

[illegible]

TABLE OF CONTENTS

01

INTRODUCTION

02

PREPROCESSING

03

DATA ANALYSIS

04

TOPIC MODELING

05

CLUSTERING

06

RECOMMENDATION
SYSTEM

07

MODELS

08

CONCLUSION



01

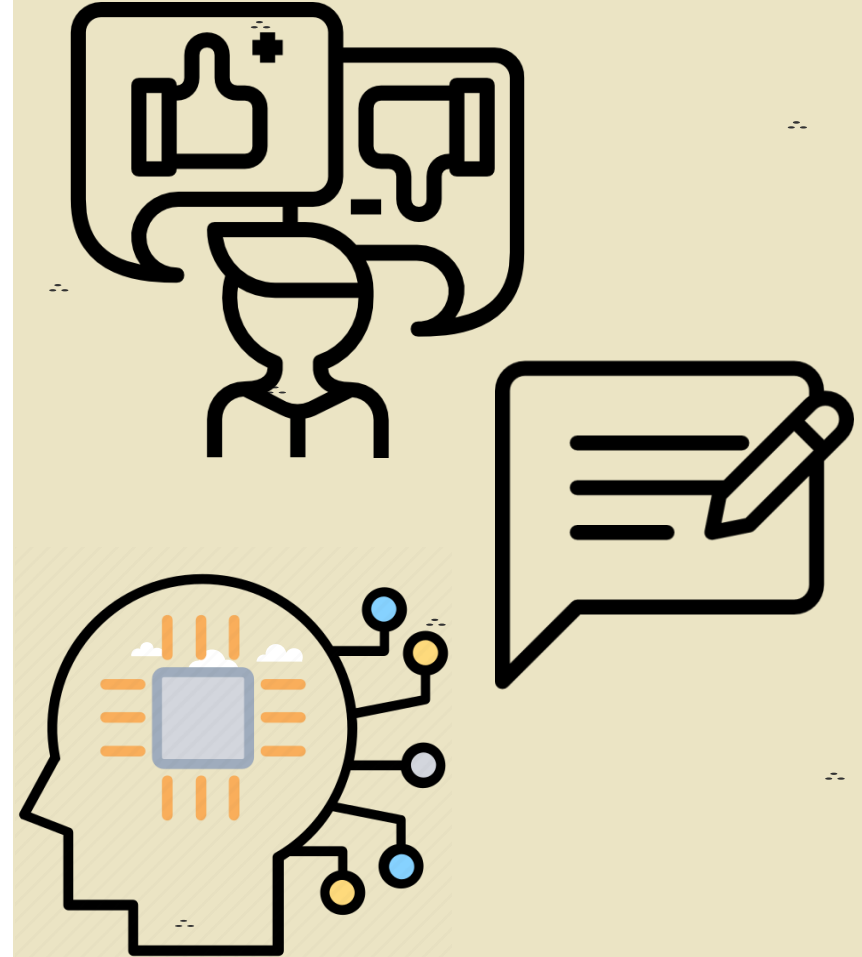
INTRODUCTION

Backstory

Users' feedback

To improve and avoid negative things and keep the positives.

Reviews can be analyzed to understand where a hotel failed to provide a good customer experience and where it succeed using NLP.



DATASET

Scraped data from [Booking.com](#)

515K Hotel Reviews Data in Europe (Kaggle)

