MDSAA

Master's Degree Program in

Data Science and Advanced Analytics

Automatic Spelling Corrector to improve Unified Registry analysis for Brazilian social development

Adjustment of automatic spelling corrector system applied in Brazilian low-income family's data

Hiromi Nakashima

Dissertation

presented as a partial requirement for obtaining the master's Degree Program in Data Science and Advanced Analytics

NOVA Information Management School

2022

Title: Automatic Spelling Corrector to improve Unified Registry analysis for Brazilian social development

Subtitle: Adjustment of automatic spelling corrector system applied in Brazilian low-income family's data

HIROMI NAKASHIMA

MDSAA

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Automatic Spelling Corrector to improve Unified Registry analysis for Brazilian social development

by

Hiromi Nakashima

Dissertation

Master's degree in Advanced Analytics, with a Specialization in Data Science

Supervisor /Co-Supervisor: Mauro Castelli

November 2022

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Hiromi Nakashima

Lisboa, November 8th of 2022

ACKNOWLEDGMENTS

Throughout the writing of this dissertation, I have received a great deal of support and assistance.

I would like to thank Caio Nakashima from the Ministry of Citizenship (“*Ministério da Cidadania*”) and Professor of the Technologic University of Paraná, who provided the challenge and problem to create this research. His insights and feedback support helped me sharpen my thoughts and look up original solutions. I also would like to thank Professor Robinson Vida Noronha from the Technologic University of Paraná, who gave great advice that led me to more professional work and help me to successfully finish this dissertation.

I would like to acknowledge University NOVA – IMS, especially Professor Mauro Castelli, who gave me this opportunity to reach this milestone.

Finally, I would like my mother, Iuri Arimura, who gave me the support to deliver this work and, my Portuguese family.

ABSTRACT

This dissertation has the goal to develop a solution to correct spelling errors when inserting the neighborhoods of Brazilian low-income families. The Brazil Government uses data from the Unified Registry (Cadastro Único) to diagnose the basic social right of low-income families and to map public policies based on the real needs of Brazilian society. Therefore, the best solution found was an adjustment of the string correction method, Automatic Spelling Correction (ASC) system, and an automatic dictionary creator, to the CECAD registry family status data and correction of the neighborhood names wrongly typed. The research will describe the algorithm's process with an explanation of the main mathematical concepts.

KEYWORDS

String Correction, Automatic Spelling Correction, ASC, Cadastro Único, CECAD.

Sustainable Development Goals (SGD):

1 NO POVERTY

2 ZERO HUNGER

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ASC	Automatic Spelling Correction
BSB	Brasília
CECAD	Consult, Selection, and Extract the Information of Unified Registry
	Consulta, Seleção e Extração de informações do Cadastro Único
CEF	Caixa Econômica Federal
CNEFE	National Registry of Address for Statistics Analysis
	Cadastro Nacional de Endereços para Fins Estatísticos
CRAS	Reference Center of Social Assistance
	Centro de Referência de Assistência Social
CWB	Curitiba
IBGE	Brazilian Institute of Geography and Statistics
	Instituto Brasiliero de Geografia e Estatística
LSC	Longest Common Subsequence
MDS	Ministry of Social Development and Fight Against Hungry
NIS	Number of Social Identification
PBF	Programa Bolsa Família
PSB	Basic Social Protection
	Proteção Social Básica
REC	Recife

1. INTRODUCTION

This project aims to assist the Brazilian Government in analyzing the data of low-income families and mapping the real needs of Brazilian society. In 2001, Brazil created a project called Unified Registry for Social Programs of the Brazilian Government (Cadastro Único) to characterize and identify low-income families based on information about each city. Currently, 5570 Brazilian cities are registered there, in addition to the dimension of the country, the difficulty of registering those needy families results in a complex registration system. (Feitosa, 2010; Torres, 2010)

The Unified Registry must be done by government agents manually to confirm the family profile. As with all manual processes, this one also has a high probability of happening grammatical, typing, and interpretation errors. For that reason, a percentage of this wrongly typed data extracted from CECAD (Consult, Select and extract the information of Unified Registry – Consulta, Selação e Extração de informações do Cadatro Único) is currently considered an outlier. Consequently, misleading analysis of the country's development directly affects the implementation of social solutions, programs, and policies. (Ministério do Desenvolvimento Social, 2022; Nakashima, 2012; Torres, 2010)

