WONDER PAL

FINAL EVALUATIONS PRESENTATION

A DATA-DRIVEN VACATION LOCATION SUGGESTION SYSTEM

IN THIS PRESENTATION, WE'LL DELVE INTO THE PROJECT I HAVE WORKED ON.

NAME : A.N.L. ILLANGARATHNA

CO3554 - DATA MANAGEMENT PROJECT

INDEX NUMBER : 19/ENG/033

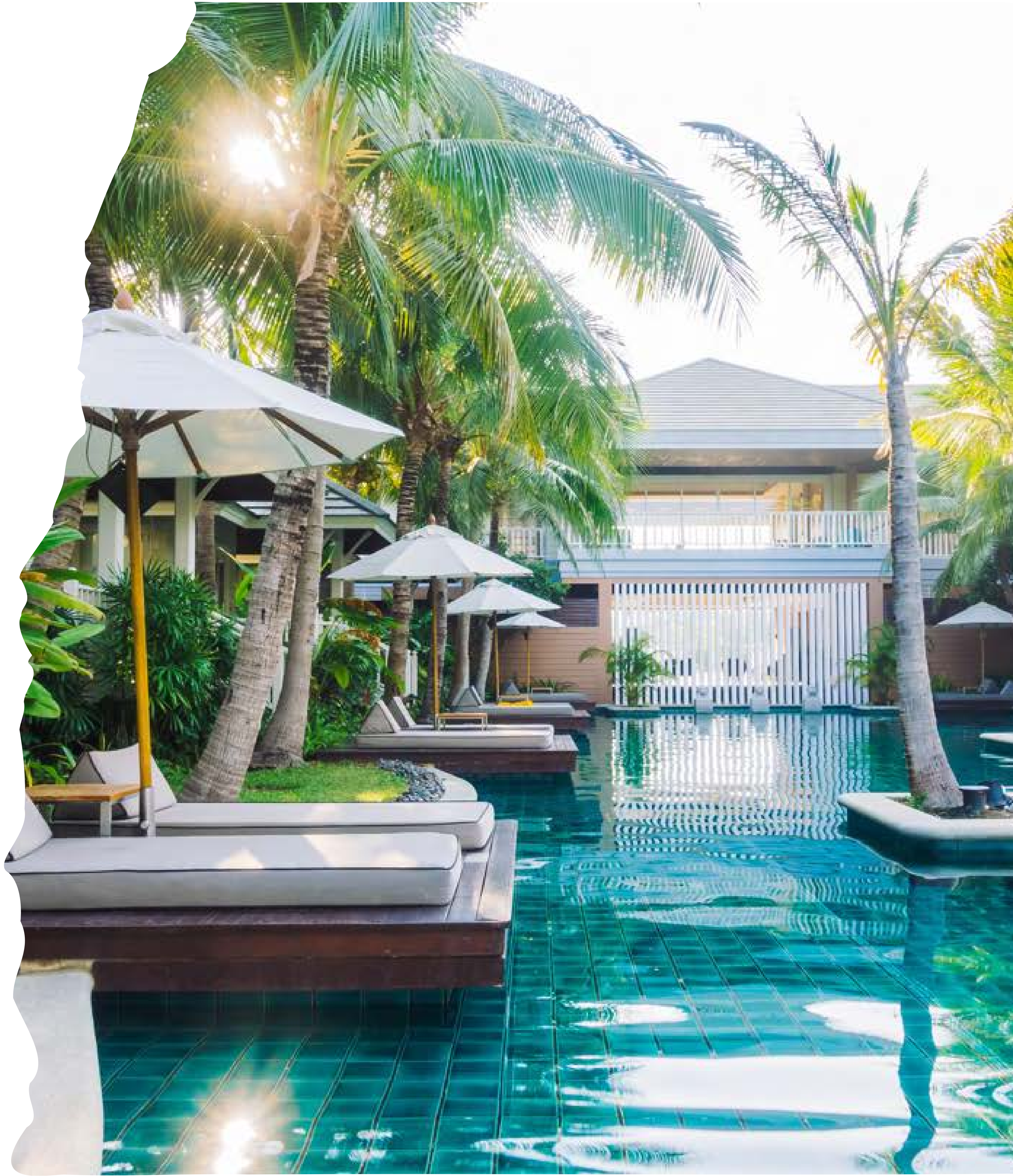
WONDER PAL

INTRODUCTION

We live in a very busy world. Everybody is in a run.
Stressful, Sleepless, BUSY society.
We all are having **VACATION** to escape from it.

BUT WHAT IF THAT VISITING LOCATION IS NOT AS YOU EXPECTED? 😢

My project aims to revolutionize the way people plan their vacations by harnessing the power of Data and Natural Language Processing.



DATASET

THIS DATASET HAS BEEN CREATED USING REAL DATA FROM BOOKING.COM.

