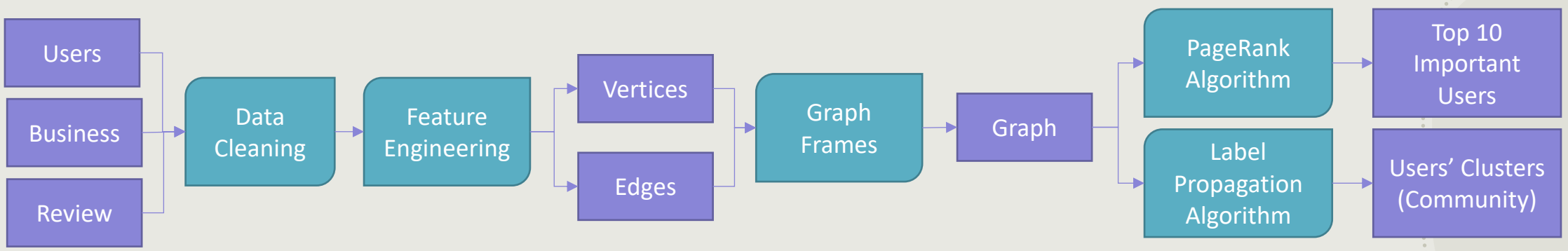
Improvement

1. Install the gensim and pyLDAvis in the future:
    - Find out the coherence score for the model, and find out the best/optimized numbers for topics for future modeling.
    - Plot out the visualization of our LDA model using pyLDAvis and check the term frequency that is used in each topic
    - Apply more variables to increase the complexity of the model
2. We could do sentiment analysis for reviews to divide the positive/ negative reviews



Model Building and Selection

– User Network Analysis



- Limit Users to Top 2000 reviewers
- Create Features: Responses
- Create Vertices and Edges

Total Vertices: 341621
Total Edges: 711209

Original Data Size: 8.8 GB
Sample 1: 18 MB
Sample 2: 531.3 MB (JSON), 75.1 MB (Parquet)

Result and Finding

Top 10 Most Important Users [PageRank]

name	since_year	average_stars	response	fans	count_friends	pagerank
Michelle	2008	4.05	55120	1353	5958	19.833261744321955
Steven	2010	3.62	66519	739	10072	17.922276178908504
Morris	2012	4.39	15651	713	3572	10.267693130330548
Dee	2010	4.13	9444	176	1448	9.381879740046559
Bruce	2008	3.77	6307	110	1546	9.227369028077154
Kimberly	2013	3.52	1385	41	1063	8.748330850125807
Chad	2015	3.7	34449	315	4732	8.302947040298289
Gabriella	2016	4.33	3895	167	2048	8.032567652224675
Karen	2008	3.69	31912	558	3708	8.03256366052581
Alisha	2016	3.97	11278	462	2617	7.641622809176325

only showing top 10 rows

Top 10 Most Important Users [DegreeCentrality]

name	since_year	average_stars	response	fans	count_friends	degree_centrality
Steven	2010	3.62	66519	739	10072	10722
Abby	2008	4.15	68127	1806	8858	8965
Niki	2014	4.59	22375	1746	6896	7075
Brittany	2011	4.33	38591	2086	6660	6854
Michelle	2008	4.05	55120	1353	5958	6627
Michael	2010	4.43	49580	1251	6481	6595
Joi	2008	4.1	61571	944	6234	6386
John	2008	3.75	69937	1309	6232	6349
Christy	2011	4.19	96429	740	5522	5716
Nate	2011	4.2	71063	1295	5607	5675

