DestinationPredict

GROUP 9 - GLOBETROTTERS

John Elvis

Sellah Otieno

Fidelis Wanalwenge

Bernadette Nganga

Lilian Ngige

Matilda Odalo

INTRODUCTION

- The aim of this project is to revolutionize the Travel industry by addressing a critical need for personalized travel destination recommendations.
- Traditional travel advisory services fall short in providing tailored experiences, leading to a lack of engagement among modern adventurers.
- The Destination Predict is a data product that allows future travelers to get a prediction for their perfect destination with the input of just a few words
- Our project aims to fill this gap by deploying a data-driven solution that enhances user engagement and rejuvenates the travel sector.

BUSINESS OVERVIEW

- In the today's landscape of travel options, modern adventurers are faced with an overwhelming array of choices. The current traditional travel advisory services are **time-consuming** and **often lack personalization**, relying on generic recommendations that fail to capture individual preferences and evolving travel trends.

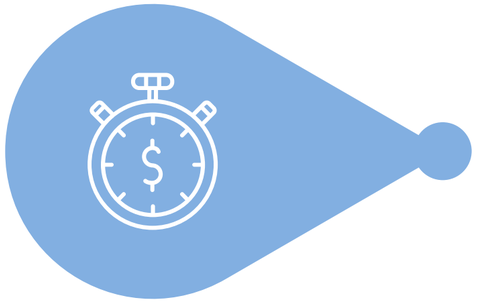
- Navigating through this information overload can be a daunting task for travelers seeking unique tailored and worthy experiences. There is a pressing need for a more personalized recommender system in the travel industry to bridge this gap and provide personalized, data-driven recommendations.

BUSINESS PROBLEM

- The existing travel advisory services fall short in delivering personalized recommendations, relying on outdated practices that neglect individual preferences and the dynamic nature of travel trends.

- This lack of personalization results in an overwhelming experience for modern adventurers who seek unique, tailored, and meaningful journeys.

PROJECT OBJECTIVES



01

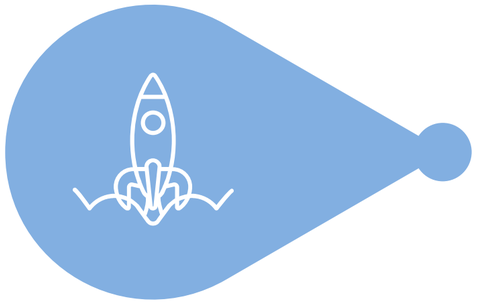
OBJECTIVE 01

Develop a recommendation system using data from top attraction sites to help travelers choose their favorable destination site based on user-provided text.

02

OBJECTIVE 02

Utilize NLP to train the Destination Dictionary on text data, allowing it to effectively interpret and predict travel preferences expressed through user-inputted text.



03

OBJECTIVE 03

Create an intuitive and visually appealing web interface for users to input their preferences and receive destination recommendations.

SOURCE OF DATA

The dataset utilized in this project was scraped from TripAdvisor's list of Traveler's Choice destinations for Popular World Destinations.

We compiled the dataset by extracting information from the attractions titles for 12 top cities featured in TripAdvisor's Top Picks per city.The final dataset included over 28,000 unique text values.