515,738 Rows → 1,031,476 Rows

17 Columns

Hotel Name

Hotel Address

Reviewer
Nationality

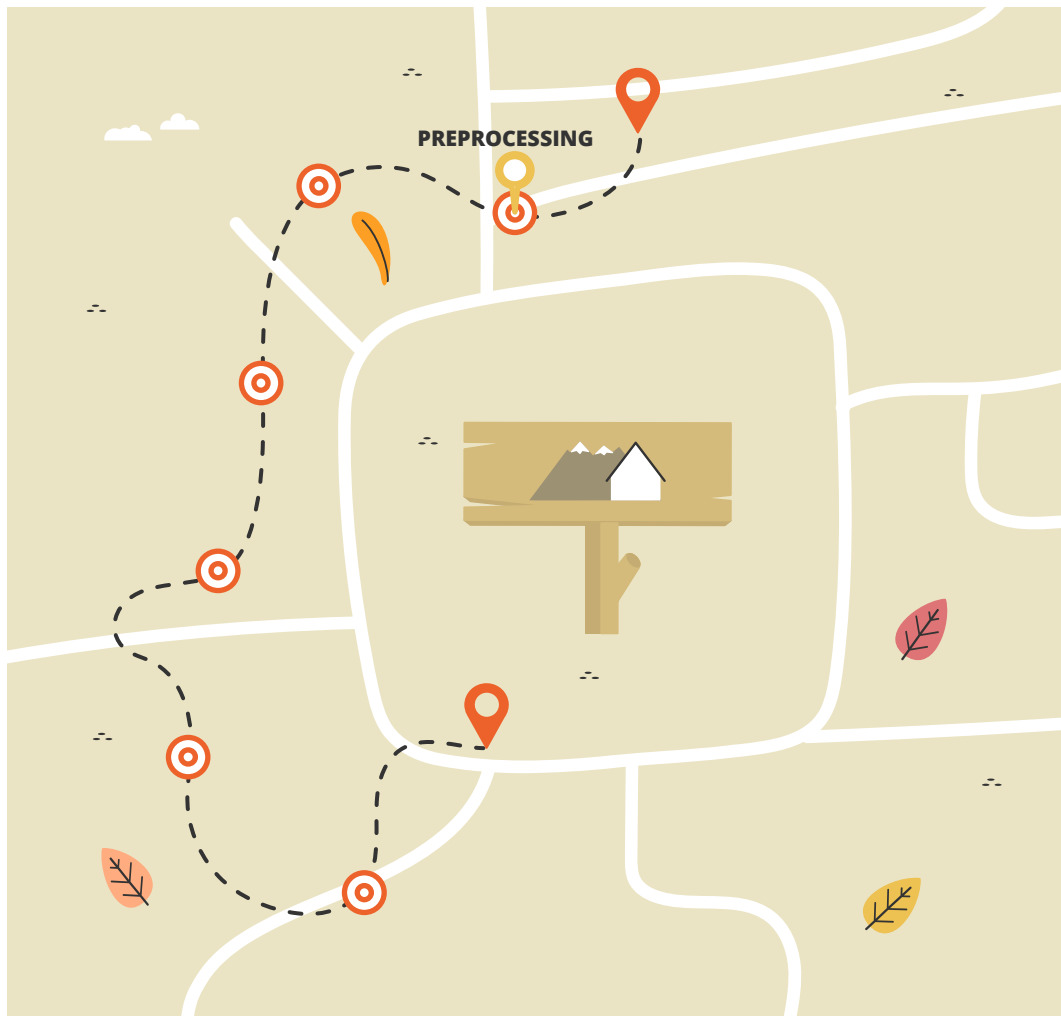
Negative
Review

Positive
Review

Reviewer
Score

Tools

- Pandas
- Numpy
- Matplotlib
- Seaborn
- WordCloud
- Sklearn
- NLTK
- Autocorrection
- CorEx
- Pickle