Neighborhoods are used for Government analysis because most of the population needing social assistance lives in Subnormal Agglomerates ("*favelas*"), regions that usually started without a basic structure and whose population does not have easy access to education. Besides the existing Subnormal Agglomerate neighborhoods, meanwhile, others are being created, increasing the complexity to map this information. (Insituto Brasileiro de Geografia e Estatistica (IBGE), 2022; Ministério do Desenvolvimento Social, 2022)

Given the context, this research intends to use an Automatic Spelling Correction (ASC) system to correct the names of wrongly typed neighborhoods in the Unified Registry based on statistical, mathematical, and programming tools to increase the precision of the Brazilian Government's analysis.

This dissertation aims for the following goals:

- Adjust an Automatic Spelling Correction (ASC) system based on the Unified Registry structure, possible to be applied in all 5570 cities.
- A GitHub with all the code to be applied by anyone who needs the current solution found at this online address: https://github.com/hnakashima96/Thesis_2022;
- Awareness of the current low-income reality of the Brazilian population.

This dissertation was organized into seven sections and a reference section (eighth).

The first section is the Introduction. That section summarizes the research with a brief explanation of the problem, its context, and the possible solution. Moreover, the main objectives and deliverables are pointed out.

The second section will present the current context of the problem and state-of-the-art of the problem's solution.

The third section will explain the main concepts to understand the problem in a detailed discussion and the problem itself.

The fourth section will describe the technical solution of the problem, the Automatic Spelling Corrector system, with a detailed theory to construct the developed algorithm for the problem context. This section also explains the string similarity methods chosen and the presentation of the tools used to develop and analyze the data.

The fifth section is the discussion. This section discusses the structural part of the Automatic Spelling Corrector system. Firstly, presents the development of the Automatic Dictionary Creator for the chosen cities for analysis in this project. Then the error method chosen and how they will be applied. The last part describes the context model.

The sixth section presents the results based on the evaluation analysis of the algorithm.

The seventh section presents the results of the previous sections. And the eighth section is the conclusion, which shows the ideas developed during the project development and further developments for future projects.

2. LITERATURE REVIEW

In the last few years, communication channels show the results of the global pandemic COVID-19, such as inflation and poverty increasing in Brazil. Many of them show, that the current income of the population is not enough to supply their basic needs, and 52% of the Brazilian population is living in Food Insecurity. Besides that, it became usual to find people looking for food in the trash or buying leftovers from groceries shop. (Carolina Mesquita, 2021; Gioras Xerez, 2021; Jornal Nacional, 2022; Pedro Rafael Vilela, 2022)

Brazilian poverty came back to be highlighted in the last years, even though the Brazilian constitution guarantees a combination of services for people in social vulnerability: "Basic Social Protection" (PSB – Proteção Social Básica). Those services have the goal to guarantee rights and improvement for Brazilian society, creating services, programs, and projects which improve family and community relations. To optimize the identification of families with social vulnerability, the Brazilian government created the Unified Registry for Social Programs of the Brazilian government (*Cadastro Único*), a database created for the identification and socioeconomic characterization of low-income families based on the information of each city. (Feitosa, 2010; Governo Brasileiro, 2022; Governo do Estado Paraná, 2022; Torres, 2010)

PSB created the Reference Center of Social Assistance (CRAS – Centro de Referência de Assistência Social), to assist in the implementation of social solutions. Through this tool, local social protection reaches the population, identifying the social gaps in cities. CRAS analyses the area to organize and create social help and other assistance for the region, which promotes access to services, benefits, and social help becoming a place of reference for the local population and sector services. Then Unified Registry adds information on the poverty population. (Ferreira, 2009; Governo Brasileiro, 2022; Minisitério da Cidadania, 2015)

Although CRAS promotes analysis, it is important to understand their solutions depend directly on their local analysis, which means, where they are based. This happens because usually, the population which needs them lives in Subnormal Agglomerates. Brazilian Institute of Geography and Statistics (IBGE – Instituto Brasileiro de Geografia e Estatística) defines Subnormal Agglomerate as a form of irregular occupation of land owned by others – public or private – for housing purposes in urban areas and characterized by an irregular urban pattern, lack of essential public services and location in areas with restricting occupation. In Brazil, these irregular settlements are known by various names, but the most common is "*favelas*" or "*invasões*" (invasions) (Figure 1). (Insituto Brasileiro de Geografia e Estatistica (IBGE), 2022)

Subnormal agglomerates are in constant development, due to this reality neighborhoods are being created every moment. The plot of figure 2 shows the number of neighborhoods with a relevant number of families registered in the Unified Registry, which means, at least 0.1% of families registered in a neighborhood. It is possible to notice that, since 2012, the number of poverty neighborhoods was almost duplicated.