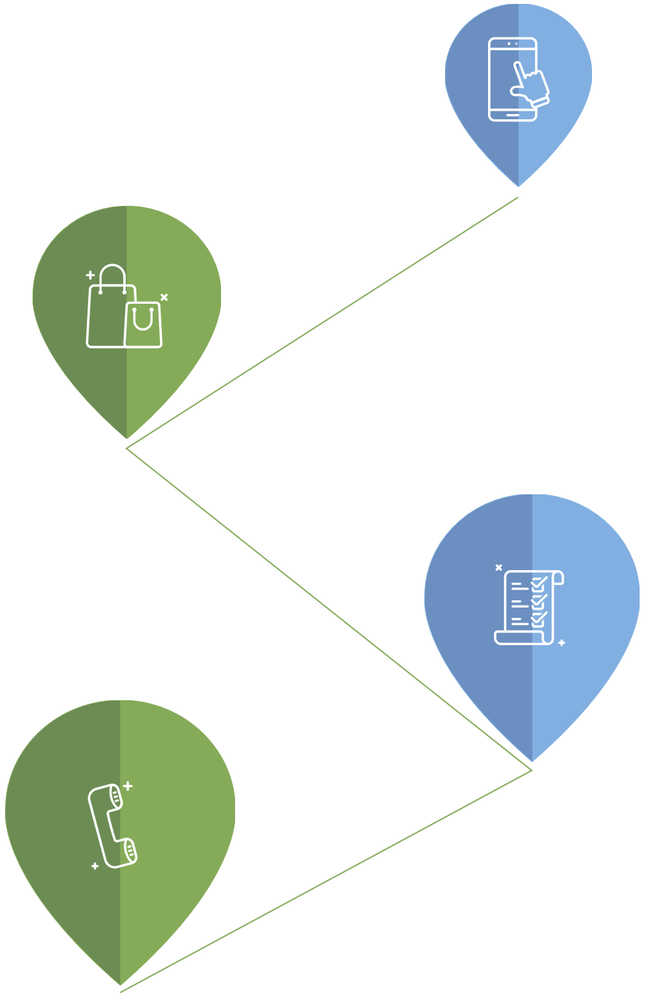
DATA COLLECTION

01

This project utilized data from 12 top cities from TripAdvisor's list of Traveler's Choice destinations for Popular World Destinations

02

data found on the website below; (https://www.tripadvisor.com/TravelersChoice-Destinations)



03

The dataset was compiled by scraping the titles from Tripadvisor 'attractions' for each of the 12 cities.

04

The final dataset included over 28,000 unique text values.

