1. Introduction

In computer science, Fuzzy Matching is a technique that finds strings that could approximately match a pattern. In real operation, there are often needs to match external data (web page data, client data) with internal data (internal database) for some tactical purposes. In Natural Language Processing (NLP) field, however, external data are not always accurate. For example, data scraped from the website can be incorrect because of the typo or the unreliability of the webpage per se. It is important and necessary to develop an algorithm that could match two strings precisely when their slight differences are due to trivial error (typo, extra space between words, etc.,) This article will match the organization names (external data) with company names in Orbis, which is a powerful data resource with 400 million global companies. Specifically, since this research purpose is to match the entities that are unique to certain country, international organizations will be thrown out in the end.

2. Algorithm

2.1.Data Description

In our external scraped data, the most essential column is call ‘organizations’, and each rows contain a list, which includes all words and their corresponding frequencies. The Orbis dataset includes 5 columns, and the most important one is ‘name_internat’. This column shows the official names of all the registered organizations. For the testing purpose, names of Nigeria Organizations are utilized. The external data have 150000 rows, and the Orbis data contain more than 3800000 different organizations.

2.2.Algorithm Description

The entire algorithm includes data preprocessing, matching implementation, and discard of international organizations. There are five parameters for the input: the length of data needed in Orbis, the index of the row in the external data to start and end matching, one file for Orbis and the other one for the external data. Firstly, basic data cleaning process was conducted. Basically, extra spaces, punctuation, and parentheses were removed. In addition, to reduce the impact of irrelevant factors on matching, the suffix of the company name was removed by a build-in package within Python called cleanco, and the 'basename' module was used to cut the words such as Limited, LTD, PLC, Plc, etc., Meanwhile, we also converted all the cleaned data to lowercase, so as to eliminate the impact of the difference between uppercase and lowercase.

For the Fuzzy Matching index, this algorithm chooses to use the simple ratio since it is effective enough. This ratio can recognize missing punctuations, case-sensitive words, and misspelled word etc. The mechanism beyond this ratio is called Levenshtein distance, which measures the minimum number of edits required to change a particular string into some other string [1]. This research chose 95 as the threshold for this ratio because of the two reasons. Firstly, the threshold cannot be set too low, as this will cause substantial matching errors, which will affect the accuracy of the model. But the threshold cannot be set too high either, since it will miss some company names that should be consistent. After multiple attempts, it was found that the ratio between "NEW YORK MATS" and "NEW YORK MEATS" was 96. This case was quite inspiring and consistent with the original intention of designing this algorithm, therefore, the

threshold was set to 95. After selecting the matching entities, international entities were thrown out at the end.

3. Result

To test how the model works, a subset of the Nigeria data set, which includes 500 observations of the lists of organization names (external input) and 20000 different company names in Orbis. Table 1 shows the results of matching after throwing out the international organizations.

Table 1:
Output of the Matching

Index	Name_Organization_Original	Name_Nigeria_Original
1	Peoples Democratic Party	Peoples Democratic Party
2	Sterling Bank	Sterling Bank Plc
3	Sterling Bank Plc	Sterling Bank Plc
4	Zenith Bank	Zenith Bank Plc
5	Access Bank	Access Bank PLC
6	Nigerian National Petroleum Corporation	Nigerian National Petroleum Corporation (NNPC)
7	Dangote Industries	Dangote Industries (Ethiopia) PLC
8	Dangote Industries	Dangote Industries Limited
9	Dangote Industries Ltd	Dangote Industries (Ethiopia) PLC
10	Dangote Industries Ltd	Dangote Industries Limited
11	Union Dicon Salt Plc	Union Dicon Salt PLC

12	Medview Airline Plc	Medview Airline PLC
13	Apollo Tyres Ltd	Apollo Tyres (Nigeria) Limited
14	Guaranty Trust Bank	Guaranty Trust Bank Plc

It can be noticed that the overall accuracy of matching is almost 100% even though there are some results for parent and subsidiary companies. After applying to the entire dataset, 1355 pairs of matching are found, and the overall accuracy is over 98%.

4. Discussion

This article mainly concerns about matching company names by an improved Fuzzy Matching algorithm. As the result showed by Nigeria dataset, it achieves extremely high accuracy. However, there are mainly two potential limitations to this algorithm. Firstly, this algorithm is very time-consuming. It took nearly 20 minutes to match these 500 company names across 20,000 pieces of data. If hundreds of thousands of data are applied, it may take plenty of hours to finish matching. In addition, the threshold may vary. Even if 95 works well in this project, it will still need to be adjusted based on different sources of data.

5. Reference

[1] Brian Young, Tom Faris, Luigi Armogida, Levenshtein distance as a measure of accuracy and precision in forensic PCR-MPS methods, Forensic Science International: Genetics, Volume 55, 2021.