2021 Big Data Final Project - Group 11

December 12, 2021

- Chun-Yen Liou, cyl625
- Tsung-Lin Yang, ty2065
- Peng-Yuan Chen, pc2973

- github repo link: https://github.com/TsungLin716/2021-Big-Data-Final-Project.git

1 Introduction

In the work of exploring enormous sets of geo-spatial data, one can often be hampered by the poor quality of name related data. These kinds of data mainly include place names (countries, cities, etc.) and street names. The problems typically include inconsistency representations of the same object, inconsistency abbreviations of the same vocabulary, misspelled names, etc.

These problems pose a devastating threat to the quality of the result of the data interpretation. The main reason is that a dataset may contain tons of different representations that actually point to the same city or street. This can greatly affect the accuracy of the statistical analysis result of the data.

In this report, we propose two improved data cleaning methods based on the openclean library for Python. One method improves the cleaning result of street names and the other improves the cleaning result of city names.

To measure the effectiveness of the proposed method, we employ precision and recall to evaluate the results. We followed the calculations of the two metrics in the study conducted by Chu et al., which are implemented in evaluating the effectiveness of their data cleaning system[Xu Chu and Ye.(2015)].

2 Problem Formulation

2.1 Street Names

Street name is one of the data types that suffer most from the problem of inconsistency. In our preliminary study of the geo-spatial data of DOB Job Application Filings[OpenData(2021)] from NYC OpenData, we found two types of problem in street names. The first one is the different representation of some common vocabularies or numbers in street names, such as avenue or ave and first or 1st. The second is the different arrangements of the words in the same street name. An example is presented in the table below.

Problems	Examples
1. Different representation of some common vocabularies or numbers.	[CROSS BRONX EXPRESSWAY, CROSS BRONX EXPY] [1ST AVENUE, FIRST AVE, 1TH AVENUE]
2. Different arrangements of the words in the same street name.	[DRUMGOOLE RD WEST, WEST DRUMGOOLE RD] [EAST CLARKE PLACE, CLARKE PLACE EAST]

Openclean library does a pretty outstanding job in standardizing street names except for a minor flaw. We observed that sometimes street names with unnecessary symbols such as comma or period are presented in the cleaned result. For example, we can see that after the cleaning, "WEST 111 ST" and "WEST 111 ST ." both exist in the dataset. As a result, some street names are still having multiple representations. This is an issue we want to cope with.

2.2 City Names

City name is another data type that not only has inconsistent format but misspelling problems. In our study from part1, we found two types of issues in city names from DOB Job Application Filings (NYC Open Data). Some of them are misspelled and some of them are mixed with lowercase and uppercase. The example are shown below:

Problems	Examples
1. Misspelled City Name	[BRKLYN, BROKKLYN] [LONG ISLAND CIT] [ENGLEWODO CFLII]
2. Lowercase mixed with uppercase	[BROOKLYN,Brooklyn] [New York, NEW YORK]

There is a library in Openclean that can compare the pronunciation of each word to see if it is correct or not. We define the correct city name and start doing the cleaning. However, there are some cases that can not be addressed. We observed that when the city name is severely scrambled, the library can not easily detect its pronunciation so that the error remains. For example, "BKRLYOON" and "BLYRKOON". Furthermore, the library can not solve the city name which is in abbreviated form. For example, "BK", "NY" and "LIC". These are the issues that we want to overcome in the improvement process.

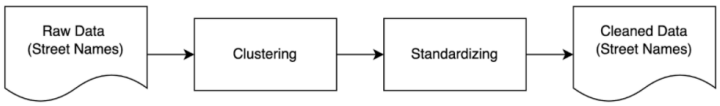
3 Related Work

Several studies were presented in the past aiming to solve the problems of poor quality of city names and street names. Lim proposed an algorithm known as "Weighted Matching Ratio" to better clean the city names in a dataset[Lim(2010)]. The algorithm is based on the Longest Common Subsequence algorithm and does not require human supervision to fix city names in a dataset. Dr. M. Ben Swarup and B. Leela Priyanka proposed an approach that using LCS algorithm for city name correction[Swarup and PriyankaAutomatic(2014)], their approach achieves a precision of 90% which is better than the traditional Levenshtein distance. Karen Kukich proposed approaches based on three problems at correcting words. The first one is nonword error detection and the second one is isolated-word error correction, the third one is context-dependent word correction. She proposed three methods which are pattern-matching and n-gram analysis techniques, general and application-specific spellingcorrection techniques and some experiments using natural-language-processing tools or statistical-language models.[Kukich(1992)]