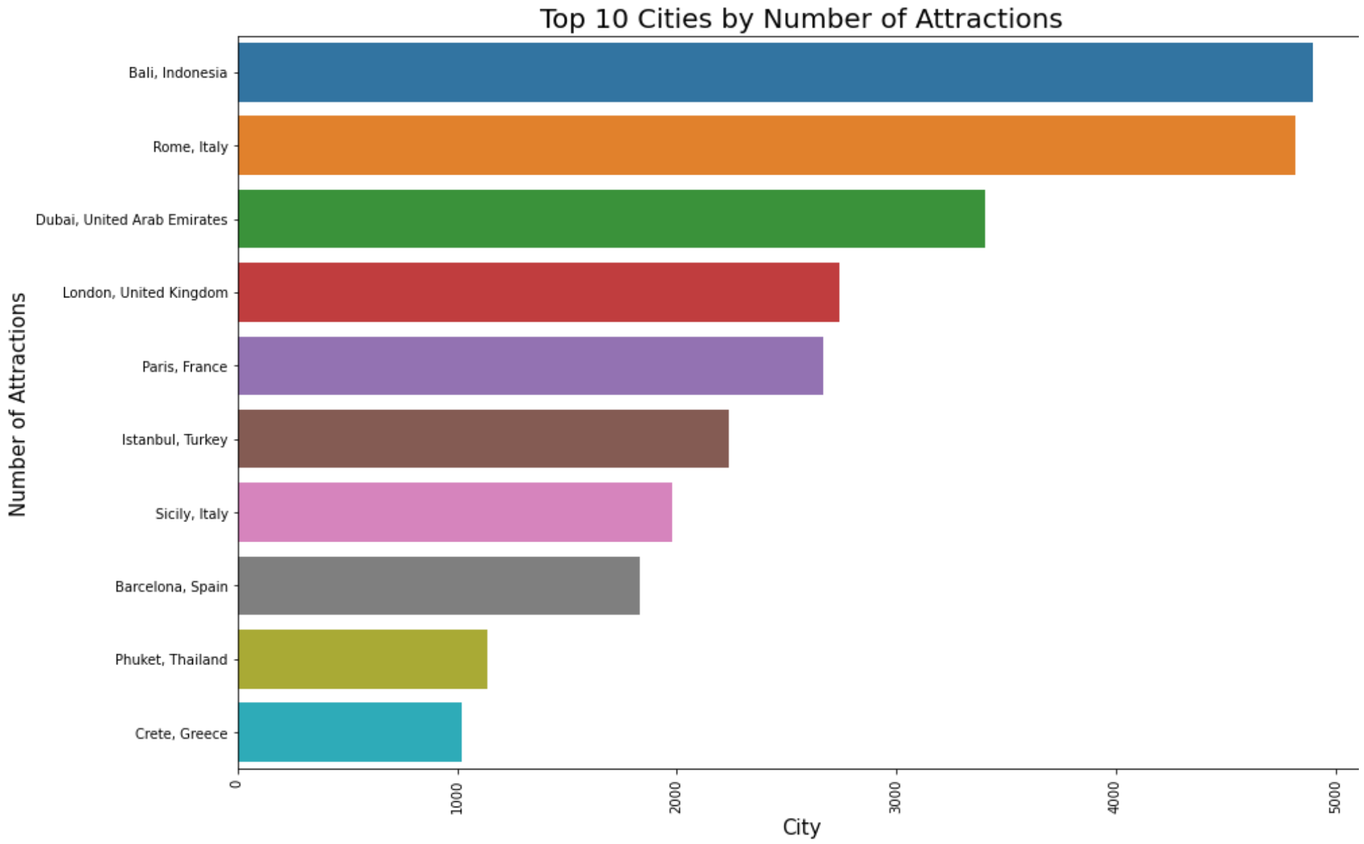
EXPLORATORY DATA ANALYSIS

- The EDA for this project was mainly devoted to exploring some cleaning and preprocessing techniques for text data using NLP. We used 3 vectorization strategies for the data and looked into specific words and phrases that needed to be cleaned from the dataset namely: Count, TF-IDF and Count Vectorization using Bi-Grams.