LINK

https://www.openml.org/search?type=data&status=active&sort=runs&id=43712

IT IS A FREE TO USE DATASET FROM OPENML SITE

DATA PREPROCESSING

DATA PREPROCESSING INVOLVED ADDRESSING MISSING ADDRESSES, VALIDATING REVIEW DATES, IMPUTING AVERAGE SCORES, AND CLEANING REVIEW TEXT. THE CLEANED DATASET WAS SAVED FOR FURTHER ANALYSIS.

ADDRESS HANDLING

Filled empty addresses by finding similar hotel name entries with close longitude and latitude.

DATE VALIDATION

Ensured review dates were in the format yyyy-mm-dd and corrected discrepancies.

AVERAGE SCORE IMPUTATION

Filled empty average scores based on other reviews for the same hotel in the same year.

TEXT CLEANING

Removed unnecessary characters, extra spaces and redundant full stops from reviews.

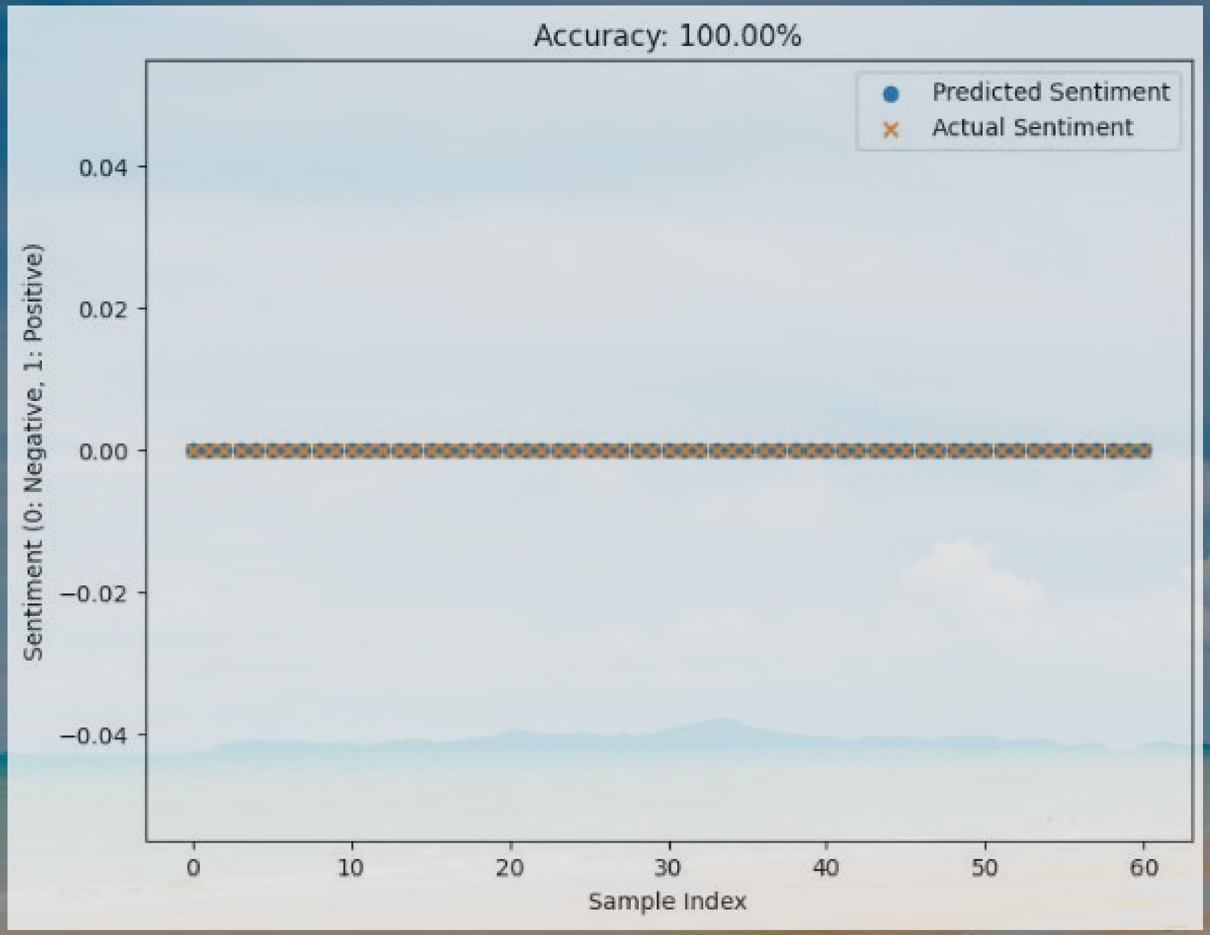
AUTOMATION CODE

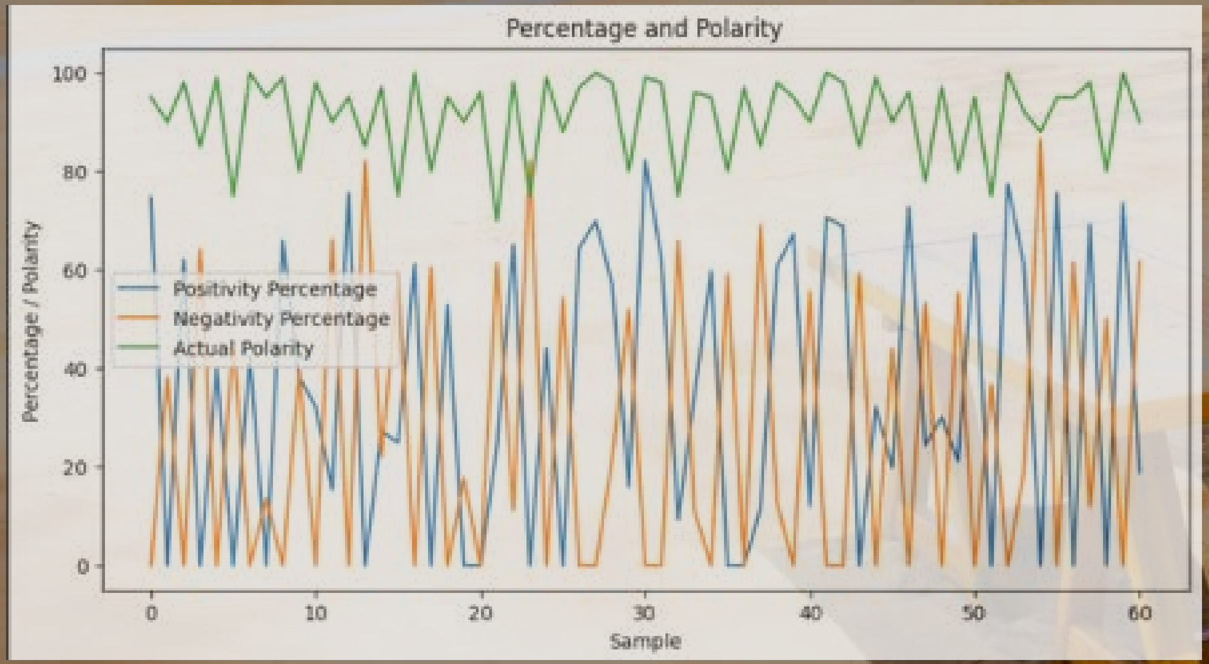
This is a massive dataset, housing over a million reviews, each with 100 to 500 words, the need for automation became clear.

