Group Project

Python For Data Analysis - Spring 2023

Due Date: May 1, 2023

Motivation

You work as a Data Analyst for a company that sells a Point of Sales (POS) system to small to medium businesses in the US. Your company has hundreds of thousands of clients and continues to grow. (Maybe you work for Square, maybe Shopify, maybe Stripe, etc…) One of the sales teams got their hands on a dataset of business names and locations and they were trying to identify the companies that are not on the platform, in other words prospective sales targets. You have been tasked with determining the overlap between the list of prospective businesses and the internal client list.

Problem Statement

You are given two datasets:

- left_dataset.csv
- right_dataset.csv

These datasets contain business names and their addresses.

The goal of this project is to find the businesses that are common to both datasets, that is, the businesses that have a **name and address** that match between the left and right datasets. Here is an example of a match. It is a nearly perfect match, ignoring case, since the name and address have identical values:

	left_dataset	**right_dataset**
id	47149	59483
name	Brothers Jewelry	BROTHERS JEWELRY
address	837 W Lancaster Ave Bryn Mawr PA, 19010	837 W Lancaster ave BRYN MAWR PA, 19010

These datasets come from different sources and so there may be some subtle differences between records, even if they do refer to the same business. For instance, in the example

below, the names and zip codes do not match exactly, but these entities clearly point to the same business:

	left_dataset	**right_dataset**
id	15883	11
name	Day's Collision Painting & Repair	Day's Collision
address	975 Florida Ave Palm Harbor FL, 34683	975 Florida Ave Palm Harbor FL, 34683-4224

Here are a few more examples of matches that are not exact:

	left_dataset	**right_dataset**
id	15925	2206
name	Jazz House Supper Club	Jazz House Supper Club LLC
address	9331 E Adamo Dr Tampa FL, 33619	9331 East Adamo Drive Tampa FL, 33619

	left_dataset	**right_dataset**
id	89855	72
name	Esposito's Italian	Esposito's 1948
address	14306 N Dale Mabry Hwy Tampa FL, 33618	14306 N Dale Mabry Hwy Ste F Tampa FL, 33618-2052

Since the business names and addresses don't align perfectly between these datasets, you will need to develop an algorithm that can find approximate matches. When your algorithm runs, it should produce a list of triplets consisting of the entity_id from the left dataset, the business_id from the right dataset, and a confidence score. The confidence score should have values between 0 and 1.0 and convey a sense of confidence of the match. An identical match should have a score of 1.0.

Your submission should consist of matches that have a high degree of confidence, eg greater than 0.8.

Here is a sample submission (as a csv file) for the examples shown above, where the confidence scores are just examples and do not come from an actual calculation.

```
left_dataset, right_dataset, confidence_score
47149, 59483, 1.0
15883, 11, 0.99
15925, 2206, 0.95
89855, 72, 0.91
```

The technical work that you are expected to conduct as part of this project is:

- conduct exploratory data analysis to understand the data, including visualizations
- develop an algorithm that can produces matches across both datasets
- assign a confidence score to your matches

Your algorithm should take no more than a few hours to run, ie you should not implement a brute force Order N^2 matching algorithm.

Deliverables

- Al zip file that contains the following:
    - your Python code: .py files and Jupyter notebooks
    - the result of your analysis: a text file or csv that contains the match IDs and confidence score
    - your presentation deck