Given this reality, it is possible to infer the context of neighborhoods by CRAS is important to impact more families. Dirce Koga points out the importance of understanding the groups and individuals who live and work in an area, considering beyond the structure of the house but also its context around, such as a neighborhood. Besides the individual analysis in social assistance, this methodology makes it harder to visualize the community and all the features that influence their reality. (Koga, 2015)

Figure 1 Example of Subnormal Agglomerate in Belo Horizonte, "Vila Senhor dos Passos"). (Peret et al., 2018)

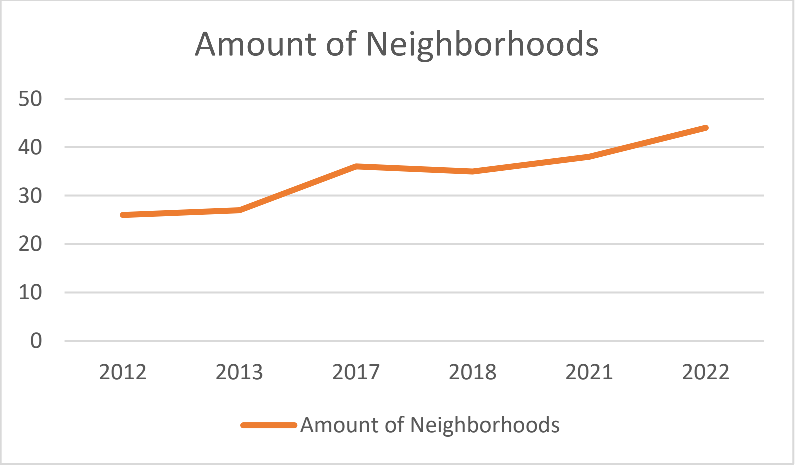


Figure 2 Growth of the number of neighborhoods with low-income families in Brasília during the last 10 years. (Ministério do Desenvolvimento Social, 2022)

All 5570 Brazilian cities have families registered in the Unified Registry database; their information was obtained through national tools. The process is created by four steps: I – Identification of target audience, II – Data gathering, III – adding the information in the registry system, and IV – data maintenance. The first two steps are made manually by city agents to evaluate and confirm the family conditions. The third step has the support of CEF, which developed the application to add the information from the previous steps to the national database. The fourth step is the merge of work from CEF with Federal Government to select the social program for each registry. (Chaves, 2021; Torres, 2010) Figure 3 shows a visualization of the processes described.

The steps II and III, the data added to the system has many errors. These errors will be described in detail in section 3. These errors are validated by CECAD (Consult, Select, and Extract the Information of Unified Registry). It is a panel that makes available to Brazilian cities quantitative information about valid/invalid registries, updates and not updated of the registry, people and families without civil registration, mapping of families without or with multiple fiscal numbers, and other identification issues. (Chaves, 2021; Ministério do Desenvolvimento Social, 2022; Sambiase et al., 2015)

It was possible to identify different typing errors in the CECAD panel, such as errors with spelling errors, wrong/mixed information, and wrong grammar among others. From this perspective, considering the importance of the data to support the validation that promotes social development, it was identified, that the correction of the wrong information of CECAD data can return relevant results for society.