- GPU processing (CUDA) - NLP-related tasks demanded significant time. Even with GPU acceleration.
- This automation, fully custom-created for my specific needs, was essential to manage the workload efficiently.

So Interruptions, Forceful stops and power failiers will not damage the progress

KEY FUNCTIONS	Managing data and ensuring data integrity.
STORAGE	Saving the last processed row on Google Drive for seamless resumption.
APPEND RESULTS	Adding batch results to the final CSV, preserving data.
WHILE LOOPS	Efficiently handling each batch of data.
TIME MEASUREMENT	Monitoring and optimizing processing speed.
OVERALL BENEFITS	Enhancing efficiency, reliability, and data integrity.





SENTIMENTS AND POLARITIES

Explored models for sentiment analysis, cleaned reviews, and used advanced tools to gauge sentiment and polarity.

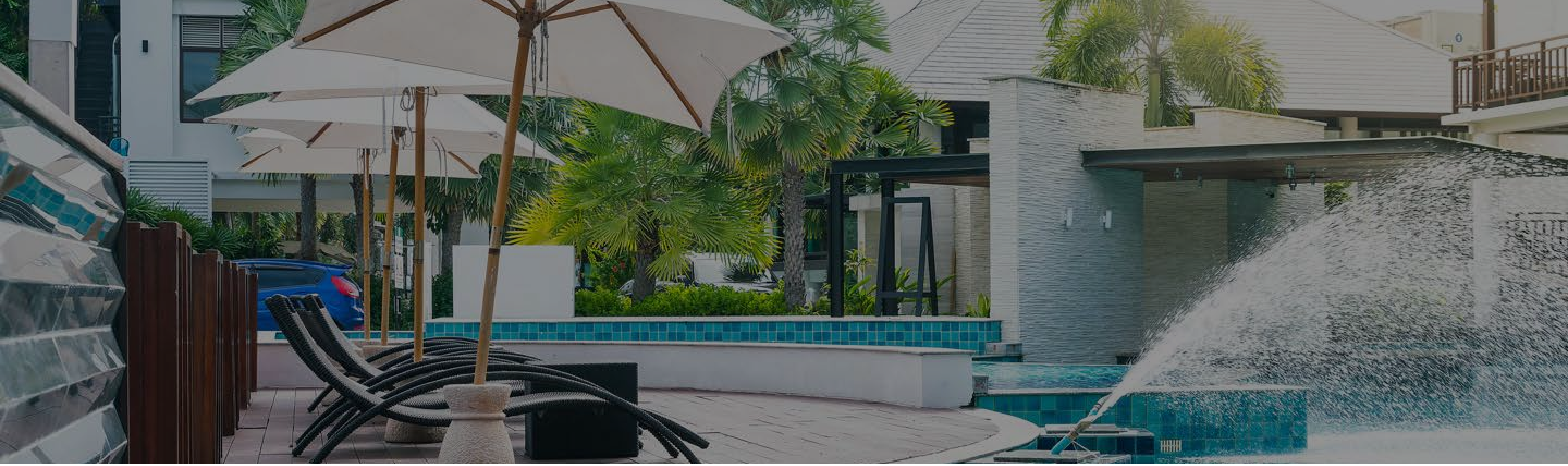
Finetuning approch

- **DISTILIBERT**, **BERT**, and **GPT2**.