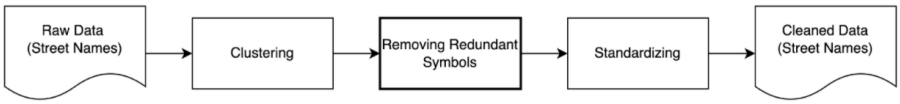
4 Method

4.1 Street Names

Originally, we implemented the clustering and standardizing function in openclean library to clean and fix the street names in our datasets. The process is illustrated in the image below.



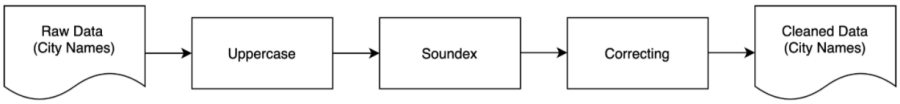
To address the issue of redundant symbols such as period and comma in the cleaned result of openclean library, we simply replace those symbols by white space after clustering and before standardizing the street names. The process is illustrated in the image below. As can be seen, an additional is added to the process.



With the additional step, the street names can be identified without the interference of the symbols. Therefore, more consistent results can be obtained.

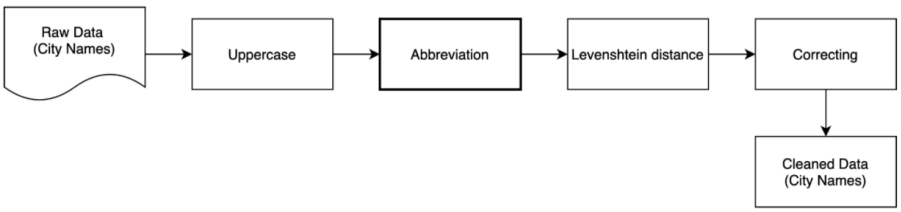
4.2 City Names

Originally, we first transformed all city names to uppercase for consistency. And we used openclean library soundex(), which is a phonetic algorithm for indexing names by sound to determine if the city name needed to be corrected. However, there are some extremely misspelled names such as BKROYLN which is supposed to be BROOKLYN that soundex cannot detect. The process is illustrated in the image below.



To handle this problem, we used Levenshtein distance to search for the minimum number of single character edits required to change in order to match the misspelled city name to the correct city name. Therefore, this algorithm can solve the extremely misspelled names problem.

In addition, we also found out that there are some abbreviations such as L.I.C., BK and NY. which are supposed to be LONG ISLAND CITY, BROOKLYN and NEW YORK. But Levenshtien distance algorithm cannot perfectly match these cases to the correct city names. These cases will just match the city names that have the least characters such as BRONX. As a result, we need to implement another function to deal with these special cases. The modified process is shown below.



With the improved method, city names can be identified correctly. And more consistent results can be obtained.

5 Result

To evaluate the effectiveness of the proposed method, we conducted a test with a sample dataset. The sample dataset was collected from 10 different datasets in NYC OpenData. The number of records collected from each dataset was determined by the portion of the total number of records the dataset takes up. Detailed information can be found in the table below.

Dataset Name	Amount of Records	Amount of Collected Samples
data-cityofnewyork-us.iz2q-9x8d.csv	6090	1
data-cityofnewyork-us.nyis-y4yr.csv	1530000	68
data-cityofnewyork-us.ipu4-2q9a.csv	3780000	174
data-cityofnewyork-us.w9ak-ipjd.csv	241000	12
data-cityofnewyork-us.bty7-2jhb.csv	2430000	113
data-cityofnewyork-us.dm9a-ab7w.csv	255000	12
data-cityofnewyork-us.kyvb-rbwd.csv	347000	17
data-cityofnewyork-us.kfp4-dz4h.csv	45700	3
data-cityofnewyork-us.8fei-z6rz.csv	27900	2
data-cityofnewyork-us.uxsz-6j5j.csv	298	1
Total number	8662988	403

We want to compare the proposed cleaning methods and the original methods in the openclean library. Thus, two cleaned dataset that only contain the city names and street names was produced. The first one is named as "unimproved_result", which was produced merely with the built-in cleaning function of the openclean library. The second one is named as "improved_result", which was produced through the proposed methods in this study. Also, a ground truth dataset was produced. This is done by manually inspecting and modifying the city names and street names in the sample dataset.

