Yelp-Sentiment-Analysis

By Abdulrahman & Turki

Outline:

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

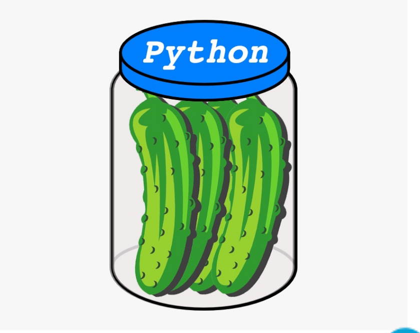
Project Scope & Tools:

- 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

pandas

scikit learn

NumPy



jupyter

matplotlib

seaborn

spaCy

NLTK

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

EDA

Removing duplicates

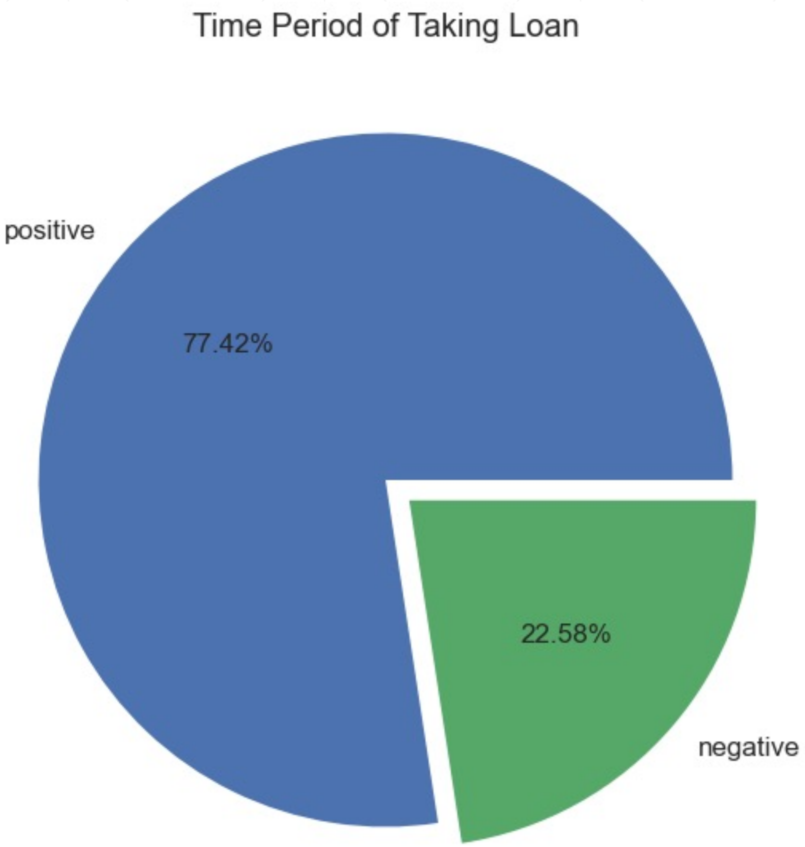