only showing top 10 rows

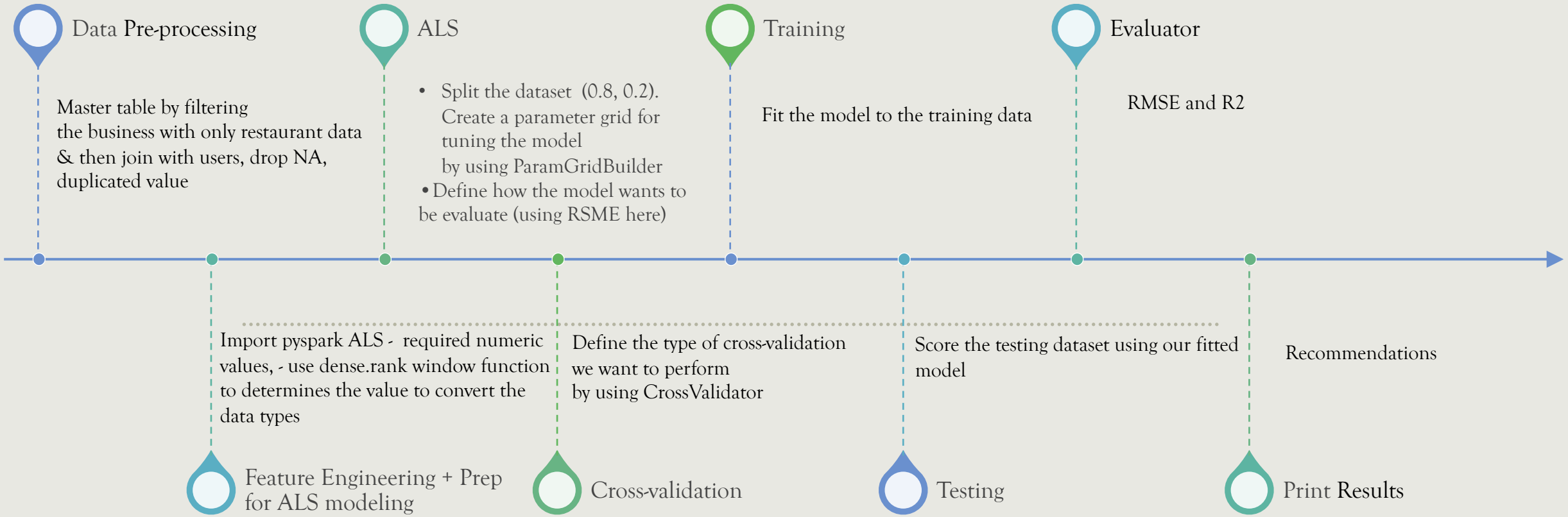
Clusters of Users [LPA]

label	count
970662608897	2
721554505733	86
1030792151045	708
936302870539	5
1022202216450	46
953482739716	182
472446402566	133
609885356039	7
326417514504	27
1005022347270	20
704374636547	10
944892805126	328
1047972020234	111
635655159814	22
558345748482	2
1039382085644	197
1709396983817	48
730144440327	1
627065225216	1
601295421445	2

Notes:

- PageRank and DegreeCentrality agrees on 2/10 important users.
- Top 10 important users – Reference to choose Influencers – Combine with Topic Clustering of their Review
- This can be used by Yelp to add more value for their business users
  1. Giving targeted suggestion of KOL to cooperate with
  2. Increase chance to get new customers from recommendation system.
- Clusters of LPA can be a reference to see community that want to be reached and find the KOL of that community based on the PageRank Algorithm.

Model Building and Selection – ALS Recommender System



Result and Finding

```
Root-mean-square error: 1.6528341288796768
r2 : 0.8181540899458146
```

- Future improvements:
  1. We can apply NLP, such as using the text reviews as another variable to analyze
  2. Since ALS model only get the input of the rating (stars), we can potentially add in more predictors in the future for more exploration as well
  3. Run on nationwide data instead of IL

user_id	recommendations
12	[{207, 2.5357628}...
22	[{33, 4.382561}, ...
26	[{207, 2.5900137}...
27	[{207, 2.814681},...
31	[{816, 4.735834},...