02

PREPROCESSING

PREPROCESSING

**Data
Cleaning**

**Spelling
Correction**

Vectorization

Lemmatization

NLP pipe

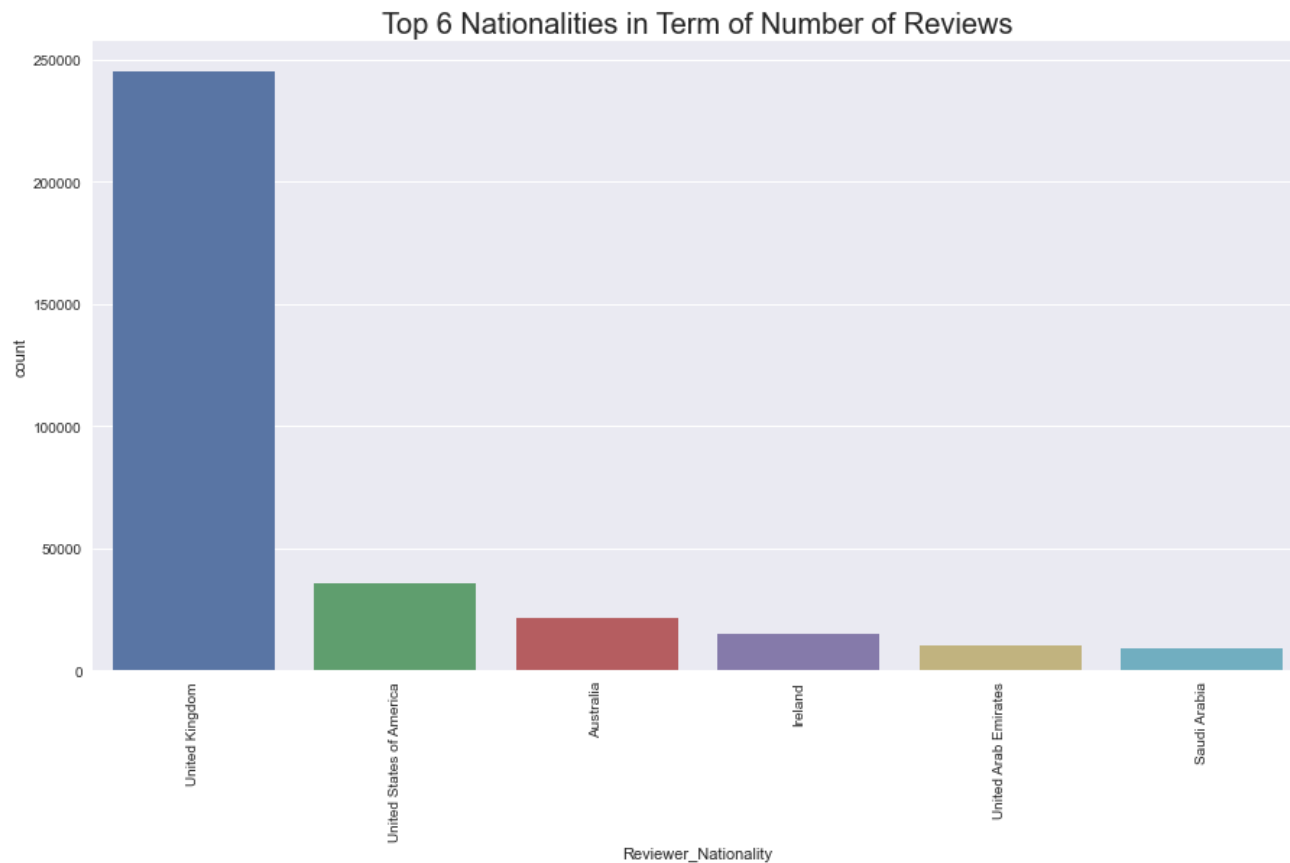
DATA ANALYSIS



03

DATA ANALYSIS

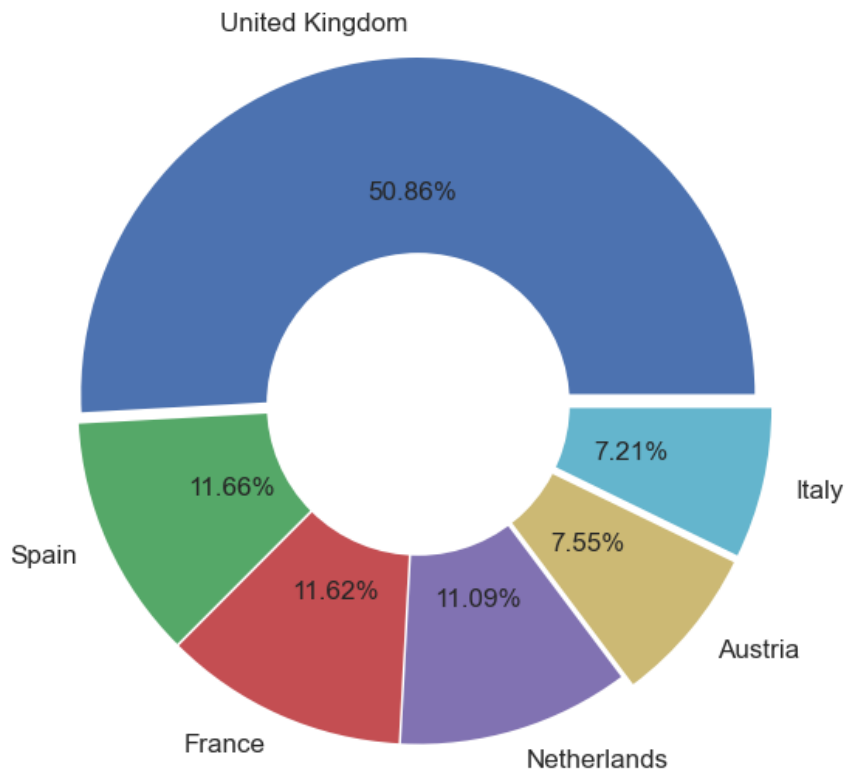
Top 6 Nationalities Plot





Hotels in Each Country plot

Comparison between reviewed hotels in each county

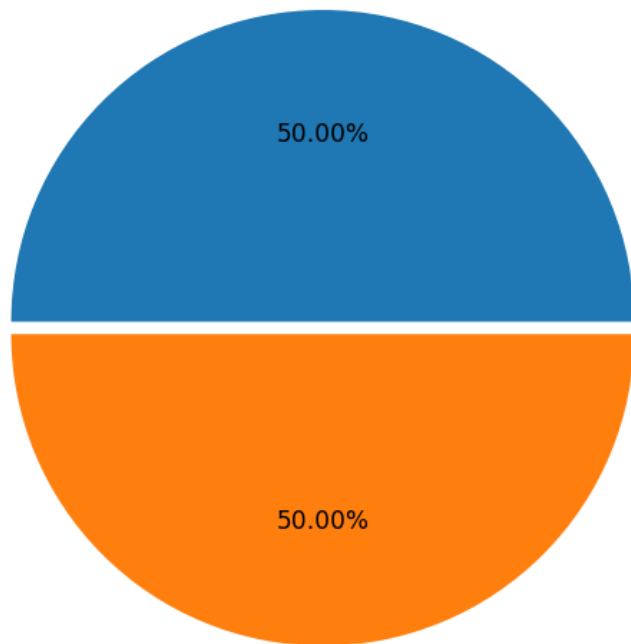


Reviews Type plot

Before Cleaning

Positive Review Vs Negative Review

Positive Review

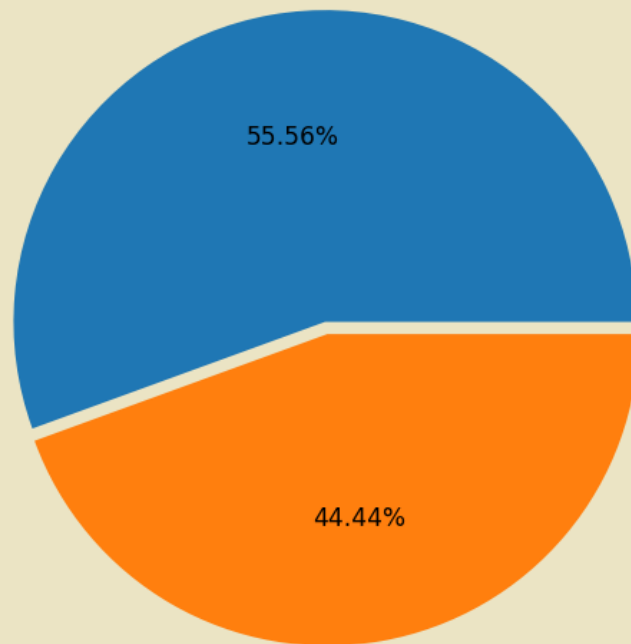


Negative Review

After Cleaning

Positive Review Vs Negative Review

Positive Review



Negative Review

—

A word cloud visualization of hotel reviews. The most prominent words are "room", "staff", "breakfast", "noise", "clean", "price", "service", "location", "comfortable", "well", "stay", "much", "one", "time", "better", "great", "good", "night", "dirty", "booked", "lack", "hot", "thing", "left", "used", "wifi", "parking", "make", "around", "really", "space", "check", "nothing", "bar", "floor", "office", "bathroom", "shower", "window", "food", "cold", "pay", "helpful", "friendly", "thank", "told", "lobby", "people", "limited", "clean", "drunk", "change", "said", "star", "noise", "guests", "seamless", "smooth", "carpet", "warm", "newer", "facilities", "usb", "port", "side", "difficult", "restaurant", "help", "couple", "arrived", "quite", "work", "free", "noisy", "little", "unpleasant".

[illegible]



04

TOPIC MODELING

LSA

NMF

CorEx

TOPIC MODELING

LSA

Topic 1

room, staff, location, helpful, clean, nice, comfortable, excellent, bed, lovely

Topic 2

room, bed, comfortable, clean, nice, bathroom, spacious, view, big, small

Topic 3

location, room, station, perfect, close, central, walk, excellent, near, tube

**Staff
Room
Location**

NMF

Topic 1

room, comfortable, clean, bed, lovely, bathroom, spacious, view, quiet, small

Topic 2

staff, helpful, excellent, lovely, clean, reception, extremely, restaurant, bar, comfortable