Removing null values

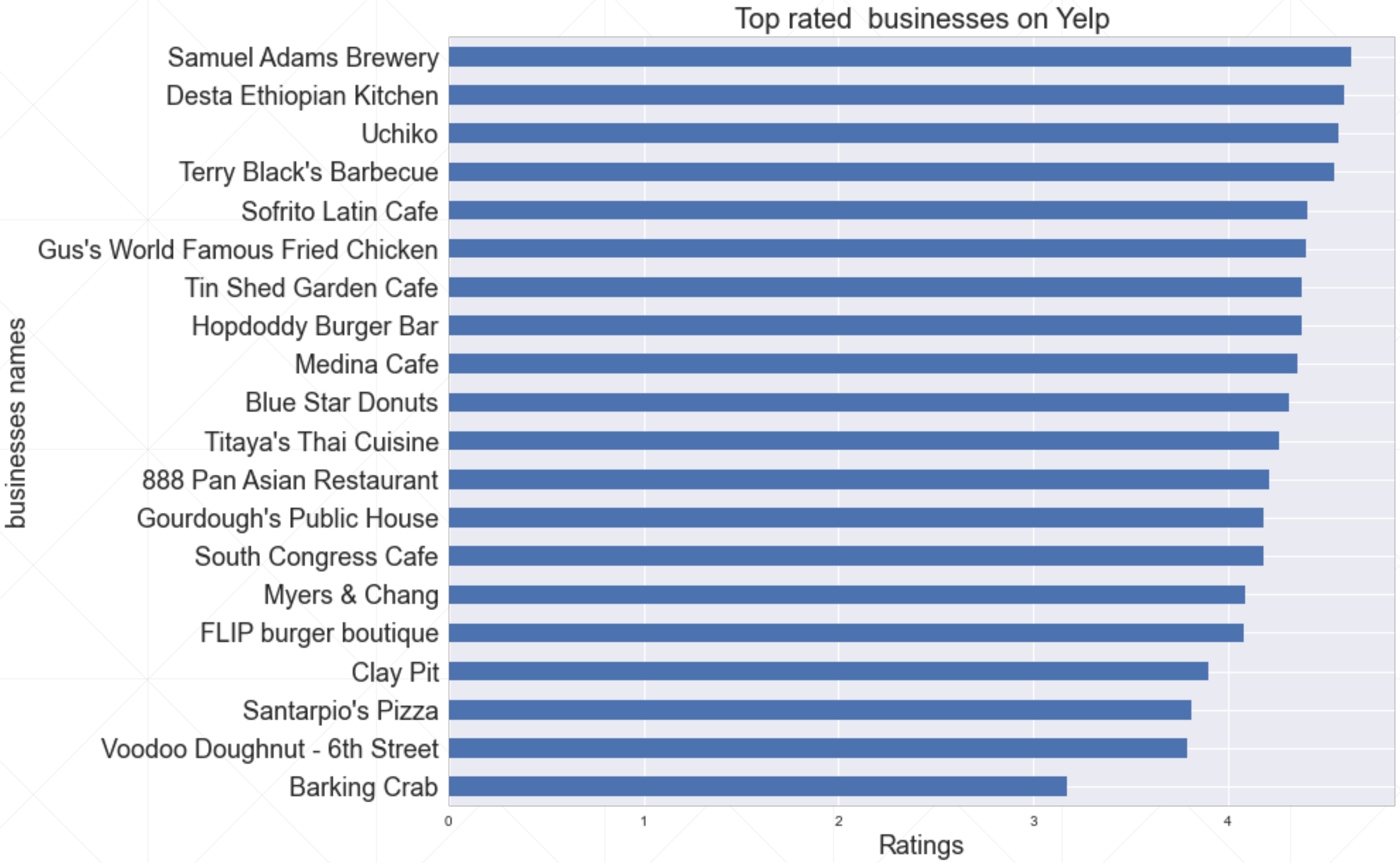
Removing columns

Striping values

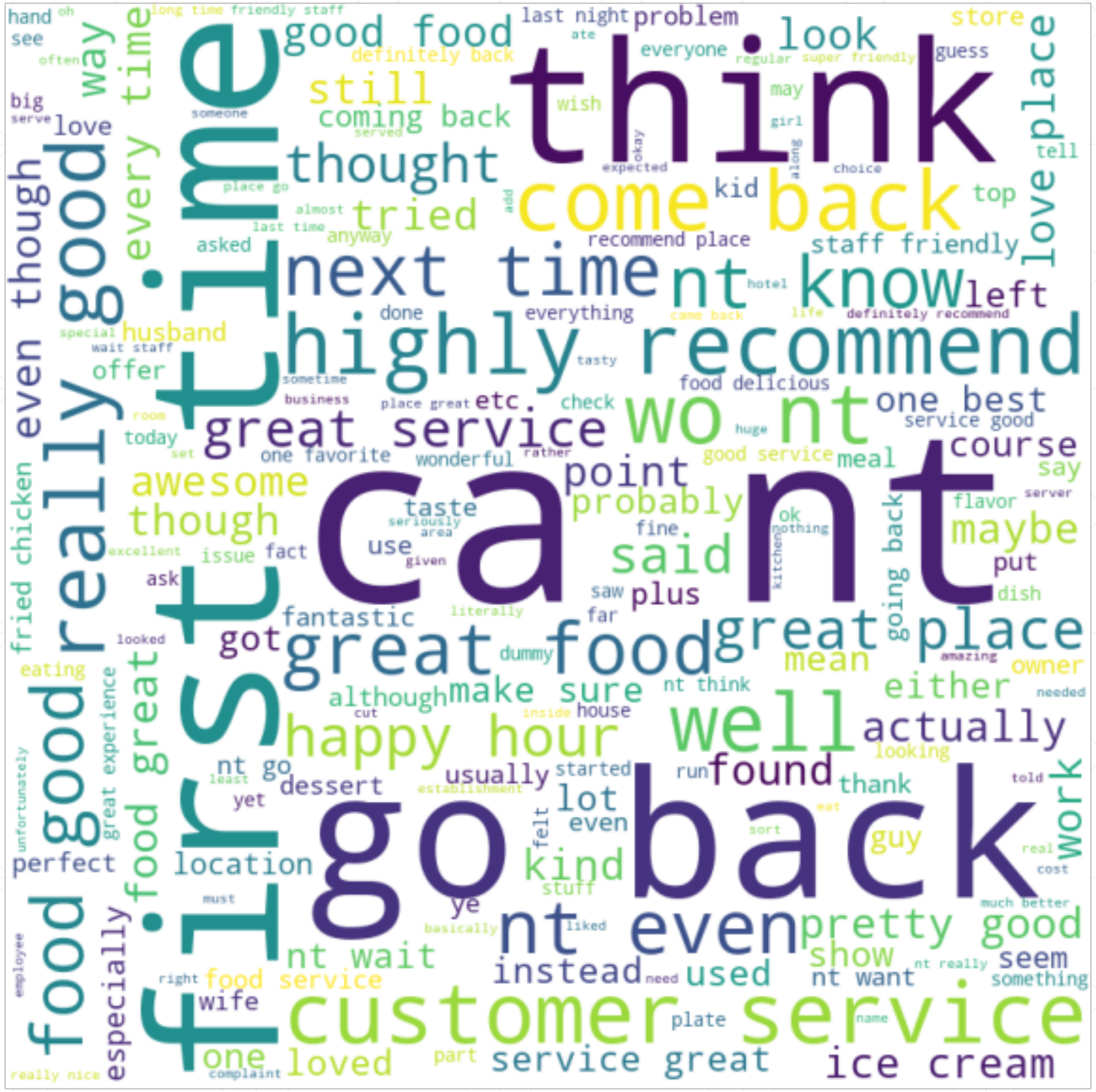
Changing stars column to positive and negative

