Junior Data Scientist Take-Home Task: Product Catalogue Creation

Objective

The objective of the take home task is to create a *master product catalogue* by merging together product-related information from different data sources. The main challenge is to identify and deduplicate identical products that have different naming conventions across the source datasets.

> As an illustrative example, one of the provided datasets has a product called *Amazon Simple Email Service* while other *Amazon Simple Email Service (Amazon SES)*.

Such a product catalogue serves as the foundation of a knowledge graph, which is critical for powering a recommender system that suggests relevant products to users based on their preferences and past behavior. Building this accurate and comprehensive catalogue is key to a functional recommender system.

Top tip: this particular process is often referred to as *entity resolution*.

Your task is to build a data pipeline that takes as an input *n* number of data sources and outputs a master catalogue of deduplicated products with information provided from the source datasets. All the necessary data will be provided. You are expected to present your process and the findings in a final interview stage to members of Dragonfly team.

Datasets Provided

We provide you two datasets in CSV format: *ts_technologies.csv* and *bd_technologies.csv* each containing information about software products. The datasets include various information about the products including *names*, *descriptions*, *URLs* and *product categories*.

Task Details

1. **Data Ingestion:** Load both CSV datasets (*ts_technologies.csv* and *bd_technologies.csv*).
2. **Data Exploration and Cleaning:** Examine the datasets for missing values, inconsistencies, and other data quality issues. Apply necessary cleaning steps.
3. **Product Deduplication:**
   - Develop a strategy to identify products that are the same but have different names or descriptions across the two datasets. This may involve fuzzy matching, text similarity measures, or other techniques.
   - Implement the deduplication logic to create a list of unique products.
   - Design your implementation so that it can be reused by others on more data.

4. **Master Catalogue Creation:** Generate a catalogue that contains the consolidated and deduplicated product information.
5. **Documentation:** Provide clear and concise documentation of the steps taken, including the deduplication strategy and any assumptions made.

Deliverables

1. A Github repository with relevant code such as a Python script or Jupyter Notebook that performs the data ingestion, cleaning, merging, and deduplication.
   - Bonus points for:
     - Setting up a reproducible pipeline and development environment.
2. The final master product catalogue in CSV format.
3. A brief document/README explaining the approach taken for product deduplication with your observations and any limitations.

Evaluation Criteria

- Accuracy of product deduplication.
- Code clarity and efficiency.
- Documentation quality.
- Ability to handle data quality issues.

You can use any tool of your choice, but you must be able to justify your choices. This applies to all aspects of the required work - methodology, choice of models/libraries and so on.

Expected Time

2-3 hours.