Topic 3

location, excellent, perfect, station, close, central, walk, near, tube, london

Topic 4

nice, station, close, bed, restaurant, walk, metro, area, minute, clean

**Room
Staff
Location
Surroundings**

TOPIC MODELING

CorEx

Topic 1

Location(station)

station,close,walk,metro,walking,tube,distance,near,walking distance,easy

Topic 2

Room

room,comfortable,clean,bed,rooms,spacious,positive,bathroom,nice,staff

Topic 3

Facilities

bar,coffee,tea,reception,water,restaurant,pool,complimentary,park,arrival



05

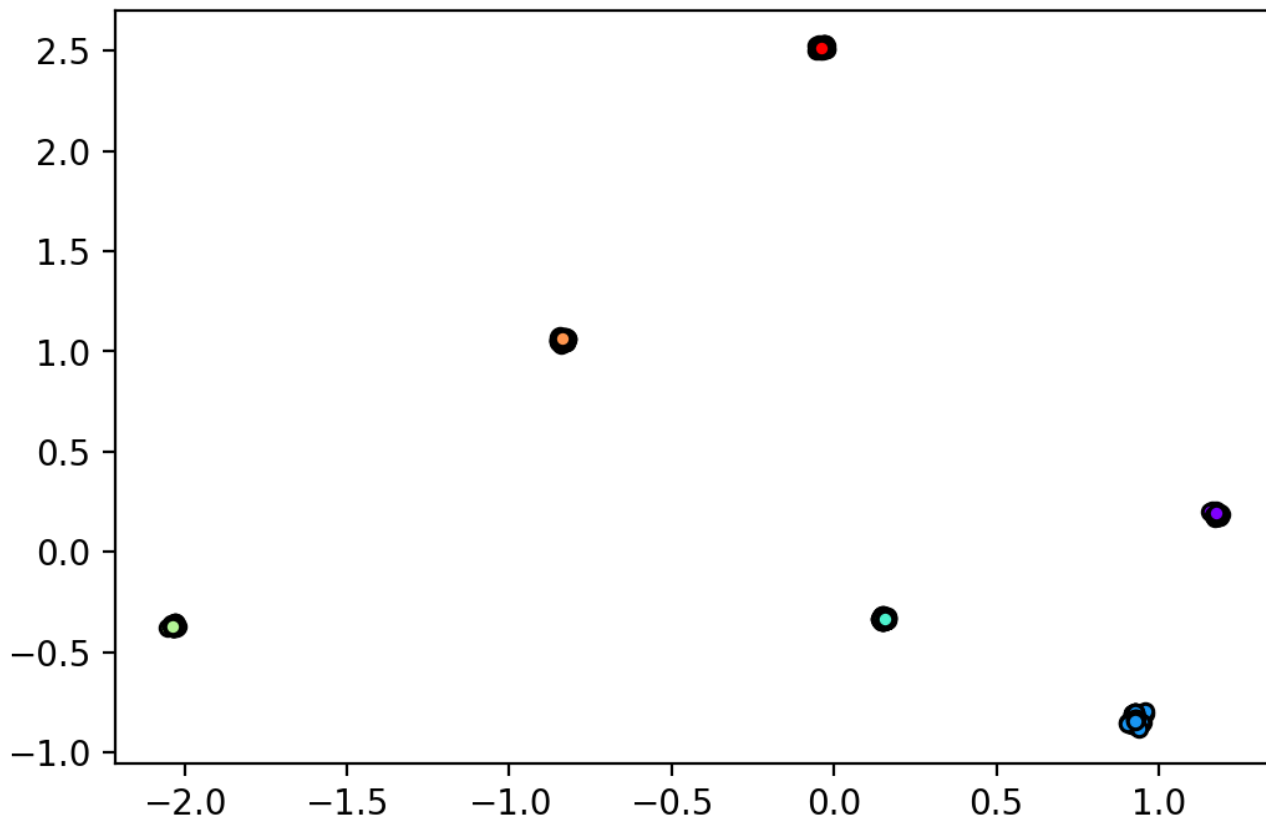
CLUSTERING

DBSCAN

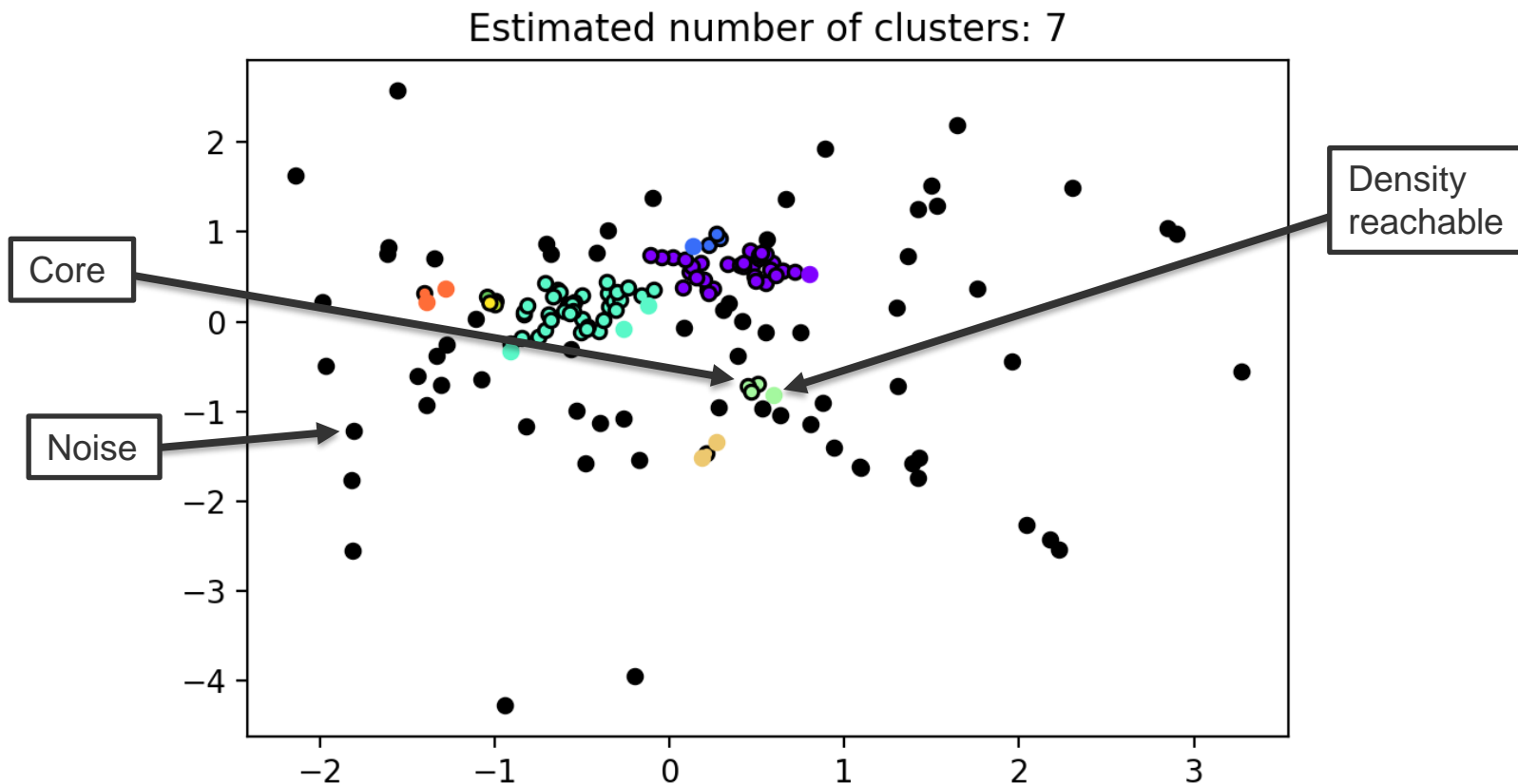
eps=0.15
min samples=3

Clustering using Lat & Lng For All Data

Estimated number of clusters: 6



Clustering using Lat & Lng (Italy)





06

RECOMMENDATION SYSTEM

SVD Recommendation

=====People from Saudi Arabia=====

===Recommended hotels in United Kingdom===

Users from Saudi Arabia are most similar to users from Singapore.

There are 50 hotels that people from Saudi Arabia did not visit, they might like

10 Hotels for people from Saudi Arabia to check out:

['Andaz London Liverpool Street', 'Bermondsey Square Hotel A Bespoke Hotel', 'Best Western Seraphine Kensington Olympia', 'Blakes Hotel', 'Caesar Hotel', 'Charlotte Street Hotel', 'Comfort Inn Suites Kings Cross St Pancras', 'Covent Garden Hotel', 'De Vere Devonport House', 'DoubleTree By Hilton London Excel']

===Recommended hotels in Spain===

Users from Saudi Arabia are most similar to users from Gibraltar.