For the "unimproved_result" and "improved_result" dataset, we respectively calculated their precision and recall upon city names and street names correction. The precision P was calculated as following:

$$P = \frac{records\ correctly\ fixed\ by\ program}{all\ records\ fixed\ by\ program} \quad (1)$$

Where "records correctly changed by program" is the number of records fixed by the improved or unimproved program that match the same records changed in ground truth dataset. And all records fixed by program is the number of records fixed by the improved or unimproved program regardless of the correctness of the fix. On the other hand, the recall R was calculated as following:

$$R = \frac{records\ correctly\ fixed\ by\ program}{records\ fixed\ manually} \quad (2)$$

Where "records fixed manually" is the number of records fixed by the manual inspection (also known as the ground truth).

5.1 Street Names

The precision and recall of the street names cleaning results done by the unimproved method and the improved method are shown in the table below. As shown, the improved method produced better results than the unimproved.

	Unimproved Method	Improved Method
Precision	0.917	0.959
Recall	0.928	0.970

The main reason for such improvement is that more records were modified to the correct forms that match with the ground truth. By manually inspecting the result, we found that the street names with redundant symbols were fixed to correct forms that do not contain symbols. Also, we observed that the records modified by the improved method are the same records modified by the unimproved method. This implies that our modification of the data cleaning strategy does not affect the decision made by the original method of what record should be fixed. In addition, the decision making strategy of the original method is adequately great. Therefore, we can conclude that, since the improved method decided to modify the same set of records, and did a better job in modifying those records, we obtained better precision and recall.

5.2 City Names

The precision and recall of the street names cleaning results done by the unimproved method and the improved method are shown in the table below. As shown, the improved method produced better results than the unimproved.

	Unimproved Method	Improved Method
Precision	1.00	1.00
Recall	0.189	1.00

As we can see from the result, the improvement is based on the fact that more data are corrected compared to the ground truth. When browsing through the result, we found out that when comparing to the result from the unimproved method, all the misspelled data was corrected which means that our algorithm[Wikipedia(2021)] works functionally and corrects the data which has the scrambled pronunciation. Furthermore, we observed that all the abbreviated city name were changed to the correct form. By such, we can conclude that we oversimplified the problems within city name while other issues remains. After the improvement, our new model does show how robust it is by looking at the new precision and recall rate.

References

[Kukich(1992)] Karen Kukich. 1992. Techniques for automatically correcting words in text. *ACM Comput Surv. 24, 4* (Dec. 1992). https://doi.org/10.1145/146370.146380

[Lim(2010)] SeungJin Lim. 2010. Cleansing Noisy City Names in Spatial Data Mining. *International Conference on Information Science and Applications. Seoul, Korea, 1-8* (2010). https://doi.org/10.1109/ICISA.2010.5480390

[OpenData(2021)] NYC OpenData. 2021. 2021. DOB Job Application Filings. (Dec. 2021). https://data.cityofnewyork.us/Housing-Development/DOB-Job-Application-Filings/ic3t-wcy2/

[Swarup and PriyankaAutomatic(2014)] M. Ben Swarup and B. Leela PriyankaAutomatic. 2014. Data Cleansing Of Incorrect City Names In Spatial Databases Using LCS Algorithm. *International Journal of Advanced Research in Computer Engineering Technology. Volume 3. Issue 4* (April 2014). http://ijarcet.org/wp-content/uploads/IJARCET-VOL-3-ISSUE-4-1393-1396.pdf

[Wikipedia(2021)] Wikipedia. 2021. Levenshtein distance. (Dec. 2021). https://en.wikipedia.org/wiki/Levenshtein$_d$*istance*/

[Xu Chu and Ye.(2015)] Ihab F. Ilyas Mourad Ouzzani Paolo Papotti Nan Tang Xu Chu, John Morcos and Yin Ye. 2015. KATARA: A Data Cleaning System Powered by Knowledge Bases and Crowdsourcing. *2015 ACM SIGMOD International Conference on Management of Data (SIGMOD '15). Association for Computing Machinery, New York, NY, USA, 1247–1261.* (2015). https://doi.org/10.1145/2723372.2749431