**Methodology Description and Comments**

To achieve the goal of this test, I started by focusing on stopword removal using country-specific lists. I figured that the most common words appearing in company names are often generic terms like "Limited," "Company," or "Incorporated," which don't help in distinguishing one firm from another. So, for each country, I compiled a list of the top 10 most frequent words found in the firm names and treated them as stopwords. This way, I could remove these common words during the cleaning process, helping to highlight the unique parts of each company's name. Once the cleaning was done, I used fuzzy matching to assign IDs.

There were some assumptions and tradeoffs in this approach. I assumed that the most frequent words in company names are generic and can be safely removed without losing important information. However, this might not always be the case, and some meaningful words could be inadvertently removed, potentially leading to different firms being incorrectly matched. The 85% similarity threshold was chosen, but it might miss legitimate matches if the names are too different after cleaning, or incorrectly match different firms if their names are too similar

If I had more time, I'd consider refining the stopword lists by looking into the actual significance of frequent words, perhaps incorporating legal terms and linguistic nuances to improve accuracy.

Since new firms will be added each year, I thought about how to make the updating process more efficient without having to redo all the work each time. Instead of running everything from scratch whenever new firms are added, I would set up a system that focuses only on the new entries. When new firm names come in, I would clean them using the same method as before to maintain consistency. Then, I would compare these cleaned new names with the ones already in my dataset to see if there's a match. If a new firm's cleaned name matches one I already have, I would assign it the same unique ID, assuming it's the same firm. If there's no match, I would assign it a new ID. This approach allows me to update the dataset by processing only the new data

Alejandro Zimmerman