Léo RINGEISSEN - Santiago MARTIN

Webscraping and Applied ML - Final Project

Link to GitHub:
https://github.com/SantiagoMartin2002/WebScrap_Project

Table of contents

- Issue and Objective
- Data Collection and Challenges
- Final Dataset
- Machine Learning Approach
- Results & Ensemble Model
- Application and Demo

Issue and Objective

Hypothesis

- **Problem**: Most long-distance travel options have a large carbon footprint
- **Solution**: Combining user preferences and ecological goals for greener travel planning

Objective

- Develop a system that recommends **eco-friendly travel itineraries** based on:
    - User’s **travel review** or description
    - **Carbon emissions** data
- Powered by:
    - **NLP algorithms** and Machine Learning
    - **Web scraping** and **API** querying

Data Collection and Challenges

Data Sources

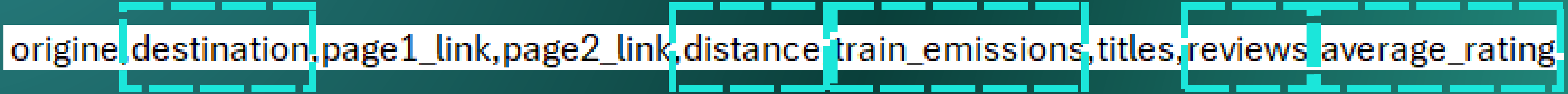
- **SNCF database API:** Carbon emissions data for train travel itineraries.
- **TripAdvisor**: Reviews of travel destinations.

Challenges

- **Web scraping difficulties:**
  - Inconsistent page scrolling and URL structure
  - Dynamic and inaccessible translations
  - Bot detection
  - Limited reviews
- **Adjustments:**
  - Focus on train trips from Paris --> more likely to have reviews
  - Use most iconic landmarks' reviews when destination reviews were missing
  - Manually map links to destinations' review pages

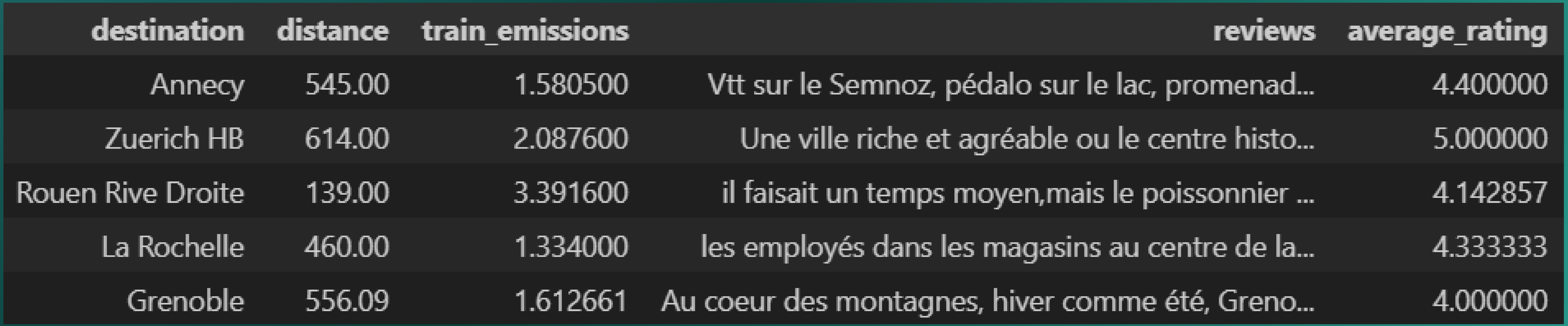
Final Dataset

Columns (ML highlight)



```
origine,destination,page1_link,page2_link,distance,train_emissions,titles,reviews,average_rating
```

Content (36 aggregated destinations from Paris)