Final Approch

- Implemented robust sentiment analysis using **spaCy** and **NLTK** libraries.

Used "**SentimentIntensityAnalyzer**" from **NLTK** for superior results since the accuracy measures are greater.



TEXT SUMMARIZATION

EFFICIENTLY SUMMARIZED TEXT USING T5 MODEL, HANDLING TOKEN LIMITS, AND SAVED PROCESSED DATA ON GOOGLE DRIVE.

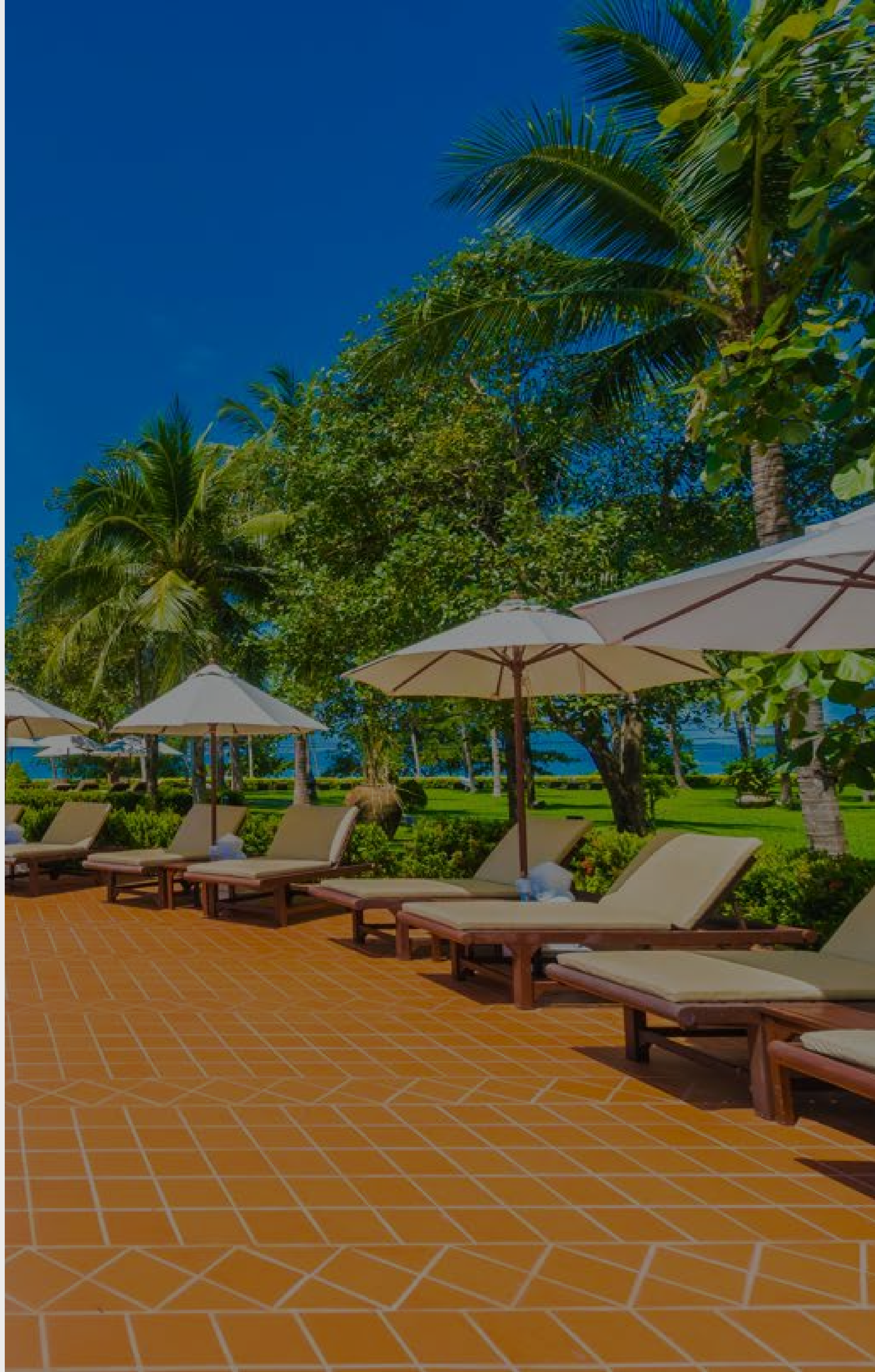
- Explored different summarization models: **BERT**, **DistilBERT**, **GPT-3**, **T5**.
- Then I tried **Transformers Summarization Pipeline** and it worked as well.
- **T5** model worked best.
- Used Hugging Face's Transformers library.
- Overcame T5's 512-token limit by **splitting long reviews**.
- Used **NLTK to break reviews into smaller parts**.
- Combined the parts to create full summaries.
- Preprocessed data saved as a CSV on Google Drive.