- After analyzing document term matrices, word clouds, and frequent words from three vectorization techniques, TF-IDF vectorization emerged as the most effective.

- This method aligns well with the DestinationPredict's aim of city-specific text classification. TF-IDF works by emphasizing the significance of words in the context of each city.

DATA PROCCESSING



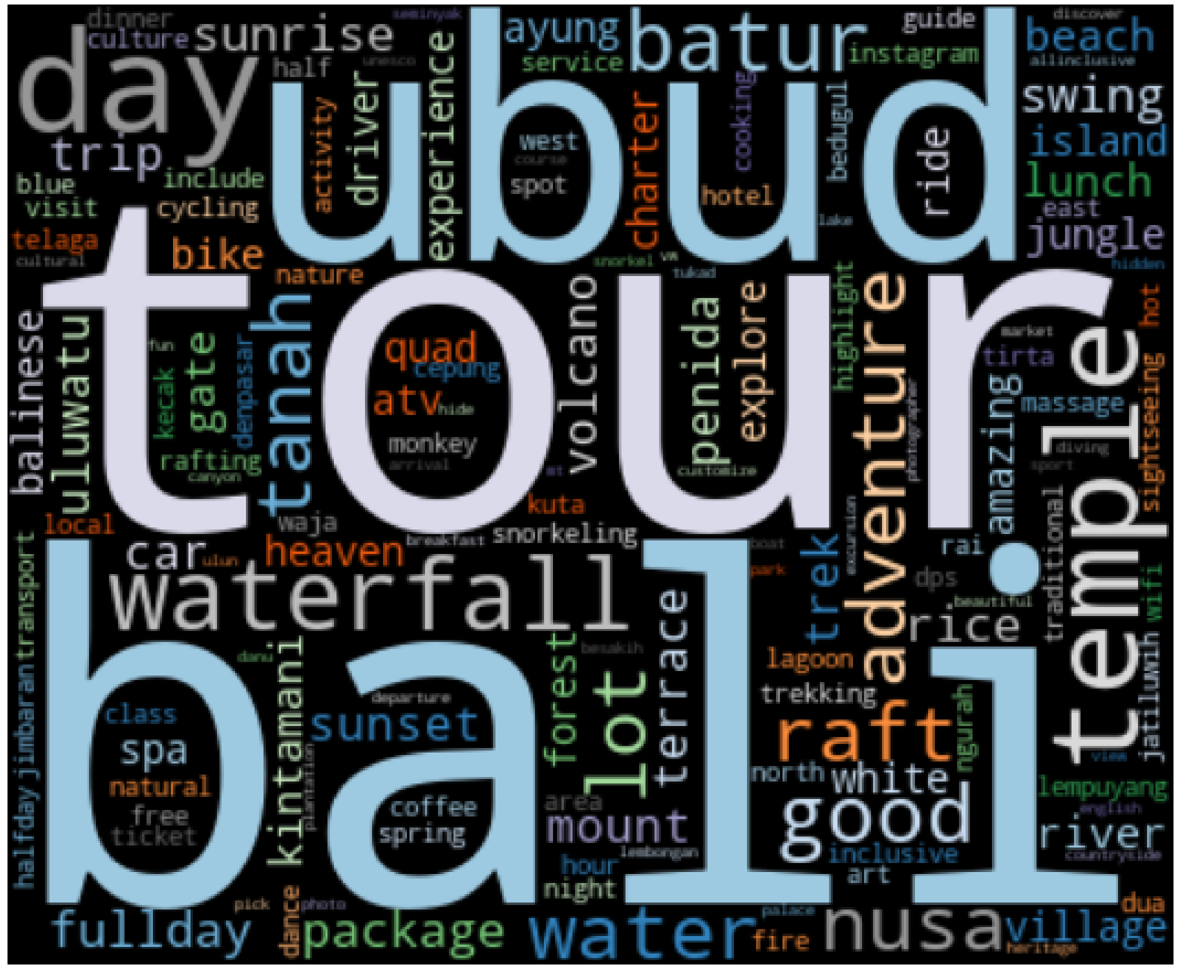
- City of Bali in Indonesia has the most number of attraction sites in the Trip Advisor dataset.

