Yelp-Sentiment-Analysis

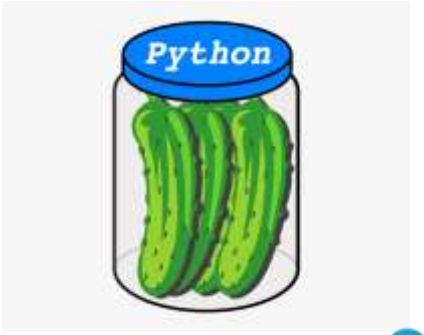
By Abdulrahman & Turki

Outline:

- Project Scope
- Data
- EDA
- NLP
- Topic Modeling Clustering
- Sentiment Review Classification
- Recommendation System

Project Scope & Tools:

- Our goal is to build unsupervised Natural Language Processing (NLP) machine learning models to predict whether a business review text is positive or negative. Also, assigns topics(clustering) based on the raw text data to find out the business domains and implementing a recommendation system



Data

Yelp is one of the most famous business review app in the Western Hemisphere countries, with more than 52 million visitors to its mobile sites as of December 2020

Two Datasets imported from Yelp website(review & business)

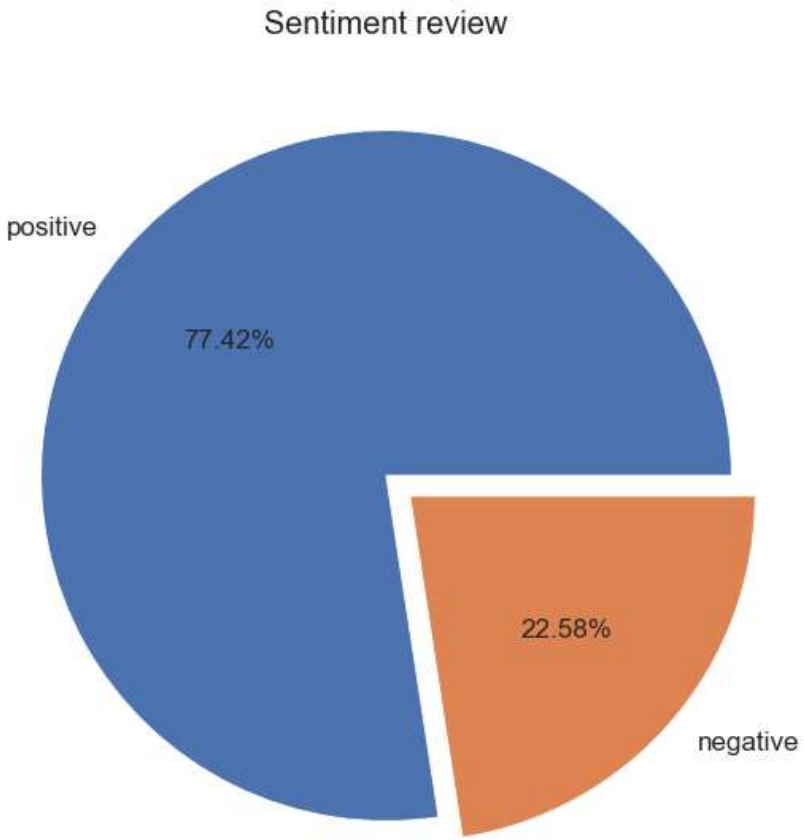
Containing 500k rows and 14 columns

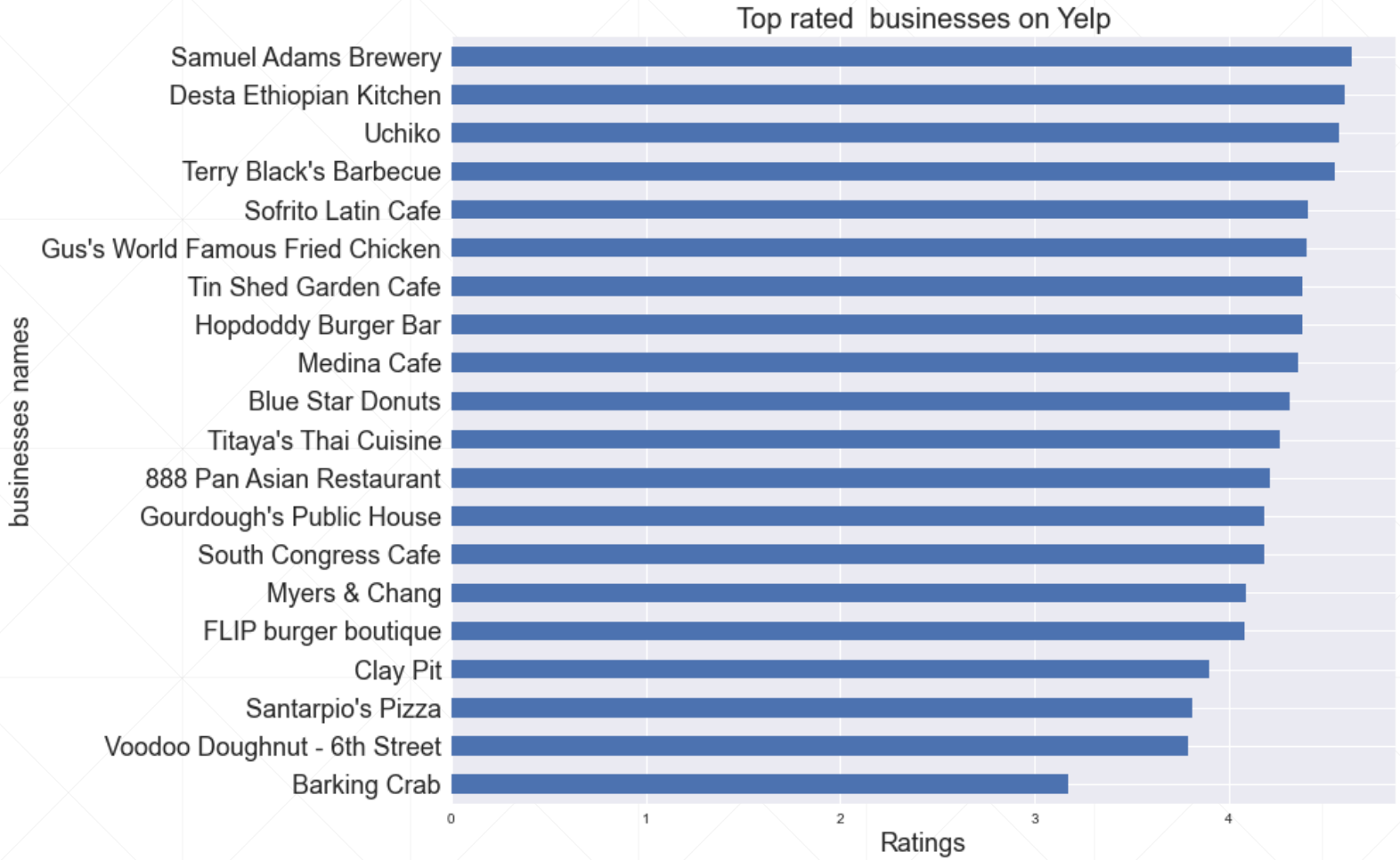
Sample with 10k rows

EDA

- Removing duplicates
- Change or remove null values
- Removing unwanted columns
- Stripping values
- Changing stars column to positive and negative

EDA


Sentiment review
positive
77.42%
22.58%
negative


Top rated businesses on Yelp
businesses names
Samuel Adams Brewery
Desta Ethiopian Kitchen
Uchiko
Terry Black's Barbecue
Sofrito Latin Cafe
Gus's World Famous Fried Chicken
Tin Shed Garden Cafe
Hopdoddy Burger Bar
Medina Cafe
Blue Star Donuts
Titaya's Thai Cuisine
888 Pan Asian Restaurant
Gourdough's Public House
South Congress Cafe
Myers & Chang
FLIP burger boutique
Clay Pit
Santarpio's Pizza
Voodoo Doughnut - 6th Street
Barking Crab
0
1
2
3
4
Ratings