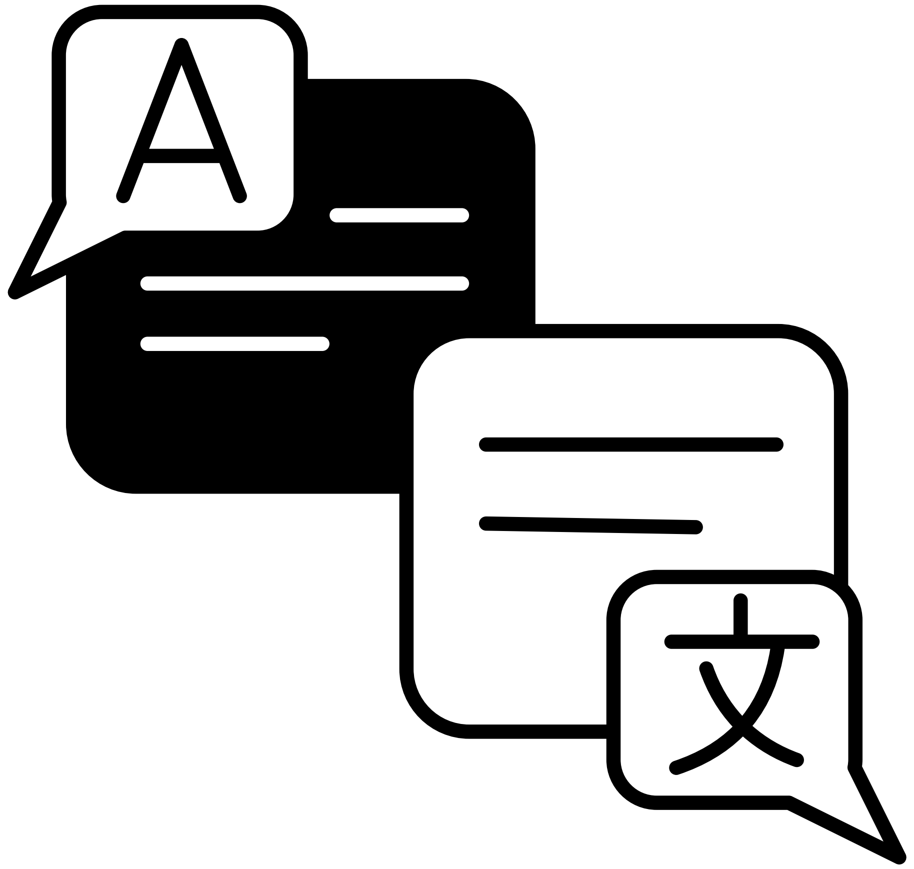


destination	distance	train_emissions	reviews	average_rating
Annecy	545.00	1.580500	Vtt sur le Semnoz, pédalo sur le lac, promenad...	4.400000
Zuerich HB	614.00	2.087600	Une ville riche et agréable ou le centre histo...	5.000000
Rouen Rive Droite	139.00	3.391600	il faisait un temps moyen,mais le poissonnier ...	4.142857
La Rochelle	460.00	1.334000	les employés dans les magasins au centre de la...	4.333333
Grenoble	556.09	1.612661	Au coeur des montagnes, hiver comme été, Greno...	4.000000