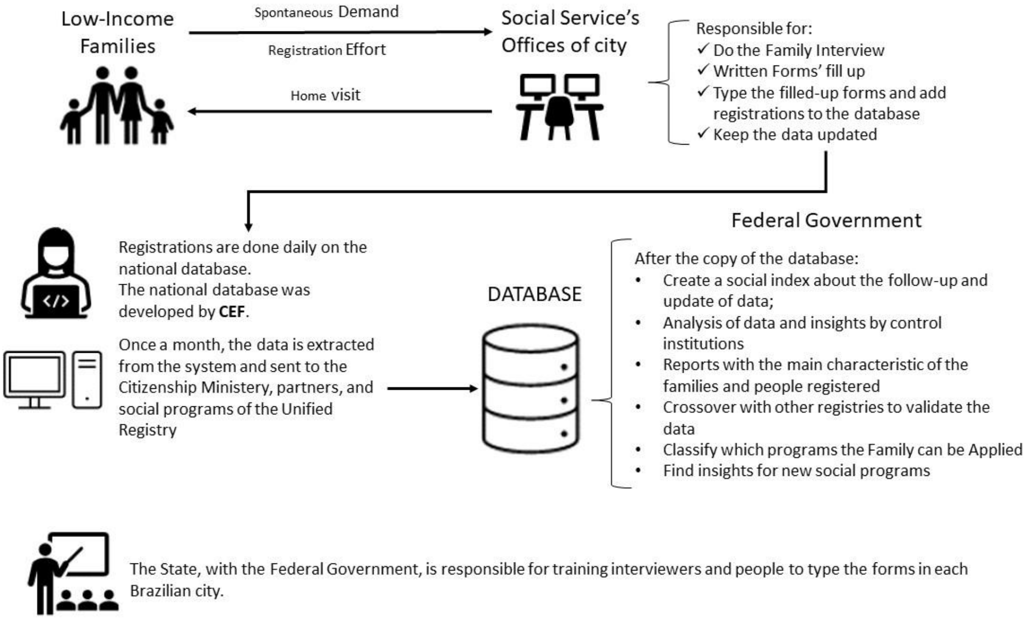

Low-Income Families
Spontaneous Demand
Registration Effort
Home visit
Social Service's Offices of city
Responsible for:
✓ Do the Family Interview
✓ Written Forms' fill up
✓ Type the filled-up forms and add registrations to the database
✓ Keep the data updated
Federal Government
Registrations are done daily on the national database.
The national database was developed by CEF.
Once a month, the data is extracted from the system and sent to the Citizenship Ministery, partners, and social programs of the Unified Registry
DATABASE
After the copy of the database:
Create a social index about the follow-up and update of data;
Analysis of data and insights by control institutions
Reports with the main characteristic of the families and people registered
Crossover with other registries to validate the data
Classify which programs the Family can be Applied
Find insights for new social programs
The State, with the Federal Government, is responsible for training interviewers and people to type the forms in each Brazilian city.

Figure 3 Visual description of the Unified Registry process and data visualization on a national level. Adapted from Nobrega, 2021.

References as Caio Nakashima, 2012; Ministério do Desenvolvimento Social e Combate à Fome, 2014; Paes De Barros et al., 2009; Petini et al., 2021 show that the data needs a solution to "correct the wrong typed neighborhood" to the "right one" for each city. Since CECAD has 5570 cities registered, the solution had to be automatized and simple to be applied for different city realities.

This problem presentation shows that a solution would be an Automatic Spelling Corrector (ASC) system with an "Automatic Dictionary creator" for each of the 5570 cities. For example, Brasília has officially 33 neighborhoods, but CECAD data shows 1293 neighborhoods. With the ASC system and the criteria to create a dictionary, the current project could optimize the 1293 neighborhood names to 38 automatically, mapping the "wrong names". The application of the solution will be presented in the following sections. (Secretaria de Estado de Governo do Distrito Federal, 2022)

Therefore, it showed relevance understand the state-of-the-art about "automatic spelling correction" system to validate the solution for the problem. For that, the query "automatic spelling correction" were researched for the years 2012-2022. This dissertation selected papers from three research sites: MDPI (https://www.mdpi.com/), ScienceDirect (https://www.sciencedirect.com/), and Scielo (https://scielo.org/). The period of review was from 2012 until 2022. It will review the relevant references about the present technique and case studies about the application of the technique.

MDPI returned 5 results, ScienceDirect 2295 and Scielo 1. In the platform ScienceDirect filtered the Article type for only "Review articles" and "Research articles", and the Subject Areas for "Computer Science" shrunk to 311 results. Subsequently, it was left 317 articles, these articles were validated to be analyzed in the first step of the analysis, where it was read the title and abstract and evaluated in if they are related to the research theme, where we found 14 relevant articles.

The second analysis was done among the 14 relevant articles, they were categorized between "Automatic Spelling Correction explanation" which characterizes articles with a good overview of the ASC structure and its variations, and "Example of application of Automatic Spelling Correction" which is articles focusing on exemplifying how to apply the solution in real problems. It returned 8 articles classified as "Automatic Spelling Correction explanation" and 6 as "Example of the application of Automatic Spelling Correction".

