Data Science Take Home

Problem statement

Our engine is responsible for measuring our customers' emissions. The engine uses two inputs:

1. Item-level data that our customers upload to our platform. For example, item-level products that the company has purchased. The most important field is the *ITEM DESCRIPTION*, which is a free-form string that describes the item (e.g. a product).

2. Our proprietary, sector-specific databases of emission factors.

The engine's main goal is to find the most relevant emission factor for every item description in our customers' data. This is needed to accurately measure our customers' overall emissions, and also provide a detailed breakdown by emission scope, business facility, location, etc.

Prompt

We've sent you samples of item-level and emission factor data for the Food & Agriculture sector. These are "InterviewDataset.csv" and "EmissionFactorsData.csv" respectively. Using these two data sets, implement a program that matches item descriptions to the most relevant emission factors.

More details

- Notes on provided datasets:
  - "InterviewDataset.csv" has the *ITEM DESCRIPTION* field which you would be trying to match with as well as the *LABELS* field which serves as a check you may wish to use in your approach.
  - "EmissionFactorsData.csv" has an ITEM field you would be trying to match to and its corresponding EMISSION_FACTOR. Do note that the EMISSION_FACTORs provided are dummy data and should not be understood as representing real world emissions.

- Solving this problem in production, at scale is difficult. The intention behind this take home exercise is to see how you would approach building a quick solution to it (while learning new topics you may not be too familiar with), so please timebox your efforts to a few hours.

- We expect you to implement your proposed solution in Python (you're free to use Jupyter notebooks), and to send it to us as an email attachment or by sharing a

repo,
with instructions on how to run it. Feel free to use any open source package for the exercise.

- Please make sure to show all your work and also include some details about your approach, observations, ideas to improve it further, how you would productionize it, etc.

- **Bonus points:** Implement a simple API endpoint that accepts a list of item descriptions, and returns a list of the corresponding emission factors using your proposed approach.

Deliverables

1. Your code / Jupyter notebook
2. Item-level emission factor matches
    a. Should contain rows for every line-item with form (item_description, matched_emission_factor)
3. Written overview of approach, observations, future improvements and how you would productionize it