EDA

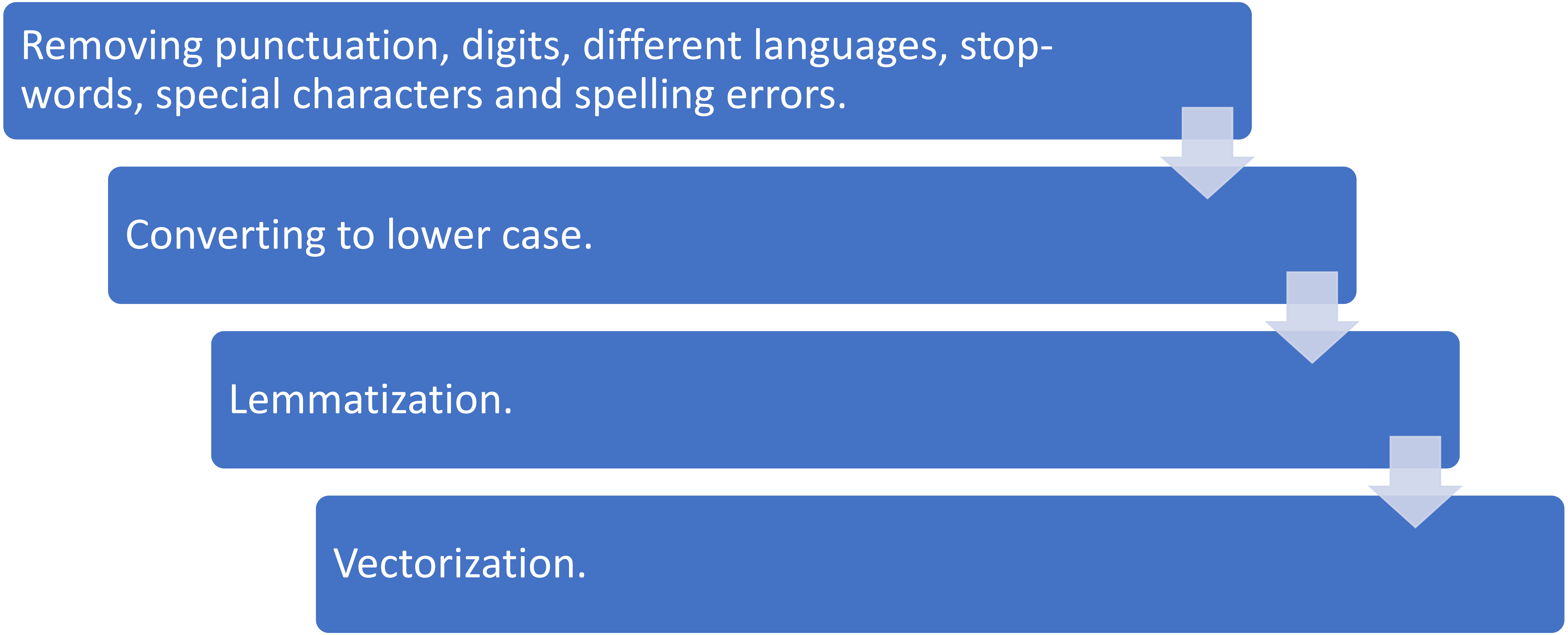

Time Period of Taking Loan
positive
77.42%
22.58%
negative


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
wo nt
really good
nt know
good food
thought
happy hour
great place
well
nt even
pretty good
food good
even though
coming back
still
tried
look
awesome
though
point
probably
said
actually
found
kind
location
perfect
ice cream
service great
loved
instead
used
staff friendly
love place
one best
course
maybe
make sure
although
fantastic
got
every time
way
husband
offer
fried chicken
food great
great experience
especially
wife
one
food service
nt wait
nt want
show
seem
something
plate
work
guy
either
mean
owner
thank
lot
even
usually
dessert
nt go
stuff
going back
put
dish
flavor
say
meal
good service
food delicious
recommend place
everything
done
asked
anyway
love
big
see
hand
last night
problem
everyone
guess
store
top
tell
left
kid
wish
may
check
etc
one favorite
wonderful
taste
use
fact
issue
ask
plus
saw
far
fine
ok
started
run
felt
ye
today
definitely back
definitely recommend
long time friendly staff
super friendly
regular
much better
really nice
eating
unfortunately
employee
complaint
part
looked
dummy
literally
nt think
house
needed
amazing
server
rather
huge
tasty
hotel
life
came back
place great
business
sometime
special
wait staff
room
set
excellent
seriously
area
given
cut
inside
establishment
least
yet
must
right
basically
liked
need
nt really
name
told
eat
real
cost
sort
kitchen
nothing
almost
last time
place go
add
expected
choice
okay
along
girl
someone
serve
often
oh
ate
served

NLP Pre-Processing



Removing punctuation, digits, different languages, 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)