The Automatic Spelling Correction explanation, given from the general point of view, the ASC system is a solution for correcting a word spelled wrongly. The researchers have usually defined the ASC in two steps: verification and correction. Verification is the analysis of whether a sequence of characters is a lexicon word or not. Meanwhile, correction is the process to find candidates from a dictionary suggested to be the correction of the analyzed word, and sometimes, the system will automatically change the wrong-spelled word to the chosen candidate. (Andrade et al., 2012; Carneiro De Araújo et al., n.d.; Hládek et al., 2020)

Hládek et al., 2020 did a literature review about spelling correction indexed in two sources from 1991 until 2019. The papers in their research were separated into three groups based on the main components of a spelling correction system. The first group uses a set of expert rules to correct spelling errors. The second group adds a context model to rearrange the correction candidates with the context. The third group learns error patterns from a training corpus. His methodology was used as a reference for the literature review, besides the guidance to lead to other references for the current research.

Carneiro De Araújo et al., n.d. did a background perspective on the evolution of similarity between words and concepts to implement automatic correctors. However, the main development is the automatic corrector based on n-grams and the affix deletes from the words to analyze the similarity. His research was great to give an understanding of how to initialize the word comparison, such as the edit distance method. Also, it supported finding references to ASC structure and understanding.

Andrade et al., 2012 proposed an automatic spell checker to be used at the end of the automatic collector for Portuguese Web text, the HASCH (High-performance Automatic Spell Checker). This work highlighted the adjustment of ASC for a specific case such as Web text. The article shows the challenges found in Web texts and which steps he had to change to keep the system structured and reach his goal.

The other articles about the explanation of the ASC technique were the application of the method in Arabian (Abdellah et al., 2020; Aouragh et al., 2021; Mohammed & Abdellah, 2018; Nejja & Yousfi, 2015) and Amazigh (Chaabi & Ataa Allah, 2022). Besides the application of the method in other language structures, Aouragh et al., 2021 presented an understanding of the N-grams language, and (Chaabi & Ataa Allah, 2022)'s work was an application of the method chosen to be used in this project.

Among the Example of the application of Automatic Spelling Correction articles Nagata et al., 2017 and Sharma & Gupta, 2015 were examples of variations in the development of ASC. Meanwhile, the other articles highlighted different field applications of ASC, such as Lai et al., 2015 and Sarker & GonzalezHernandez, 2018 applied in medicine, Neto et al., 2020 in an Offline handwritten text recognition system, and Alwabel, 2021 in code compilers.

3. DATA EXPLANATION

UNIFIED REGISTRY

The Unified Registry for Social Programs of the Brazilian government (*Cadastro Único*) is a database created to identify and characterize the socioeconomic status of low-income families based on the information of each city. The registration is done through the federal government, cities, and Caixa Ecônomica Federal (CEF) to understand which are the poorest and most vulnerable families of the population. The Unified Registry was based on the Brazilian social program "*Programa Bolsa Família*" (PBF) and currently supports more than 20 social programs and policies. (Barca, 2017; Feitosa, 2010; Torres, 2010; World Without Poverty, 2022)

The project was developed during the mandate of present Fernando Henrique Cardoso, in 2001, by the law "Unified Registry for Social Programs of the Federal Government" with the CEF's support. Although the real implementation was in 2003, with the PBF program. Among the requirements for the project Unified Registry would be the development of beneficiary identification, nowadays known as NIS, number of social identification. (Chaves, 2021; Torres, 2010)

NIS represents an individual identification code orientated to register ideal social supports based on the *per capita* wage in a family. Also, it considers the sociodemographic situation of each person registered. Although, the highlight of NIS among other Brazilian identifications is the target of the low-income population which need extra support. (Torres, 2010)

From the methodological perspective, the Unified Registry is composed of a mapping of each family member's information, and family lifestyle from the population in which the monthly wage is half the minimum wage *per capita*. This information is not considered in census analysis, due to it a questionary has been created and tests to validate the real status of the family. (Torres, 2010)