SUMMARY CLEANING

CLEANING SUMMARIES FOR CLARITY AND CONCISENESS, IMPROVING READABILITY WITH THE ALL-MINILM-L6-V2 MODEL.

- Cleaning summaries to enhance quality:
- **Removing redundant sentences** with similar meanings.
- Eliminating **unnecessary full stops**.
- Handling uppercase words for **readability improvement.**
- Utilizing the **All-MiniLM-L6-v2** model for cleaning.
- Highlighting the transformation of uncleaned summaries into more concise and coherent versions.
- Ensuring that cleaned summaries are easier to understand and interpret.

Summeries column	Cleaned summary
The staff were TERRIBLE UNHELPFUL SLOW AND WITH A COULDN t CARE LESS ATITUDE PARKING OF THE LACK OF IT WAS HORRENDOUS. Rooms were small. No wi fi in room when states on ad that it was standard size of room and ridiculous prices in bar.	The staff were terrible unhelpful slow and with a couldn t care less atitude parking of the lack of it was horrendous. Rooms were small. No wi fi in room when states on ad that it was standard size of room and ridiculous prices in bar.
. Very small. Very noisy air conditioning. Pool. The room. Location. The room. Wi Fi connection poor. Pool. The room. Location. The room. Location. The room. Location. The room. Location. The room. Location. The room. Location. The room. Location. The room. Location. The room. Location. The room. Location. The room. Location. The room. Location. The room. Location. The room.	Very small. Location. Pool. The room. Wi fi connection poor. Very noisy air conditioning.
a tsxi. a tsxi......... The wine...... The staff could smile... The staff could smile. The pool.. The staff could smile... Room view was of generator..	Room view was of generator. The wine. The pool. A tsxi. The staff could smile.
Poor service from staff Poor service from staff Emails not returned. No water hous. Poor service from staff Emails not returned. Poor service from staff Emails not returned. No water hous. Poor service from staff Emails not returned. Poor service from staff Emails not returned. Poor service from staff Emails not returned. Not much space in the rooms but expected that in central london.	Poor service from staff poor service from staff emails not returned. Not much space in the rooms but expected that in central london. No water hous. Poor service from staff emails not returned.
The staff were awful. The pool wifi was a bit too expensive The pool wifi was a bit too expensive. The pool wifi was a bit too expensive. Breakfast was not good.......... Bathrooms a bit small... Small bathroom. Small room. Small bathroom. Small bathroom. Small bathroom... Appalling customer service... .	Appalling customer service. The pool wifi was a bit too expensive the pool wifi was a bit too expensive. Breakfast was not good. Small room. Small bathroom. The staff were awful. Bathrooms a bit small. The pool wifi was a bit too expensive.
The staff were TERRIBLE UNHELPFUL SLOW AND WITH A COULDN t CARE LESS ATITUDE PARKING OF THE LACK OF IT WAS HORRENDOUS. Rooms were small. No wi fi in room when states on ad that it was standard size of room and ridiculous prices in bar.	No tea coffee or water in room. No water in room. No tea coffee coffee or water in room. No coffee. No tea coffee. No tea or coffee in room.
.... No tea coffee coffee or water in room. No tea coffee or water in room. No tea coffee or water in room. No tea coffee or water in room. No tea coffee or water in room. No tea coffee or water in room. No tea coffee or water in room. No tea coffee or water in room. No coffee. No water in room. No tea or coffee in room.. No tea coffee or water in room. No tea coffee.	Location. Bathroom very small. Bathroom is small. Bathroom was tiny. Not applicable. Bathroom to small shower was a trickle of water. Bathroom is small and dated.
.. Bathroom to small shower was a Trickle of water. Bathroom to small shower was a Trickle of water. Bathroom very small....... Bathroom is small and dated........... Bathroom was tiny. Not applicable.. Bathroom was tiny.. Location..... Bathroom is small.	The staff were terrible unhelpful slow and with a couldn t care less atitude parking of the lack of it was horrendous. Rooms were small. No wi fi in room when states on ad that it was standard size of room and ridiculous prices in bar.



NAMED ENTITY EXTRACTION

Generated labeled sentences using GPT-2 for fine-tuning a custom Event NER model, aimed for wider sharing.

LABELED DATASET CREATION

- Methodology for labeling events:
  - Utilized **GPT-2** model to generate sentences.
  - Created a dataset with 300,000 labeled sentences.
- Challenges faced during labeling:
  - Ensuring grammatically correct sentences.
  - Developing a random content generation mechanism
- This dataset is very good.
  - Planning To Publish It On Hugging Face In The Future under my account.

