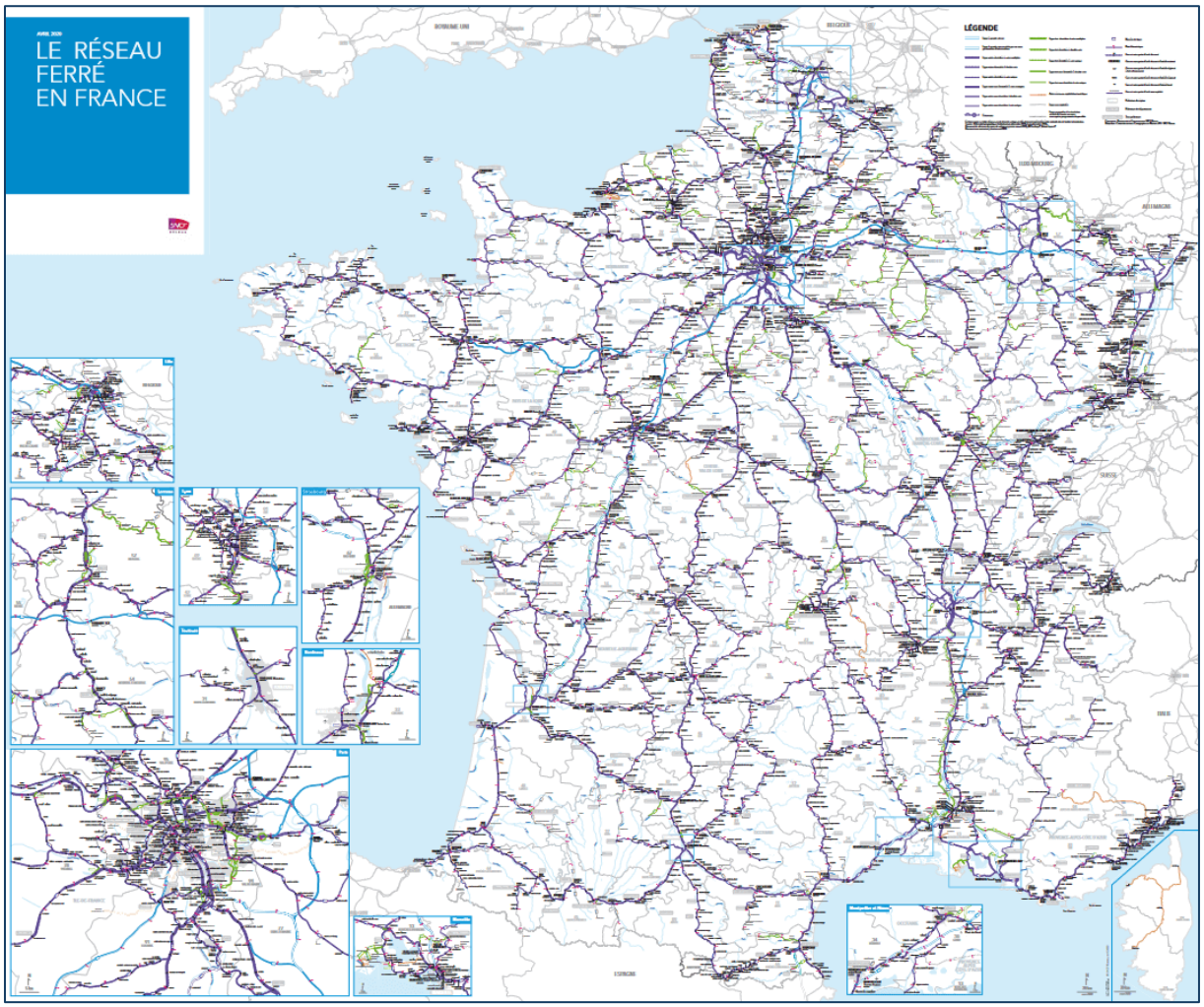
Web scraping Project Plan

Léo RINGEISSEN and Santiago MARTIN – DIA3



Issue

In today's more climate conscious culture, people are more aware of how their everyday actions play a role in the potential contamination of our planet. For these reasons, many of these people seek to live greener lives, driven by healthy decisions for the environment.

Unfortunately, when it comes to travel it's more difficult to be fully environmentally friendly with the options we have today. Most modes of transportation across long distances have an inevitably large carbon footprint.

It is our mission to help climate conscious travelers mitigate the contaminating effects of their travels, by facilitating the process of planning a fun trip that remains earth-friendly.

Objective

We want to develop a system that allows a user to enter a past trip they enjoyed, along with a brief description of that trip, and then makes a travel itinerary recommendation based on their input and minimizing the user's carbon footprint.

Method

We will use two sources of information to base our system on :

- An API with carbon emissions data on different travel itineraries provided by SNCF
- Web scraping reviews of travel destinations from TripAdvisor

We will use an NLP information retrieval model like BM25 to process the different reviews of the locations and return a destination recommendation based on matching reviews/descriptions. It will be trained on a corpus composed of the concatenated reviews of destinations we web scraped from TripAdvisor.

Next, the ranked recommendations returned by the model will be sorted based around the carbon footprint of said travel itinerary, which we attain the information for thanks to our API retrieval. This will provide the user with a travel itinerary that both matches their initial trip description/review and maintains a planet-friendly level of emissions.

Links to sources

The links to our sources are the following :

API : https://ressources.data.sncf.com/explore/dataset/emission-co2-perimetre-complet/information/

Example page to web scrap : https://www.tripadvisor.fr/ShowUserReviews-g187253-r86823297-Marseille_Bouches_du_Rhone_Provence_Alpes_Cote_d_Azur.html

First API calls and web scraps

See attached python notebook to view our first API call to the SNCF database and our first web scrap of reviews on TripAdvisor.