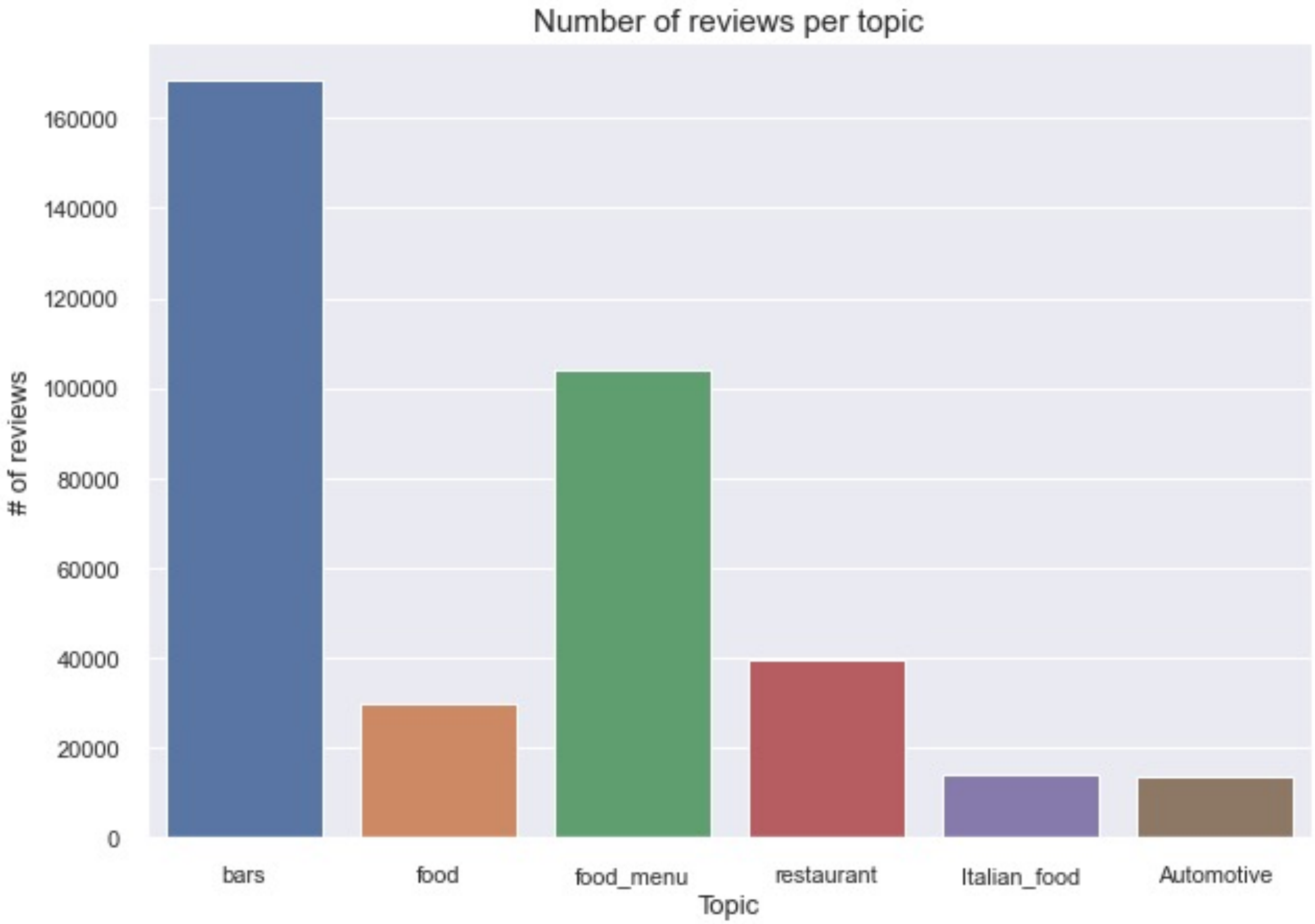
CV & TF-IDF

Best model was Count Vectorizer NMF with six topics

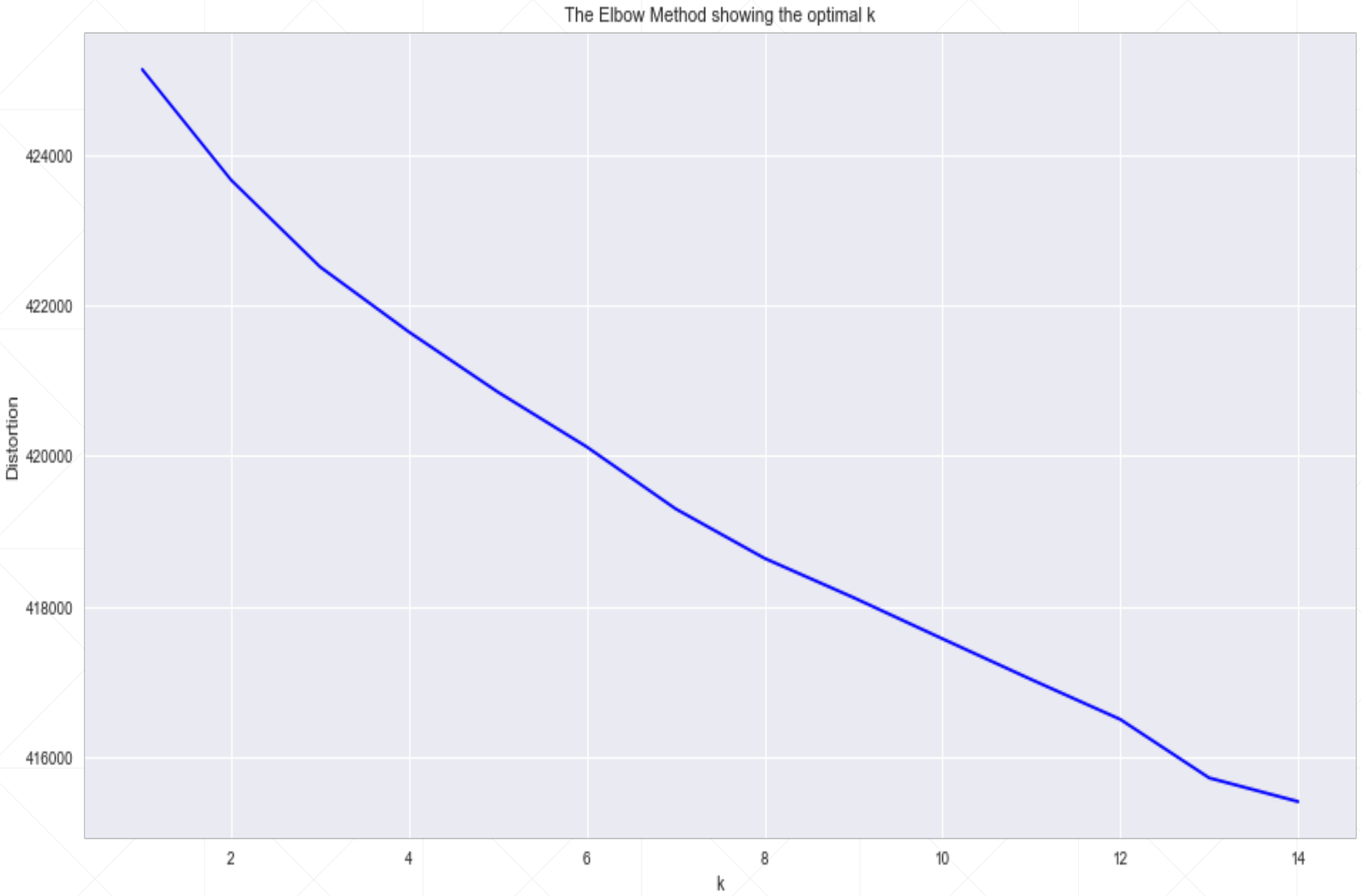
Count Vectorizer NMF

- 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(food) chicken, fried, rice, sandwich, salad, spicy, sauce, crisp, lunch, fries, hot, wings, sides, curry, thai, soup, juicy, beans, cheese, dry