Sentence
I had a wonderful time trying sommelier guidance, Having a \n"Limbo" competition, Joining a local hiking, Playing billiards (pool). I was so excited to be able to do this. I had to go to the local store and buy a bottle of wine. It was a great experience.I was very impressed with the quality of the wine and the service. The wine was good and I would recommend it to anyone. We had the same experience with

Entities
{'the pool': 'ACTIVITY', 'sommelier guidance': 'ACTIVITY', 'Having a \\n"Limbo" competition': 'ACTIVITY', 'Joining a local hiking': 'ACTIVITY', 'Playing billiards (pool)': 'ACTIVITY'}

Sentence
I enjoyed every second of trying Joining a local book reading and discussion group, Doing a virtual art gallery tour. I also enjoyed the opportunity to meet some of the best artists in the world.I am a huge fan of all things art and I am looking forward to seeing you all again soon!.

Entities
{'discussion group': 'ACTIVITY', 'Joining a local book reading and discussion group': 'ACTIVITY', 'Doing a virtual art gallery tour': 'ACTIVITY'}

MODEL FINE-TUNING FOR NER

- Fine-tuning process:
  - Fine-tuned **BERT** model for Event NER.
- Fine-tuning BERT model for Event NER
  - accuracy and performance was acceptable.

NAMED ENTITIY EXTRACTION

- Models Used:
- Custom **Events NER Model**:
    - Fine-tuned specifically for extracting EVENTS from text.
- **Babelscape/wikineural-multilingual-ner:**
    - Extracts PERSONS, ORGANIZATIONS, and LOCATIONS.
- **Dizex/FoodBaseBERT:**
    - Specialized in identifying FOODS.

This system learns your preferences, such as your favorite foods, events, and places. Then, it suggests vacation spots that match what you like. It's like having a travel buddy who knows your tastes and recommends destinations just for you, making your trip planning a breeze.

EMBEDDINGS

Transforming user **input** and **review data** into **numerical embeddings**, the system enhances location suggestions. By comparing these embeddings for **cosine similarity**, the system provides **practical** and **accurate** recommendations, ensuring a personalized vacation experience.

EMBEDDINGS AND COSINE SIMILARITY

- Breaking each review into sentences to create embeddings.
- Proper storage and organization of sentence embeddings.
- Calculating **cosine similarities** between user input and stored embeddings.

PRACTICAL APPLICATION OF EMBEDDINGS

- Seamlessly **compare user input** with the vast **review** database.
- Discover matching locations, regardless of phrasing differences.
- Deliver personalized, accurate suggestions based on unique user input.
- Can give the feeling of this system understands the users inputs.

```
Similarity with MCQ 1: 0.8084
Similarity with MCQ 2: 0.7713
Similarity with MCQ 3: 0.6954
Similarity with MCQ 4: 0.3288
```

```
Most similar sentences based on combined similarity:
Combined Similarity with Sentence 9: 4.5851 - 'The beach offe
Combined Similarity with Sentence 4: 4.5162 - 'A vacation by
Combined Similarity with Sentence 8: 4.3884 - 'Spending leisu
Combined Similarity with Sentence 10: 4.3687 - 'In the embrac
Combined Similarity with Sentence 5: 4.1097 - 'The beach alwa
Combined Similarity with Sentence 1: 3.8641 - 'On a bright ar
Combined Similarity with Sentence 3: 3.8079 - 'Taking a break
Combined Similarity with Sentence 2: 3.6394 - 'The sound of t
Combined Similarity with Sentence 7: 3.0858 - 'The ocean wave
Combined Similarity with Sentence 6: 2.8143 - 'With a basket
Combined Similarity with Sentence 11: 2.2603 - 'vacation.The
Combined Similarity with Sentence 12: 1.4648 - 'beach, and th
```

MODEL CREATION

The recommendation system transforms user input and review data into numerical **embeddings**. Then consider the **Cosine Similarity** and then **Rank** them.

Utilizes Sentence Transformers to encode textual information.

All combined **16 APIs** are there based on these **16 functions**

```python
import ast

def get_recommendations(query_sentence, dataframe, num_recommendations=5):
    rank = 0
    # Load the Sentence Transformers model
    model = SentenceTransformer('jinaai/jina-embedding-t-en-v1')

    print(len(dataframe))

    try:
        dataframe['Positive_Review_Embeddings'] = dataframe['Positive_Review_Embeddings'].apply(ast.literal_eval)
    except Exception as e:
        # print(f"Error: {e}")
        problematic_rows = dataframe[dataframe['Positive_Review_Embeddings'].apply(lambda x: not isinstance(x, dict))]['Positive_Review_Embeddings']
        # print("Problematic rows:")
        # print(problematic_rows)

    # Extract necessary columns
    positive_reviews_data = dataframe[['Positive_Review', 'Positive_Review_Embeddings', 'Cleaned_Positive_Summary']]

    # Filter out rows with missing embeddings
```

GET DATA

1. GET_RECOMMENDATIONS
2. INITIAL_RECOMMENDATIONS
3. GET_HIGHEST_RANKED_RECOMMENDATIONS
4. GET_DATASET_SIZE
5. READ_FEEDBACK_LOOP_CSV