Word Cloud

think
ca nt
go back
first time
customer service
highly recommend
next time
come back
great service
great food
really good
good food
wo nt
nt know
happy hour
well
great place
nt even
pretty good
ice cream
service great
look
still
coming back
thought
tried
staff friendly
love place
awesome
though
said
probably
maybe
point
actually
either
make sure
found
kind
instead
used
nt wait
location
perfect
food good
food great
especially
even though
fried chicken
every time
way
love
problem
last night
one best
course
meal
fantastic
got
dessert
usually
lot
plus
mean
put
taste
use
fact
issue
ask
asked
husband
offer
today
one favorite
wonderful
etc
check
food delicious
recommend place
everything
done
kid
top
left
guess
tell
store
big
see
hand
dish
flavor
say
guy
work
thank
show
seem
nt want
something
wife
loved
one
plate
going back
although
house
started
even
felt
yet
nt go
ye
food service
great experience

NLP Pre-Processing

Removing punctuation, digits, different languages, custom stop-words, special characters and spelling errors.

Converting to lower case.

Lemmatization.

Vectorization.

Topic Modeling

LSA

Topics(2-10)

CV & TF-IDF

NMF

Topics(2-10)

CV & TF-IDF

Corex

Topics(2-10)

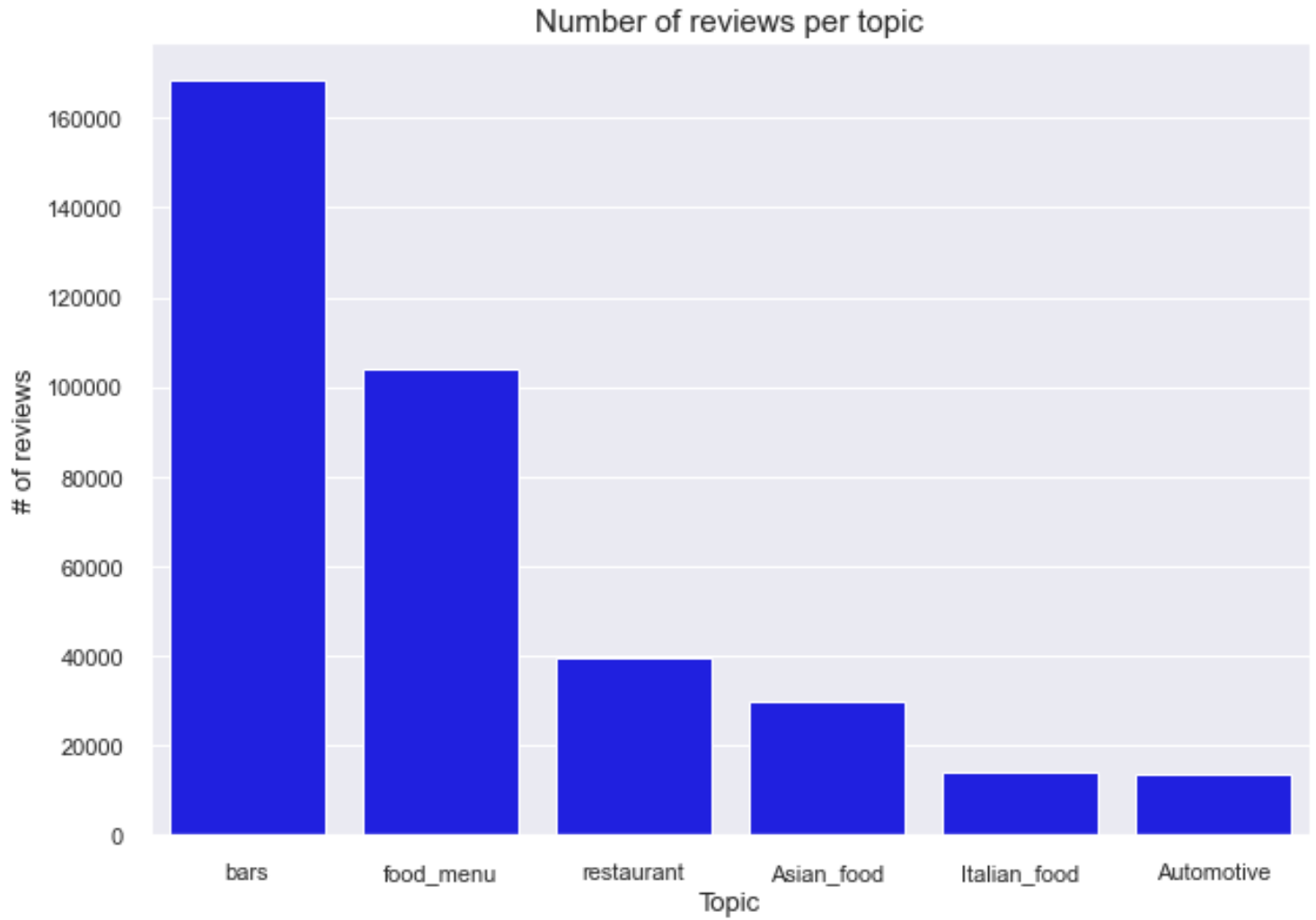
CV & TF-IDF

Best model was Count Vectorizer NMF with six topics

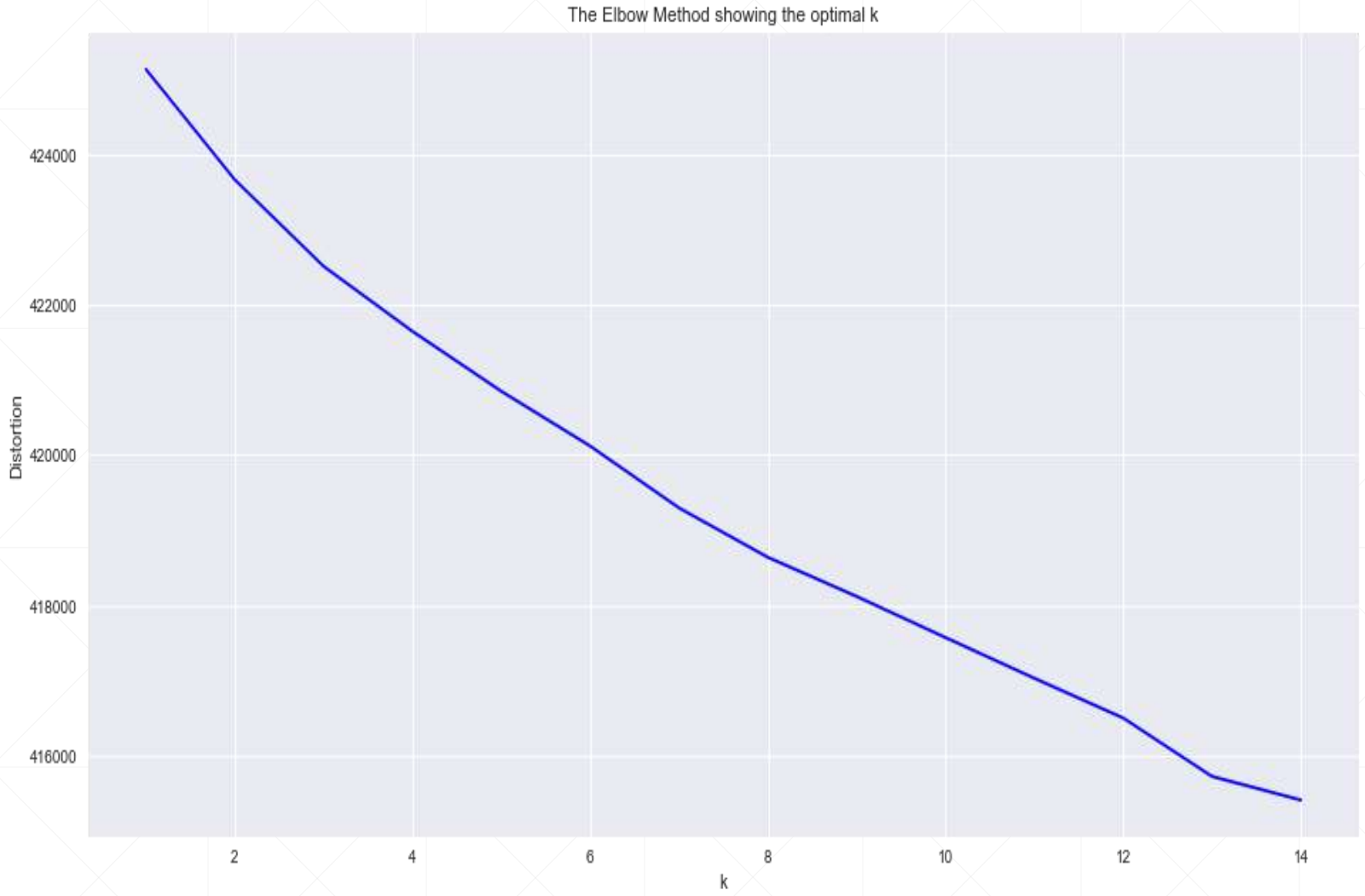
Count Vectorizer NMF

- Topic: 0(**Food menu**) - sauce, menu, cheese, fresh, dish, sweet, pork, flavor, salad, meat, taste, beef, meal, rice, fish, spicy, lunch, soup, cream, hot
- Topic: 1(**bars**) - bar, staff, night, drinks, table, drink, beer, area, coffee, hour, friends, happy, work, location, free, line, bartender, server, parking, friend
- Topic: 2(**Automotive**) - car, customer, work, manager, rental, honda, phone, cars, company, dealership, hours, vehicle, business, days, drive, oil, sales, guy, change, appointment
- Topic: 3(**Italian food**) - pizza, cheese, crust, sauce, topping, slice, sausage, salad, pepperoni, garlic, pie, slices, fresh, italian, delivery, bread, oven, half, dough, pasta
- Topic: 4(**restaurant**) - restaurant, table, menu, meal, server, dinner, waitress, restaurants, waiter, seated, dining, dishes, wine, manager, tables, dish, reservation, dessert, party, family
- Topic: 5(**Asian food**) chicken, fried, rice, sandwich, salad, spicy, sauce, crisp, lunch, fries, hot, wings, sides, curry, thai, soup, juicy, beans, cheese, dry