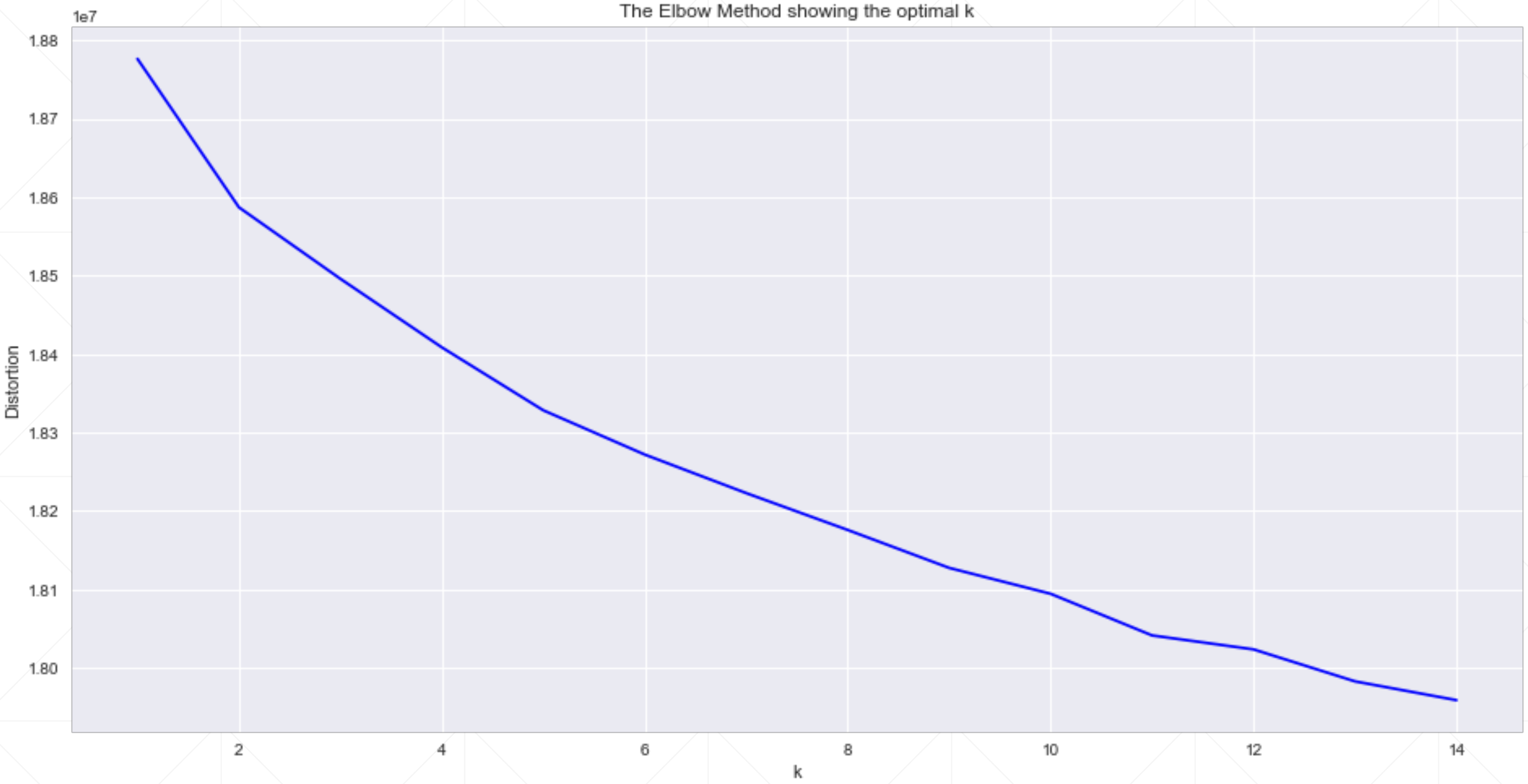
Topics


Number of reviews per topic
# of reviews
160000
140000
120000
100000
80000
60000
40000
20000
0
bars
food
food_menu
restaurant
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