FILTERING

1. FILTER_ROWS
2. GET_ITEM_BY_RANK_FROM_RECOMMENDATIONS
3. GET_ATTRIBUTE_KEYS
4. GET_ATTRIBUTE_VALUE
5. GET_COLUMN_NAMES_FROM_ENTIRE_ROW
6. GET_COLUMN_VALUE_FROM_ENTIRE_ROW

UPDATE DATA

1. ADD_TO_PREVIOUS_PREFERENCES
2. GET_RECOMMENDATIONS_FROM_PREVIOUS
3. EDIT_COLUMN_VALUE

ADD DATA (FEEDBACK LOOP)

1. ADD_NEW_HOTEL_DATA_FEEDBACK_LOOP
2. READ_FEEDBACK_LOOP_CSV

PRESERVING

1. SAVE_MODEL

MODEL MANAGEMENT AND FEEDBACK LOOP

This slide outlines the **model management** and **feedback loop** functionalities integrated into the recommendation system.

FEEDBACK LOOP:

- Incorporated a feedback loop by saving user preferences to a CSV file.
- Implemented a mechanism to read and use feedback for future recommendations.

MODEL SAVE:

- Utilized **pickle** to save the model and functions for future use.
- Ensures **easy retrieval and deployment** of the recommendation system.

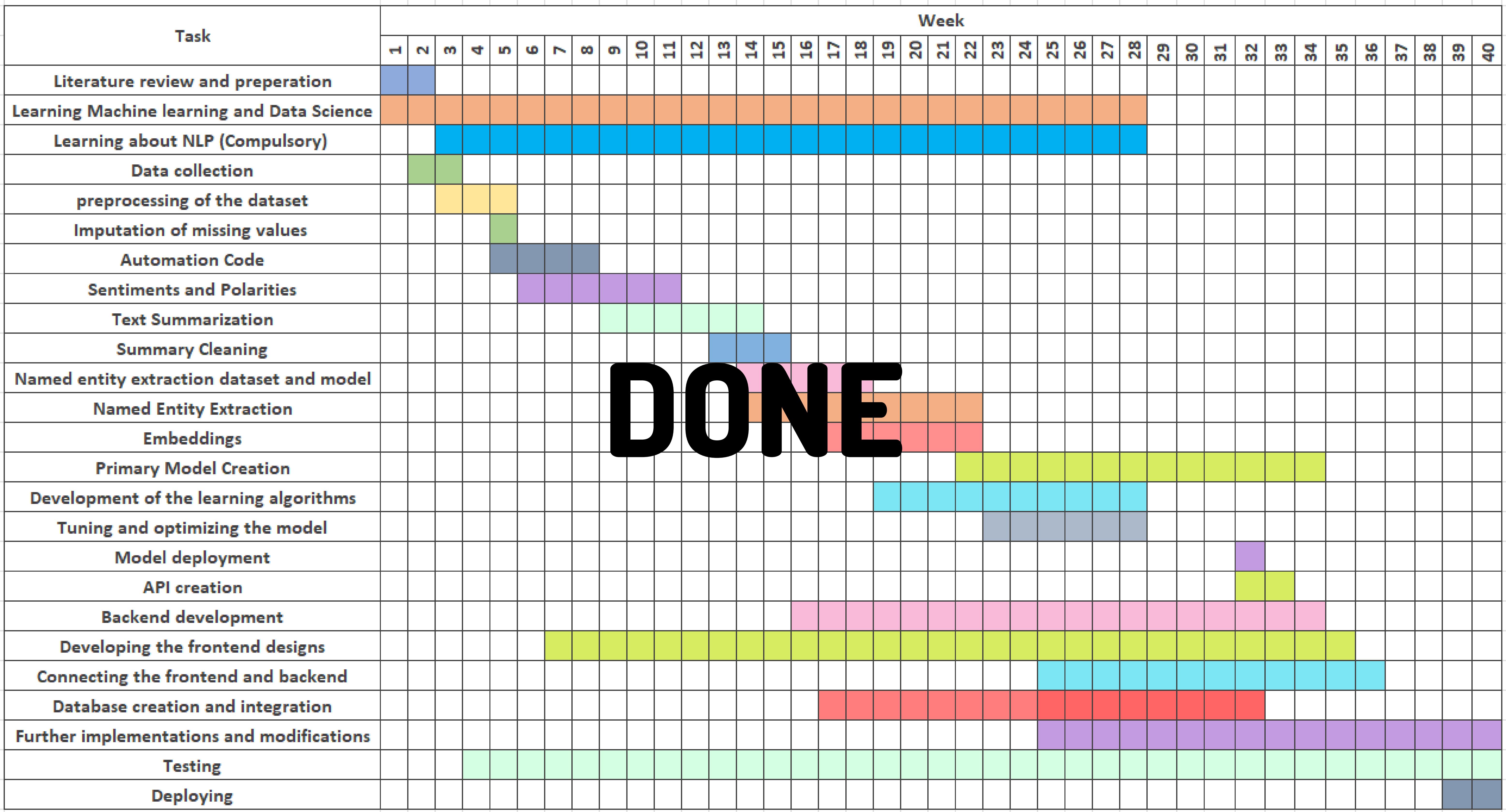
NEXT STEPS:

- Explore further enhancements, such as **collaborative filtering**.
- Continuously refine the model based on **user interactions and feedback.**

FRONTEND DEVELOPMENT

- Components
  - Splash screen
  - Helper screen
  - App Drawer
  - Homepage with creative designs, Input forms, Location cards.
  - suggestions
  - Recommandations
  - Visited places
  - Feedbacks
- Deployment Goal: Focus on deploying on Google Playstore with design excellence.