Machine Learning Approach

Strategy

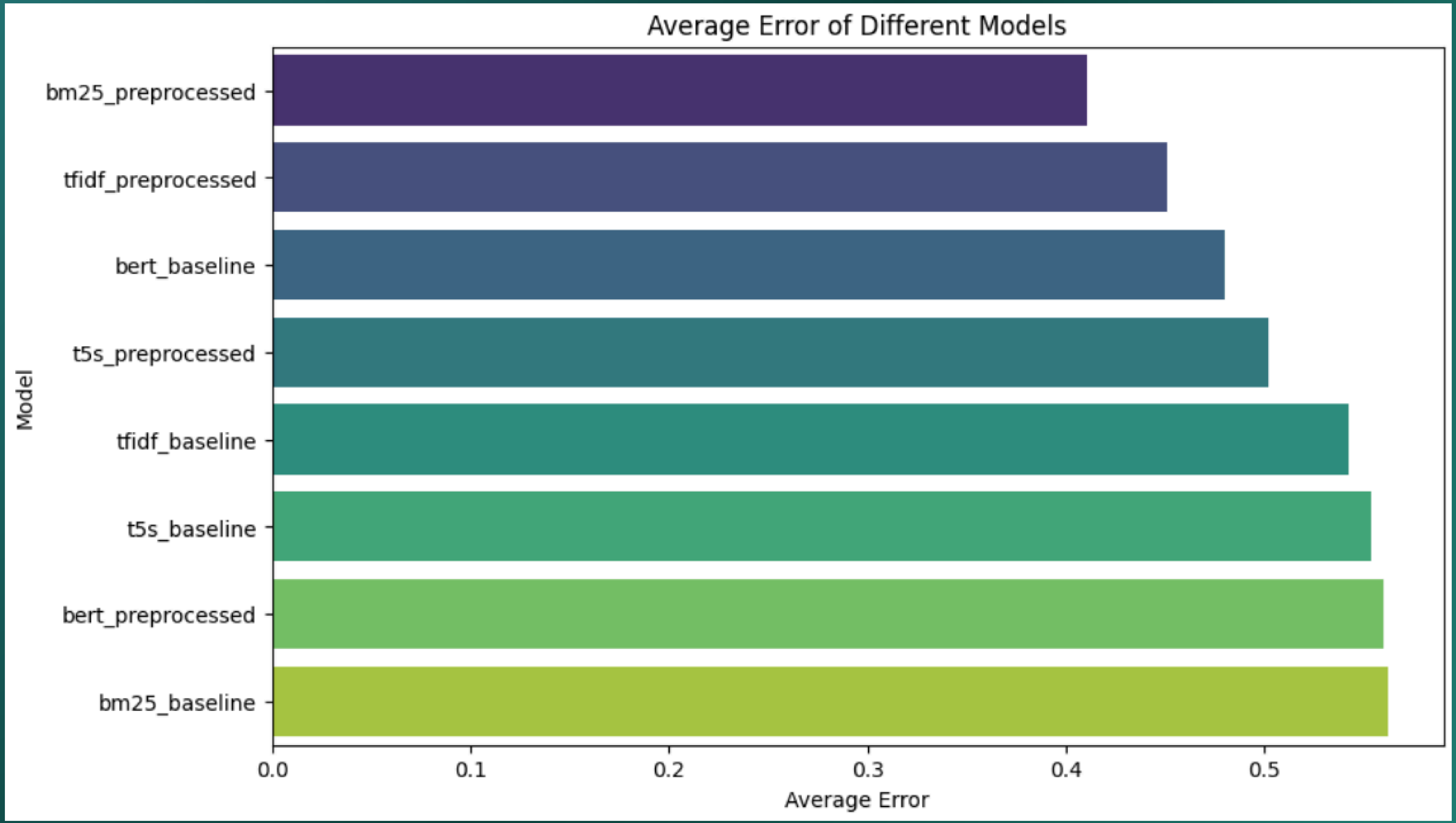
- **Information Retrieval Models:** BM25, TFIDF, T5 Flan, BERT.
- **Corpus:** Standard (raw reviews) vs. Preprocessed (NLP Techniques)
- **Evaluation:** Compare recommendations’ ratings to user query ratings.



NLP Preprocessing on reviews

- Make **all text lowercase**
- **Remove stopwords** with **NLTK**
- **Remove punctuation** and **other non-text characters** with **Regex**
- **Lemmatize** with **WordNetLemmatizer**
- **Apply techniques** to **each review** in the dataframe