Top 5 business recommendations for user

business_id	recommendations
12	[{1963, 4.8463387...
22	[{14476, 3.115163...
26	[{2764, 4.3978014...
27	[{13190, 2.541835...
28	[{9620, 4.478644}...

Top 5 user recommendations for business

Additional Learning – Interactive map

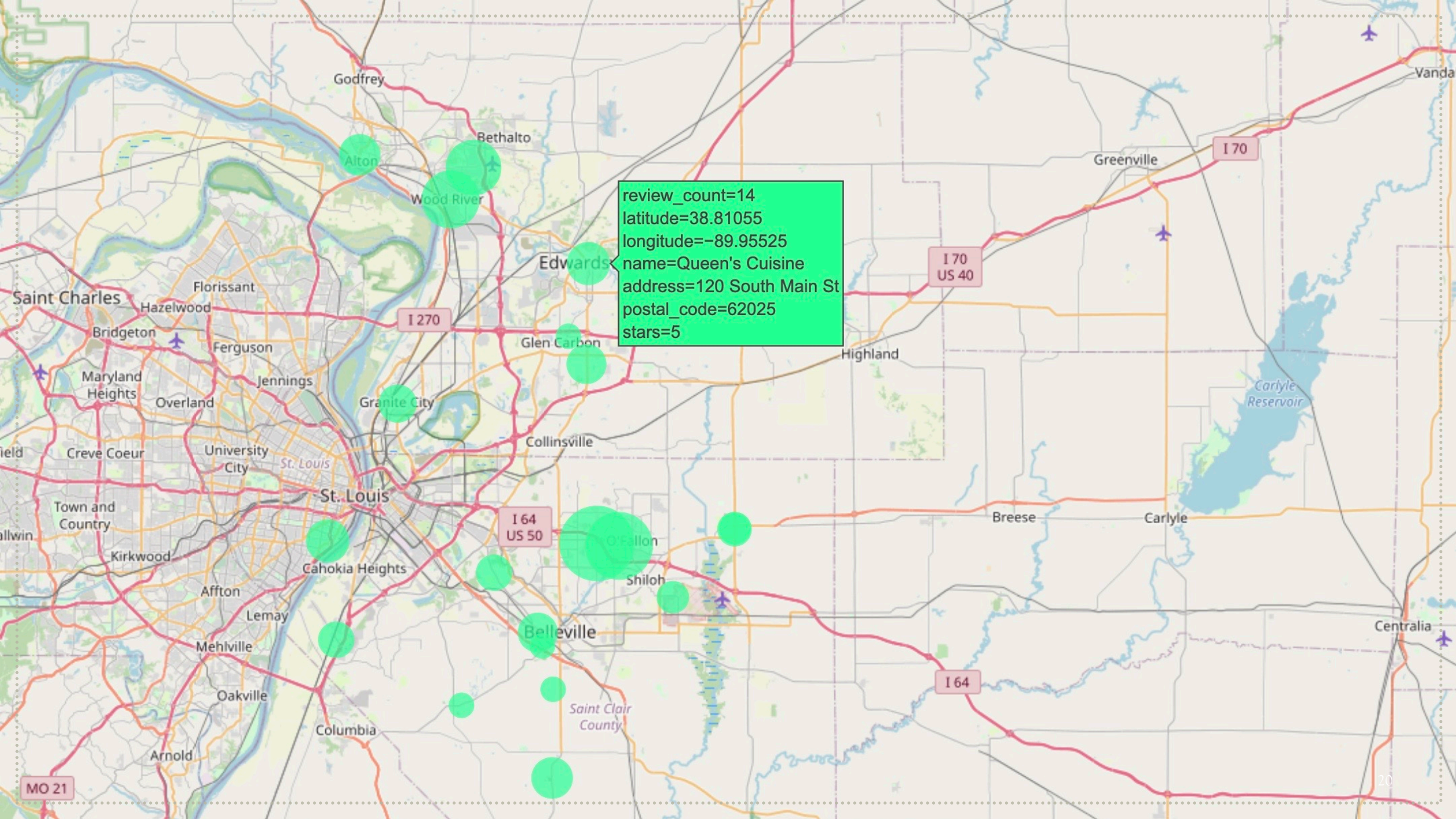
state: IL
stars: 5.0
categories: Restaurants
Run Interact