One of the main challenges of registration is to keep the data updated to prevent fraud and risk of validation about the information from the tool. It is expected the updated each registry every two years after the registration by each city. Those, which are not updated have the invitation to do an annual process: Registration Review. The control is more precise in bigger cities that have more technology. That process is relevant due to the automatic process to be registered in Unified Registration, as Bolsa Família: if the family has the wage profile, the Federal Government selects a family which will receive to withdraw the money support. Not only for these main social programs the Unified Registry is relevant but also for social programs, to which families can reach out such as Social Fare of Electric Energy (*Tarifa Social de Energia Elétrica*), a program that provides a discount on electricity bills on the respective light companies. (Chaves, 2021)

Besides the Registration Review, the Unified Registry has more two methods to prevent fraud: the Registry Investigation and Logical Exclusion. The Registry Investigation started in 2005, and since 2016 the inspection become semiannual. The process consists in compare the data of the Unified Registry per period with other official registries composed of information on wage or pension benefits and social support. This analysis was created to identify false positive registries, defined as people identified as poverty profiles but receiving wages above the limit defined. (Chaves, 2021)

The Logical Exclusion happens after the inspection methods mentioned were applied. In each step, the deadline of the registry is updated, then families which are not updated have their data deleted by the Citizenship Ministry, and the social benefits canceled. (Chaves, 2021)

The Constitution of the federative republic of Brazil in 1988 was an important goal for the public laws of social protection. It considered the health rights of social welfare, pension, and social assistance, therefore, organizing the strategies of application based on cover power (item I of article 194) and help and selection and distributivity of supply basic benefits and services (item III of article 194). Social assistance is written in article 203 and highlights who need it, along with social health and pension. Rules of social assistance are on the Organic Law of Social Assistance (LOAS), 1993, because of the movement of social service for its regulation. (Castro & Modesto, 2010)

As mentioned, in 2003, the creation and enforcement of the Unified Registry made it possible to map the low-income population of Brazilian Cities with registry precision. However, it was created for a specific social program (PBF), Unified Registry has a range of information about a family's lifestyle evaluated in six dimensions: i) vulnerability; ii) education access; iii) work access; iv) resources availability (wage and expenses *per capita* and food expenses); v) child develop; vi) home conditions (Figure 4). (Paes De Barros et al., 2009)

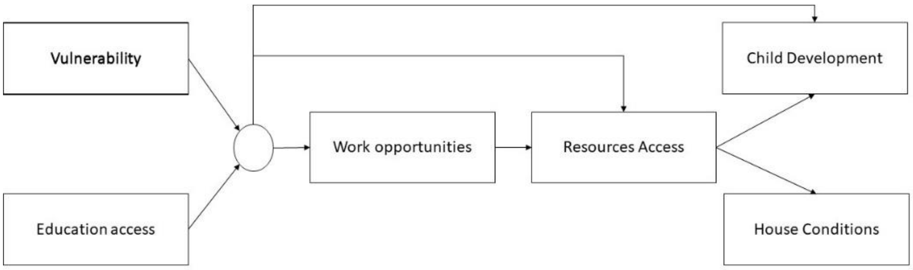


Figure 4 Interrelations between the dimensions of life conditions. Adapt from (Paes De Barros et al., 2009)

Among the benefits of cash transfer, the Unified Registry dimensions take on a leading role due to the diagnostic function of the basic social rights of families, it helps governments map public policies based on the real needs of society. "Fight the poverty and the social gap have a fundamental politic goal in a system of social protection, which must guarantee equality of access and opportunities for everyone". Programs such as the Unified Registry created a repercussion of poverty decrease and was responsible for a relevant reduction of the income gap in Brazil until 2018. (Paes De Barros et al., 2009; Petini et al., 2021)

The analytical perspective of the Unified Registry demands a unique characteristic that validates the integration of social interventions. Continuous use promotes quality due to the identification of flaws and improvements of the necessary information for the database. (Paes De Barros et al., 2009)

The data gathering is the city's responsibility, which facilitates the diagnostic of each family member to evaluate the poverty evolution and identify the basic needs of Brazilians. Zooming out on local analysis, the Unified Registry possibilities the follow-up of communities' life conditions which the government interviewed with specific solutions (ex.: the creation of schools, and hospitals) and global performance of local administrations. Finally, from a regional perspective, the results add to the census which happens every 10 years. (Paes De Barros et al., 2009)