- We note the number of attraction sites are not even in the cities. To mitigate the effects of an imbalanced dataset we assigned higher weights to under represented classes.
- By setting these weights, we mitigate the effects of imbalanced datasets and achieve a more equitable performance across all destination classes.

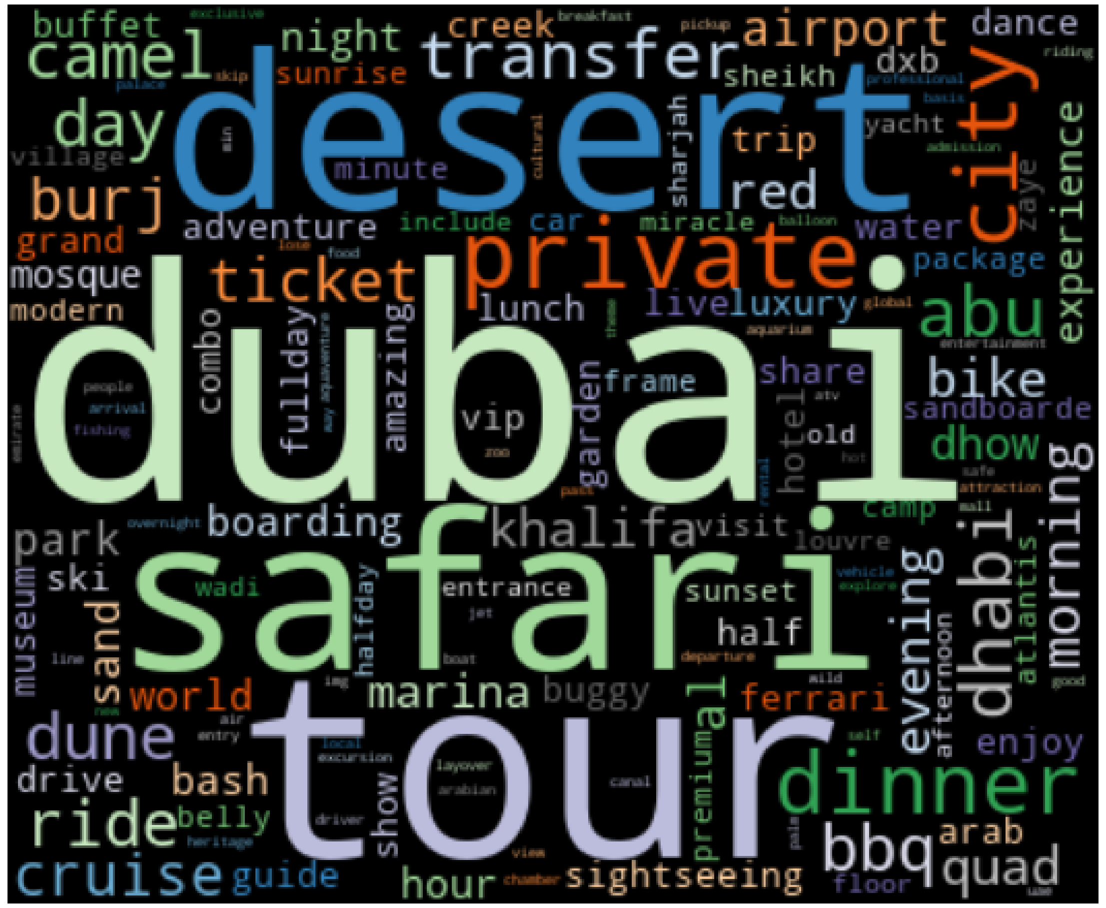
Text Data Preprocessing

- Filtering of the text data was done by removing stopwords(is, the and,an,in) ,numbers and punctuations.
- Different methods of vectorization used: count vectorization & tf-idf vectorization.
- We created a document term matrix based on the processed text dataset which was used to create wordclouds for attraction sites and their respective cities