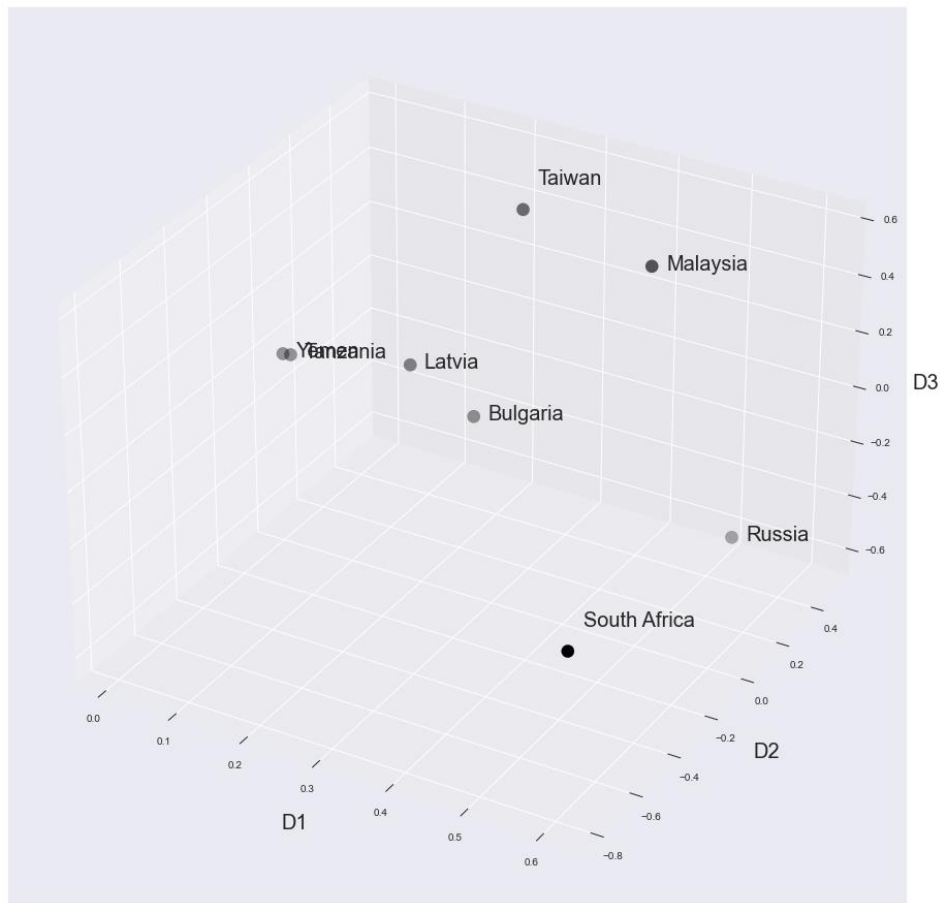
There are 9 hotels that people from Saudi Arabia did not visit, they might like

9 Hotels for people from Saudi Arabia to check out:

['AC Hotel Sants a Marriott Lifestyle Hotel', 'Aparthotel Atenea Barcelona', 'Attica 21 Barcelona Mar', 'Catalonia Diagonal Centro', 'Catalonia Eixample 1864', 'Hotel Omm', 'Hotel Spa Villa Olimpica Suites', 'Mercure Barcelona Condor', 'NH Sants Barcelona']



3D Similarity Plot for 8 Nationalities



Recommendation

Similar Hotels

The Kensington Hotel:

Park Grand Paddington Court

Park Plaza Westminster Bridge London

Best Western Premier Hotel Couture

Similar Users

Kuwait:

