***We recognize and appreciate the considerable time and effort required to complete the initial task. We will compensate the selected candidates for the time dedicated to this task.***

***Please note that this is a multi-line position, and we aim to hire all candidates who satisfy a certain accuracy threshold in their matching algorithm and demonstrate responsible and detail-oriented traits. You will not be competing with other candidates in the sense that we will not solely choose the candidate with the highest accuracy. Instead, we value and reward all individuals who meet our established criteria for quality and exhibit strong work ethics.***

**Objective**

You will find two datasets in the shared Google Drive folder. Your task is to to match the firms listed in the "Data1\_Sample.csv" dataset with their corresponding entries in the "Data2\_Sample.csv" dataset.

There are numerous GitHub repositories and open-source projects that provide code and algorithms for name matching and entity resolution tasks. Feel free to explore and utilize any of these resources to assist you in developing your matching approach. Some popular libraries and techniques include fuzzy string matching, Levenshtein distance, and machine learning-based methods. We encourage you to leverage existing tools and adapt them to the specific requirements of this task. We will consider the submissions that demonstrate a well-thought-out and effective matching strategy, regardless of whether they use existing GitHub code or custom-built solutions.

***While achieving a 50% accuracy rate in matching might be possible with minimal effort, our goal is to attain an accuracy rate of 85% or higher. Please keep in mind that an accuracy rate of around 50% would not provide sufficiently meaningful insights for our analysis. Therefore, we aim for a significantly higher accuracy rate to ensure the reliability and usefulness of the matched results. This requires careful attention to detail, thorough data preprocessing, and a well-designed matching algorithm. We appreciate your effort in keeping this high accuracy target in mind as you approach the task and strive for precision in your matching results.***

**Two datasets**

**Data1\_Sample.csv**

This dataset originates from Data Vendor 1 and features firms along with specific attributes collected from their profiles.

**Columns to use:**

* company: The name of the firm as it appears on Data1\_Sample.
* Please disregard remaining columns. They are necessary for our end, but would not help with your matching.

**Data2\_Sample.csv**

Sourced from Data Vendor 2, this dataset provides a comprehensive list of firms with a range of details pertaining to their operations and status.

**Columns to Use:**

* Company\_Name: The official name of the firm as recorded in Data2.
* Business\_Description: A brief overview of the firm's primary activities and market focus.
* Please disregard remaining columns. They are necessary for our end, but would not help with your matching.

**Example**

Consider the firm "Luminex Corp." as listed in the Data1\_Sample dataset:

* company: Luminex Corp.

In the Data2\_Sample dataset, the corresponding entry might be less straightforward and appear as:

* Company\_Name: Luminex Corporation (NasdaqGS:LMNX)

**Challenges and Cleaning Required:**

* Company Name Variation: The name appears as "Luminex Corp." in data1 and " Luminex Corporation (NasdaqGS:LMNX)" in data2. To match these effectively, you would need to standardize company names by perhaps removing common legal identifiers (e.g., "Corp.", "Corporation") and punctuations.

**Suggestions**

* Use firm names (“company” in Data1\_Sample and “Company\_Name” in Data2\_Sample) as your primary basis for matching. Given potential variations in naming (e.g., "Inc." vs. "Incorporated"), a degree of cleaning and standardization will be necessary.

**Evaluation Criteria**

Your matches will be evaluated based on the following criteria:

* **Match Coverage**: The vast majority (over 95%) of companies in the Data1\_Sample should have a corresponding match in the Data2\_Sample. For each company in the Data1\_Sample, you may find multiple potential matches in the Data2\_Sample. However, you should only include the single best match for each Data1\_Sample company in your final Excel file.
* **Match Accuracy**: Each match you identify will be assessed for its likelihood of being correct. Focus on precision and ensure that your matching approach minimizes false positives.

**Further instructions**

* In the absence of a ground truth for company matches, you will need to rely on your business intuition to evaluate the precision of your matching models and identify false positives. For example, 'Barnes & Noble Education, Inc.' from Data1 should be matched to 'Barnes & Noble Education, Inc. (NYSE:BNED)' in Data2. In Data2, you may notice that some company names are structured as 'Company Name (NYSE: XXX)'. This indicates that the company is publicly traded on the New York Stock Exchange (NYSE), one of the largest stock exchanges in the US, and 'XXX' represents its ticker symbol. Keep this in mind when matching company names and use your judgment to determine the best matches based on the available information.
* To continuously improve your matching code, we recommend the following iterative approach: After running your initial matching algorithm, identify the unmatched firms from Data1. Manually search for these firms in Data2 and analyze how you could modify your data cleaning and standardization process to capture these matches. One common technique is to compare the string similarity between two firm names and select the pair with the highest similarity score. By iteratively refining your data preprocessing and matching logic, you can improve the overall accuracy and coverage of your algorithm. For example, if 'Acme Inc.' from Data1 remains unmatched, you might find 'Acme Incorporated' in Data2. By calculating the string similarity between these two names and setting an appropriate threshold, you can determine if they should be considered a match. Continuously review and adjust your data cleaning, standardization, and matching techniques based on these findings to enhance the performance of your algorithm.

**Submission Requirements**

* Matched List in CSV: **Create a CSV file that has the same structure and number of observations as the Data1\_Sample.csv file. Include all columns from Data1\_Sample.csv in your submission.** For each company in Data1\_Sample, find the best-matched counterpart from the Data2\_Sample dataset and **include the corresponding "Excel\_Company\_ID" in a new column named "Matched\_Excel\_Company\_ID**". If no suitable match is found for a particular Data1\_Sample company, leave the "Matched\_Excel\_Company\_ID" field blank. The output CSV file should have exactly 1,252 rows, matching the number of observations in the Data1\_Sample.csv file.
* **Please ensure that the column name "Matched\_Excel\_Company\_ID" is exactly as specified, as any mismatch will lead to processing errors on our end. Our evaluation code relies on this specific naming convention**.
* **Python Code**: Provide the Python code used to perform the matching process. The code should be capable of generating the submitted csv file. Include comments in your code to explain your approach and make it understandable to others.
* **Please submit the code and matched list in csv file via Google Forms (https://forms.gle/Kgjfx67zKLFTkFZi9) by the end of this week (Feb 28th, 2025)**. Your application will be considered only if you complete this task.