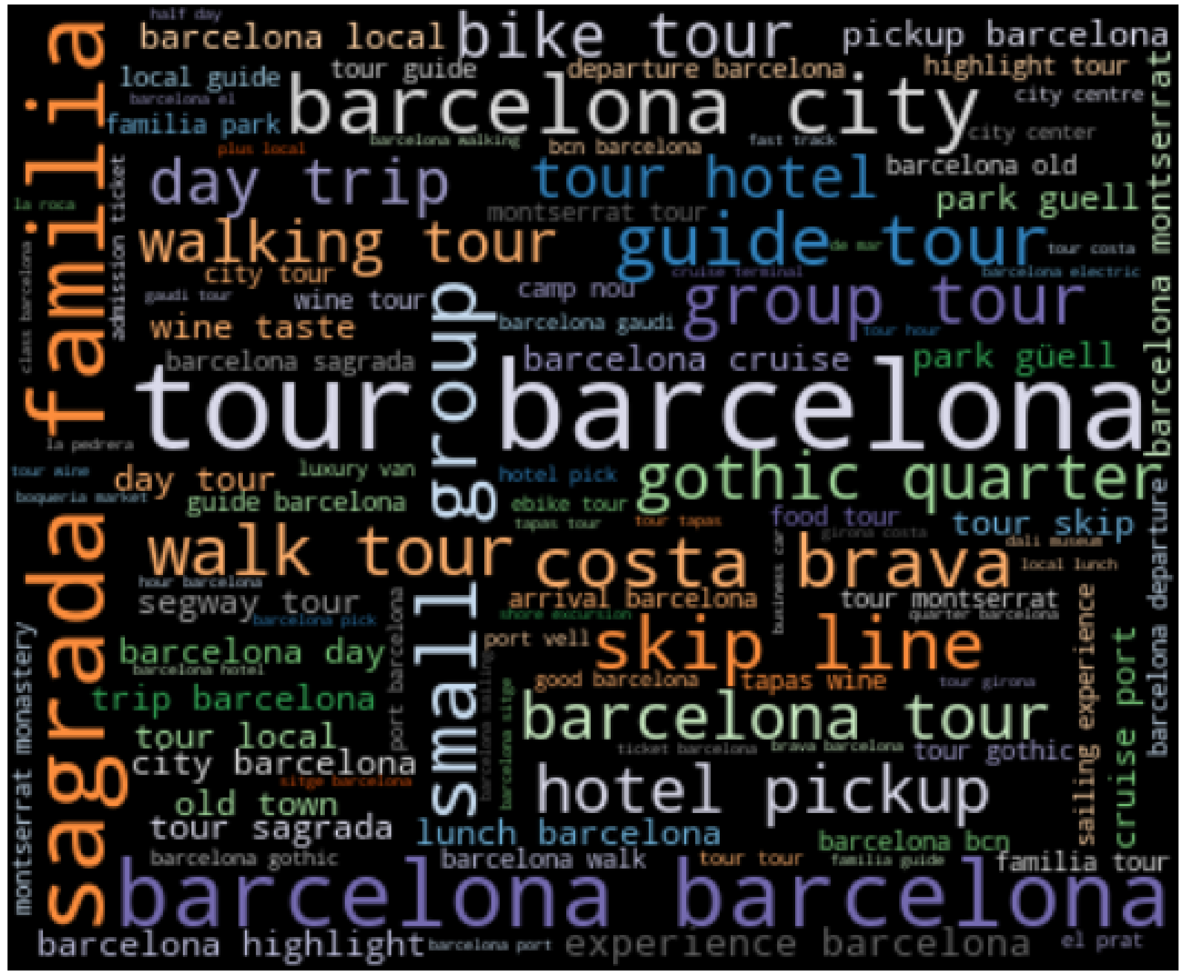
Bali, Indonesia



Dubai, United Arab Emirates



Barcelona, Spain

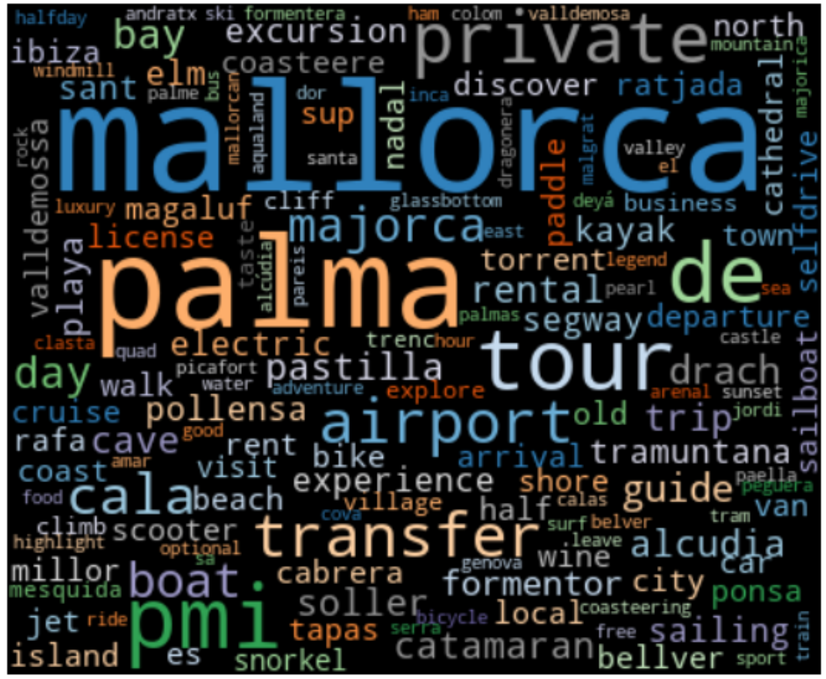


From the wordclouds,there is a lot of repetition with words like 'tour', 'private', 'transfer' between cities because all of these cities will have airport transfers, and different tours. However, these words were not unique to the cities.

We later chose to use tf-idf vectorization because it is located more words that are unique to the cities.. See wordclouds below:

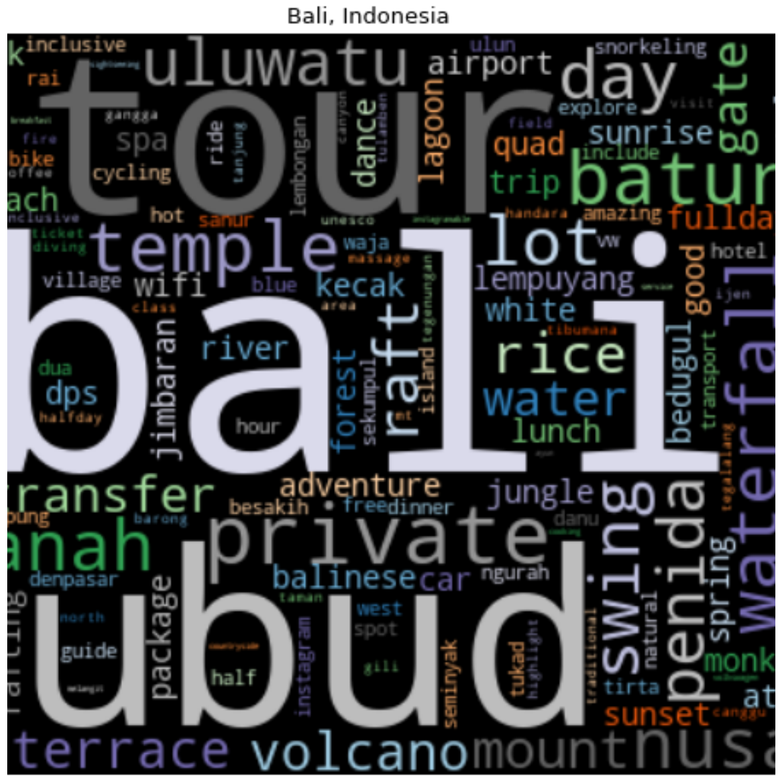
Dubai, United Arab Emirates

Majorca, Balearic Islands



Masked the word cloud in iconic features of the cities:

i.e Bali is quite scenic with palm trees



For Paris we masked the Eiffel Tower…etc

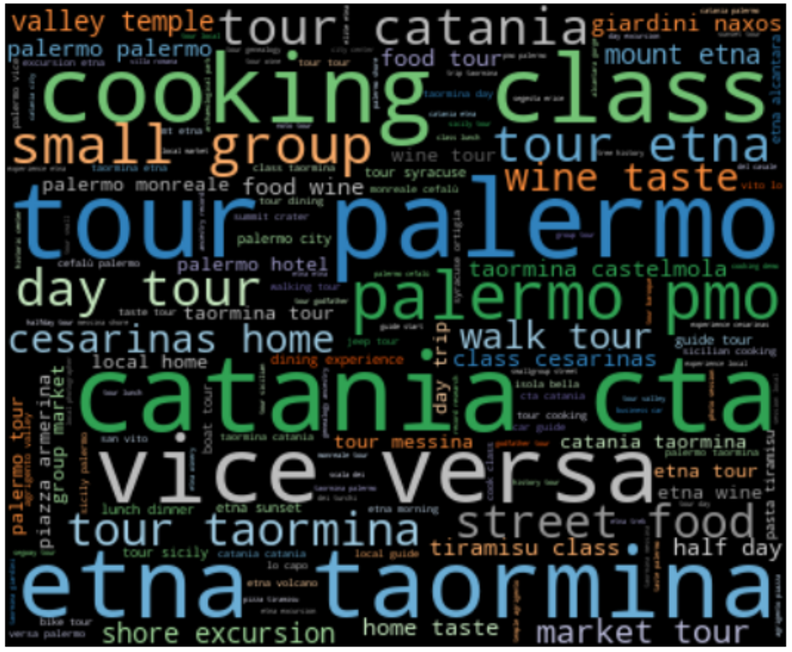
Paris, France

Paris, France

..other cities and their wordclouds

London, United Kingdom

Sicily, Italy



Rome, Italy

MODELLING

Our baseline model for building the recommendation system was the **Naive Bayes Model**

- Why: dealing with textual data and it involved frequency count for each city.

- The advantages of using this model include:
  -effective in processing meaningful patterns and preferences
  -well suited to handle multiclassification of the many potential destination text data

- Also employed the different models ie Count Vectorization, TF-IDF Vectorization & Bi-Grams to come up with the best model.

MODELLING

MultinomialNB

Count Vectorization

Training Accuracy: 0.8399
Testing Accuracy : 0.7981

MultinomialNB

Bi-Grams vectorization

Training Accuracy: 0.9020
Testing Accuracy : 0.7308

overfit

MultinomialNB

TF-IDF

Training Accuracy: 0.8552
Testing Accuracy : 0.8097

MultinomialNB

TF-IDF without Lemmatizing

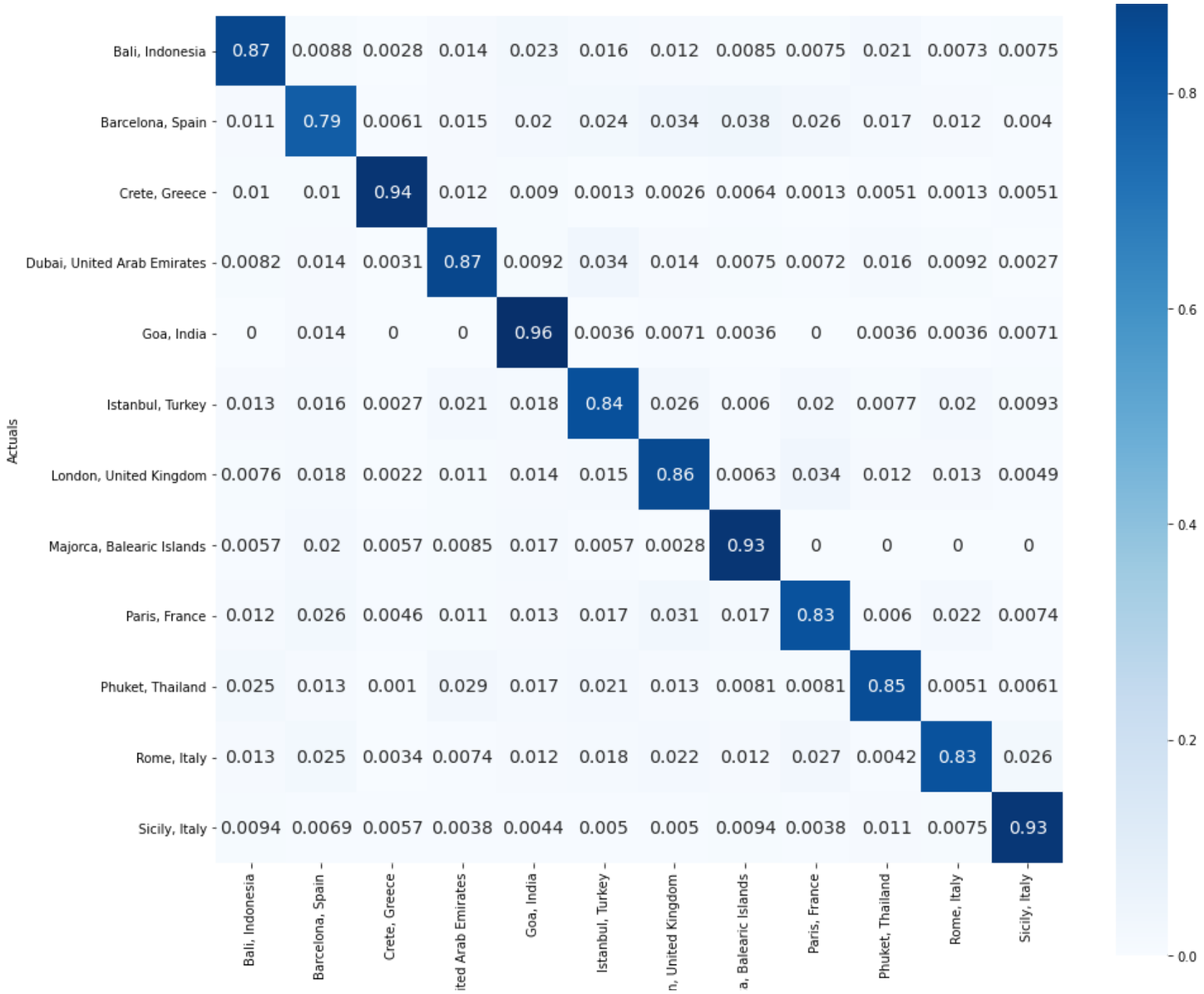
Training Accuracy: 0.8552
Testing Accuracy : 0.8097