Results & Ensemble Model

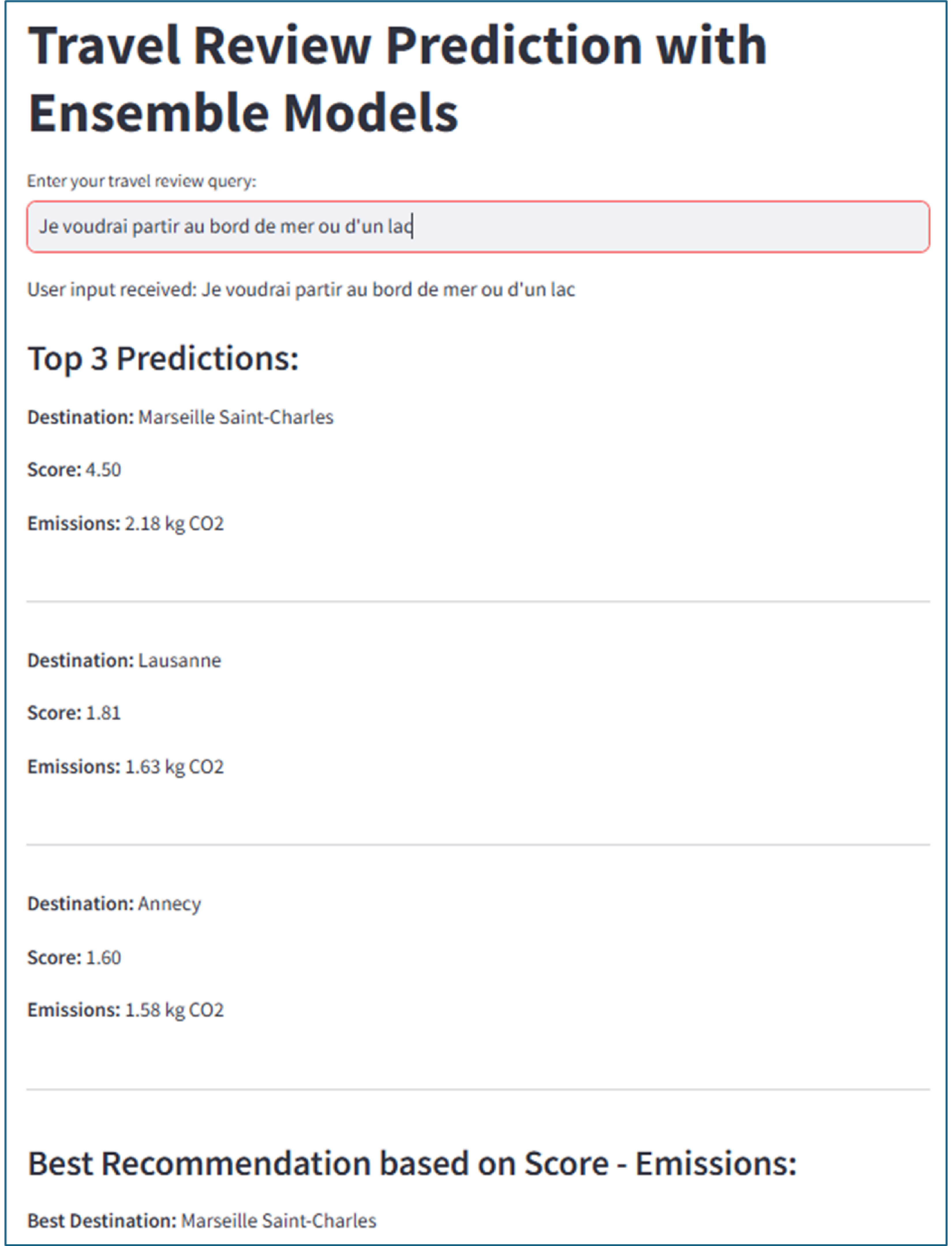


Performance is generally good, with errors trending towards 0.6 or less. Most models perform best when trained on preprocessed data. **Ensemble Model is built from best 3 models: BM25 and TFIDF on Preprocessed Data and Baseline BERT**

Application and Demonstration

Streamlit App

- Input trip review.
- Output: Top 3 recommended destinations with distances, emissions, and similarity scores.
- Highlights the most eco-friendly destination.



Travel Review Prediction with Ensemble Models

Enter your travel review query:

Je voudrai partir au bord de mer ou d'un lac

User input received: Je voudrai partir au bord de mer ou d'un lac

Top 3 Predictions:

Destination: Marseille Saint-Charles

Score: 4.50

Emissions: 2.18 kg CO2

Destination: Lausanne

Score: 1.81

Emissions: 1.63 kg CO2

Destination: Annecy

Score: 1.60

Emissions: 1.58 kg CO2

Best Recommendation based on Score - Emissions:

Best Destination: Marseille Saint-Charles