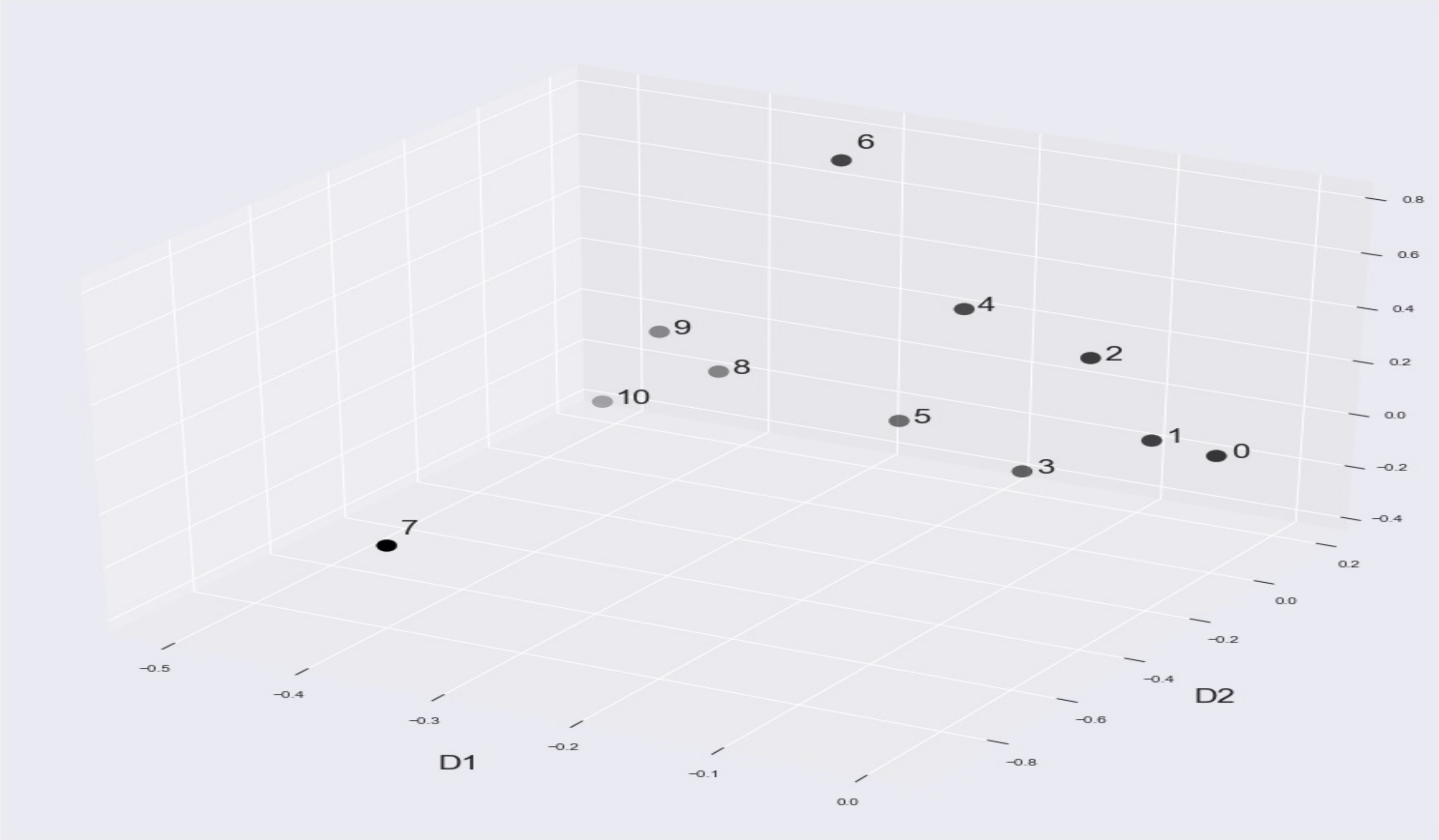
1 businesses for User ID# 2 to check out:

['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		0.940	0.937	
MultinomialNB	0.891	0.891		0.898	0.898	
BernoulliNB	0.870	0.871		0.870	0.871	
Logistic Regression Weighted	0.938	0.927		0.934	0.927	
Ada Boost	0.868	0.869		0.868	0.868	
Random Forest	0.8580	0.8616		0.9035	0.9067	
Extra Tree	0.8916	0.8856		0.9266	0.9179	

Conclusion

Testing selected model on testing data

- Combining train and validation data
- Training the model
- Using testing data to get scores

Model name	Training	Testing	F1 scores	Precision	Recall
Logistic Regression	0.938	0.939	0.544	0.520	0.569

Thank you

Any question?