Naive Bayes Model metrics

```
Training Accuracy: 0.8598863586310179
Testing Accuracy: 0.8146582100070472

---------------

Training F1: 0.8631488745947244
Testing F1: 0.8190645588523411

---------------
```

The final model had the following training and testing accuracy and F1 scores:
* Testing Accuracy Score 0.81 | F1 Score 0.82
* Training Accuracy Score 0.86 | F1 Score 0.86

F1 score is also important to consider since there is some class imbalance in the dataset and to account for the model's false positives and false negatives



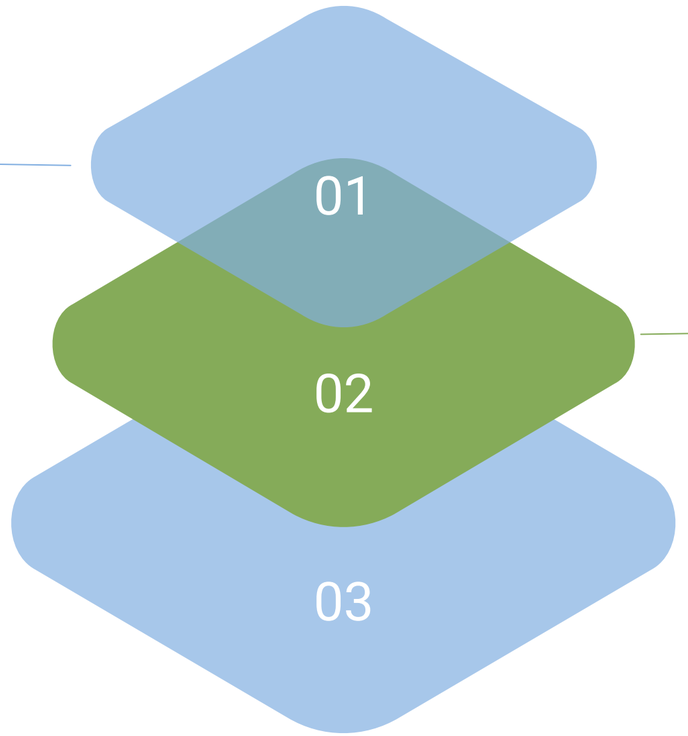
Model Deployment

Link:
https://destinationpredict-8pym.onrender.com

PROJECT RECOMMENDATIONS

Integrate the Destination Predict technology into pages where Top Destination lists are published to drive engagement with future travelers and drive traffic to affiliate links.

Offer paid sponsorship of the 'default' city-- example. Town of Bali can pay be the first recommended city when you open the page.



Use the Destination Predict model as an API to be used on travel websites to act as a virtual travel agent.

Given additional resources for exploration, we would delve into the following areas:

- Expand the model training to encompass a broader range of attraction sites and cities, to include live data from multiple travel advisories.

- Enhance the functionality and features of the Dash deployment app for an overall improvement in user experience.


THANK YOU
GLOBE TROTTERS,
GROUP 9