'Top 10 recommendations are as below'

[Stage 150:> (0 + 2) / 2]

address	categories	city	hours	is_open	name	postal_code	review_count	stars	state
1334 Central Park...	Korean, Restauran...	O'Fallon	{11:0-22:0, 11:0-...	1	bb.q Chicken - O'...	62269	42	5.0	IL
711 W US 50	Pizza, Italian, R...	O'Fallon	{11:0-20:0, 11:0-...	0	Alex's Pizza	62269	34	5.0	IL
144 E Ferguson Ave	Caribbean, Cuban,...	Wood River	null	0	Big Johns Cuban Cafe	62095	25	5.0	IL
611 E Airline Dr	Pizza, Sandwiches...	Rosewood Heights	{7:0-15:0, 0:0-0:...	0	First Stop Bake Shop	62024	23	5.0	IL
120 South Main St	British, Tea Room...	Edwardsville	{11:0-14:0, 0:0-0...	1	Queen's Cuisine	62025	14	5.0	IL
201 W Mill St	Restaurants, Bake...	Waterloo	null	1	Ahne's Bakery	62298	14	5.0	IL
1820 Cherokee St	Historical Tours,...	St. Louis	{null, null, null...	1	St. Louis Paranor...	63118	14	5.0	IL
101 W 9th St	Sandwiches, Resta...	Alton	{11:0-21:30, null...	1	Shake Rattle & Ro...	62002	13	5.0	IL
408 South Main St	Food, Burgers, Re...	Smithton	{11:0-21:0, 11:0-...	1	Walton's Ice Crea...	62285	13	5.0	IL
1926 West Main St	Southern, America...	Belleville	null	1	C And C Food For ...	62226	12	5.0	IL

only showing top 10 rows


review_count=14
latitude=38.81055
longitude=−89.95525
name=Queen's Cuisine
address=120 South Main St
postal_code=62025
stars=5
Godfrey
Alton
Bethalto
Wood River
Edwards
Saint Charles
Florissant
Hazelwood
Bridgeton
Ferguson
Maryland Heights
Jennings
Overland
Granite City
Glen Carbon
Collinsville
Creve Coeur
University City
St. Louis
Town and Country
Kirkwood
Cahokia Heights
Affton
Lemay
Mehlville
Oakville
Columbia
Arnold
MO 21
I 270
I 64
US 50
O'Fallon
Shiloh
Belleville
Saint Clair County
Greenville
I 70
US 40
Highland
Carlyle Reservoir
Breese
Carlyle
Centralia

Conclusion & Learnings

Conclusions

- By using topic modelling, we are able to extract insight that triggers User's review, such as: taste, type of food, services (staff friendliness)
- By graph analysis, we identified top 10 KOL and the respective communities to reach out
- By recommender system, yelp can recommend personalized restaurant to the current user and restaurant can get recommendation of user for their more targeted promotion.

Improvements

- Increase the dataset since we're only currently running the data for restaurants in IL
- Encounter many problems during NLP, can pre-installing some libraries in the University of Chicago Cloud account earlier in the future (ex: NLTK, TextBlob, WordCloud...)
- Adding more machine learning methods, NLP to better improve our results
- Set the permissions for our own cluster so no one can delete it by accident.
- Performing Graph Analysis at Cloud so it can process Millions of edges.

Thank You

Rolamjaya Hotmartua, Yichin Tzou, Zoey Chen

University of Chicago

MSc in Analytics

2023