TIMELINE COMPLETED



REFERENCES

MACHINE LEARNING

[1] Kirill Eremenko, Hadelin de Ponteves, SuperDataScience Team, and SuperDataScience Support, “Machine Learning A-ZTM: Hands-On Python & R In Data Science,” Udemy, 2015. https://www.udemy.com/course/machinelearning/

OPEN ML

[2] “OpenML,” www.openml.org. https://www.openml.org/search?type=data&status=active&sort=qualities.NumberOfInstances&id=43712 (accessed Dec. 05, 2022)

HUGGING FACE

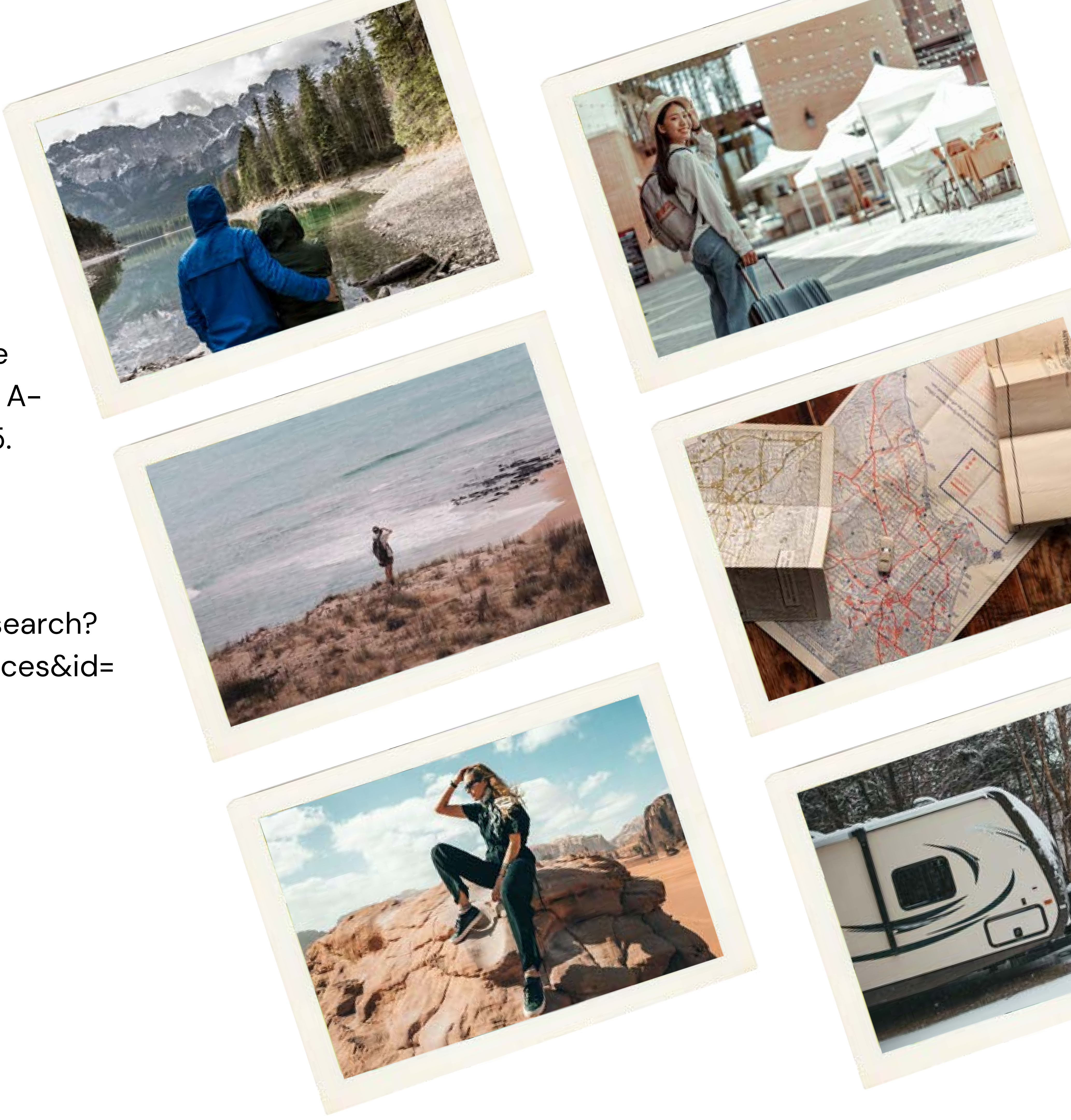
COLABORATORY

PYTHON

PYCHARM

FLUTTER

NODE.JS



WONDER PAL

THANK YOU

Q&A