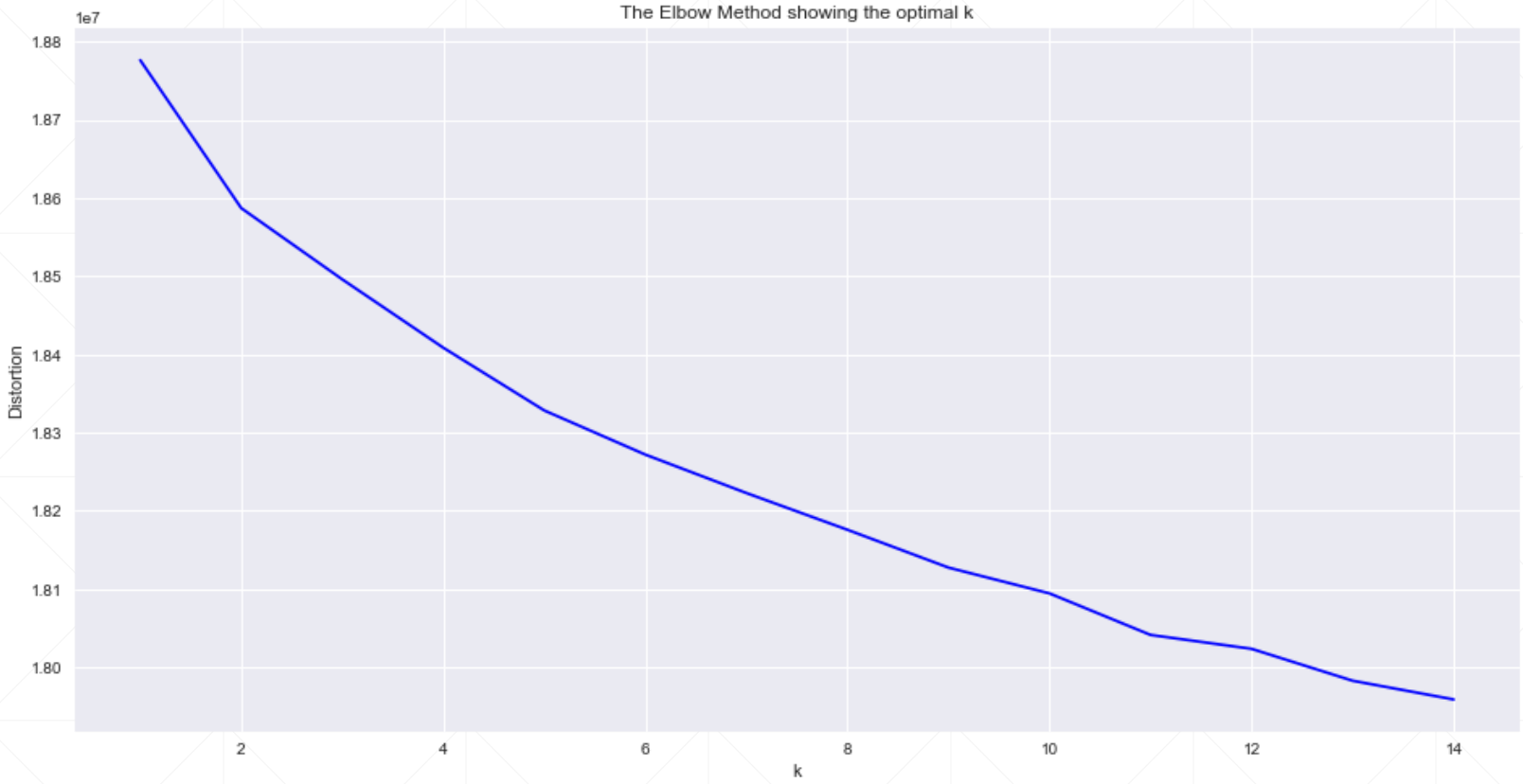
Topics


Number of reviews per topic
# of reviews
160000
140000
120000
100000
80000
60000
40000
20000
0
bars
food_menu
restaurant
Asian_food
Italian_food
Automotive
Topic

K-means clustering TF-IDF



K-means clustering Count Vectorizer



Recommendation System:

- Positive Recommendation System
- Negative Recommendation System

Negative Recommendation System:

- simple metric
- Example:

User ‘t5SRIRU6INiAyVkiMJhRPA’

Don’t go to these businesses :

[('Prides Osteria'), ('Bonchon Salem'),

("Santarpio's Pizza")]

business ‘Finz Seafood & Grill ‘

Similar businesses :

[('Scratch Kitchen', 2), ('Howling Wolf Taqueria', 2), ('Engine House Pizza', 2)]

Positive Recommendation System:

- SVD
- Example:

User ID# 2 is most similar to User ID# #1335

There are 1 businesses that user ID# 2 did not visit

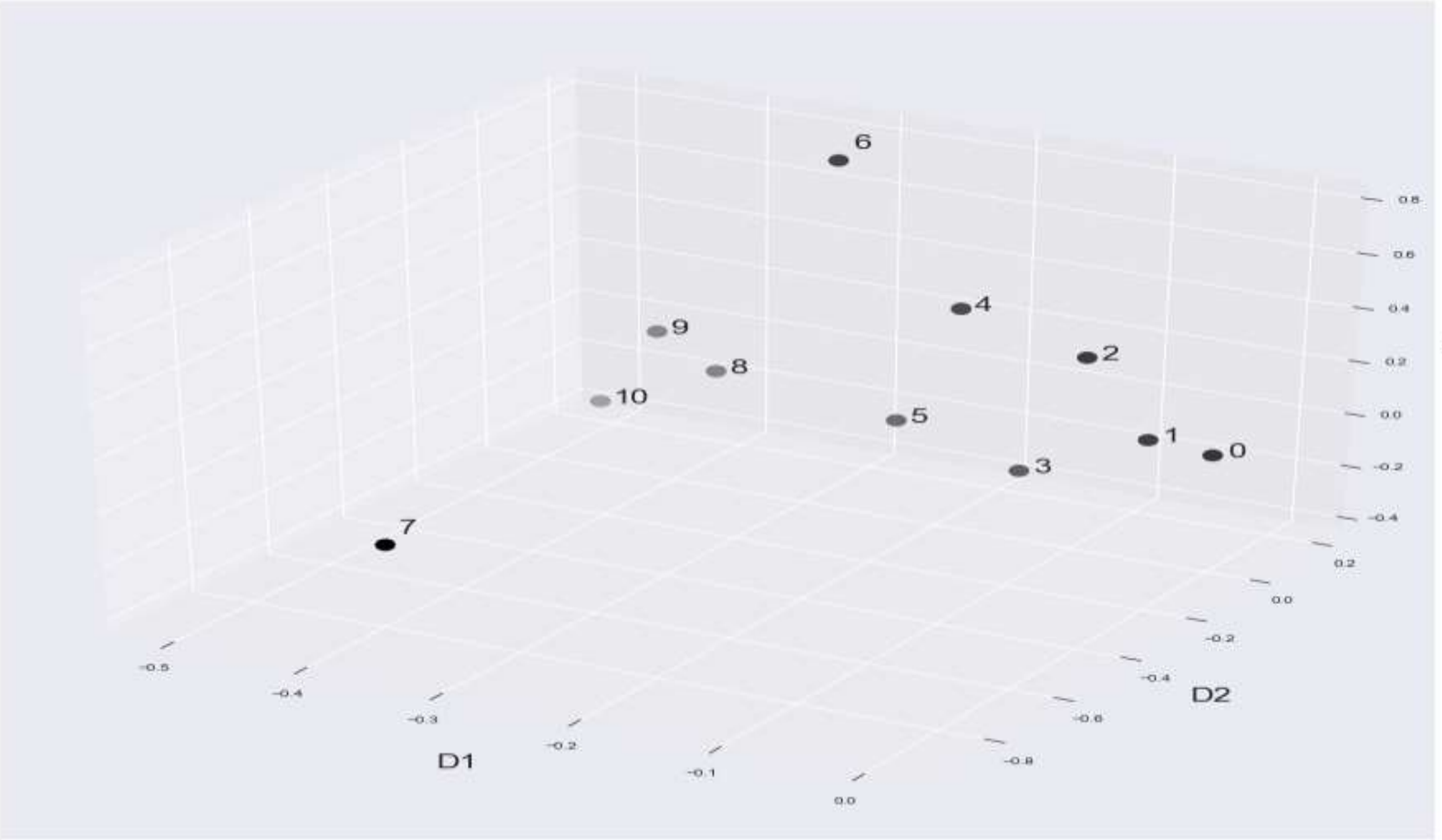
1 businesses for User ID# 2 to check out:

['Yamato Sushi Restaurant']

Recommendation System:

SVD

- Sample of 11 user.
- User 9,8 and 10 are close



Sentiment Review Classification

Model	Count Vectorizer			TF-IDF		
	Train	Validation	F1-SCORE	Train	Validation	F1-SCORE
Logistic Regression	0.944	0.936	0.959	0.940	0.937	0.960
MultinomialNB	0.891	0.891	0.930	0.898	0.898	0.937
BernoulliNB	0.870	0.871	0.918	0.870	0.871	0.918
Logistic Regression Weighted	0.938	0.927	0.952	0.934	0.927	0.953
Ada Boost	0.868	0.869	0.919	0.868	0.868	0.919

Conclusion

- There was no overfit with high f1-score.
- Logistic regression was best model.
- For topics NMF(6) with count vectorizer was the best.
- Yelp reviews is more on restaurants and food.

Model	Count Vectorizer		
	train	test	F1-SCORE
Logistic Regression	0.944	0.934	0.958

Thank you

Any question?