United Arab Emirates

Saudi Arabia

Canada

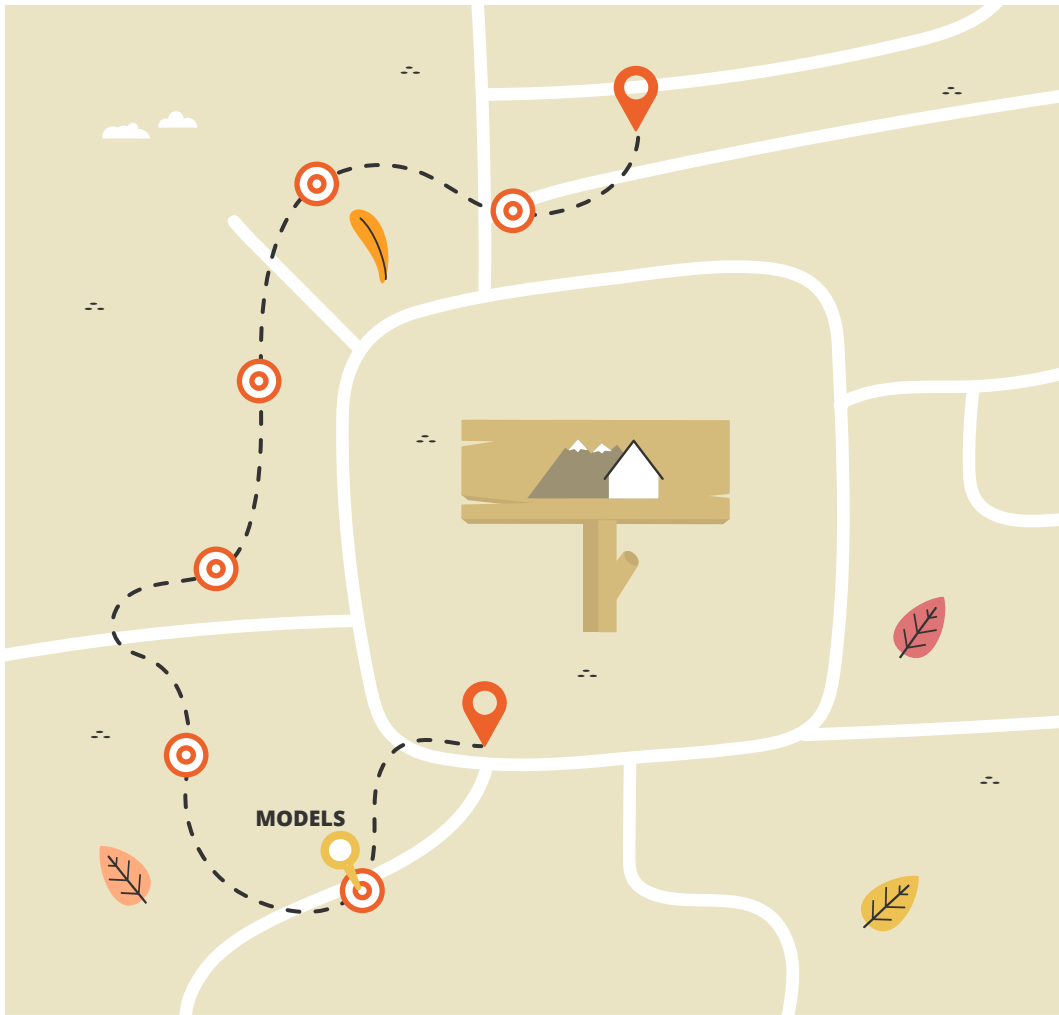
Recommended Hotel

Saudi Arabia:

Park Plaza Westminster Bridge London

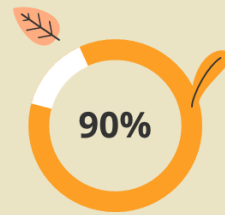
The Student Hotel Amsterdam City

Best Western Premier Hotel Couture

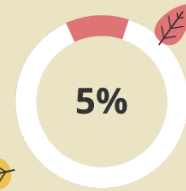


07

MODELS



Train



Validation



Test



Classification Models

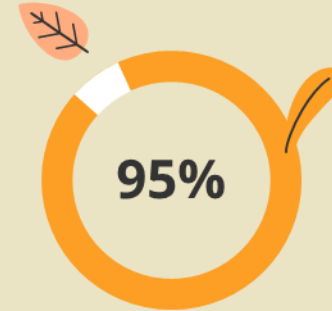
Model	Accuracy	
	Train	Validation
Counter Vectorization		
Logistic Regression	0.955	0.941
Bernoulli NB	0.889	0.877
Decision Tree	0.991	0.877
TF-IDF		
Logistic Regression	0.948	0.941
Bernoulli NB	0.880	0.877
Decision Tree	0.982	0.881



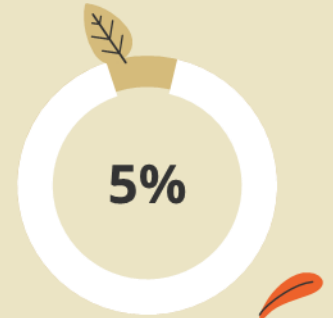


08

CONCLUSION



Train
Validation



Test

CONCLUSION

Logistic Regression TF-IDF

Accuracy

Train & validation: 0.948

Test: 0.942