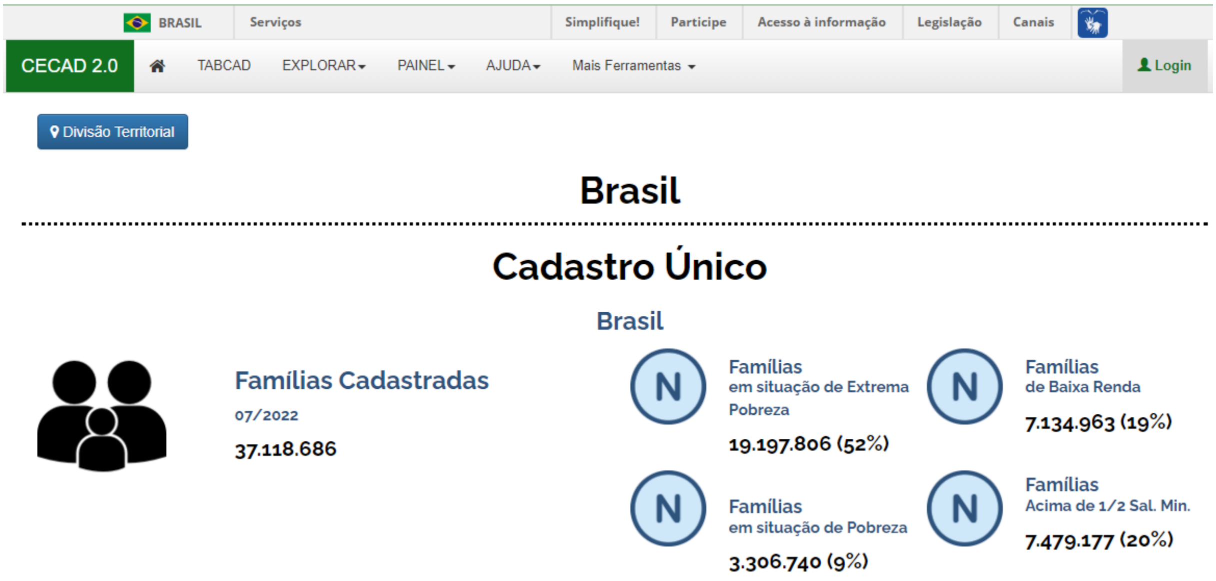
In 2014, the Ministry of Social Development and Fight Against Hungry (MDS), announced the exit of Brazil from the Hunger Map. The main factors that led Brazil to this achievement were: i) an increase in food access; ii) a real increase of income in 71,5% of the extreme poverty ones and new jobs; iii) Programs of the Federal Government to Income Access; iv) 43 million children and teenagers with meals. Thus, 20% of the poverty Brazilian had their income increased. The MDS report confirms the importance of the Unified Registry for Brazilian society and the impact of the development of research in the field.

> “Campello points out, based on data, ‘we reached a percentual of 1.7% of Food Insecurity in Brazil. This means that 98.3% of the Brazilian population has access to food and food security highlights. ‘This is an important winning’. From 2002 until 2013, decreased in 82% of the Brazilian population considered Food Insecurity.”
>
> (Ministério do Desenvolvimento Social e Combate à Fome, 2014)

The site CECAD, a tool developed by Senarc and Sagi to facilitate the access of Unified Registry data and assist planning and implementation of social solutions, programs, and policies, is possible to validate how the status of the low-income of the Brazilian population. (Nakashima, 2012)

CECAD facilitates cross attributes of forms’ registry with people and families registered to create analysis tables. The results present on the platform are aggregate and unidentified from cities, states, and countries to make possible public access. Due to all mentioned reasons, the tool represents a bigger variety of consults, adding graphics, temporal series, and reports. The data update happens monthly and can have from 30 to 60 days of lag compared to the online System of Unified Registry. (Ministério do Desenvolvimento Social, 2022)

Access to the platform is done through: https://cecad.cidadania.gov.br/.Figure 4 shows the initial front page of CECAD.



Brasil

Cadastro Único

Brasil

Famílias Cadastradas
07/2022
37.118.686

Famílias
em situação de Extrema Pobreza
19.197.806 (52%)

Famílias
de Baixa Renda
7.134.963 (19%)

Famílias
em situação de Pobreza
3.306.740 (9%)

Famílias
Acima de 1/2 Sal. Min.
7.479.177 (20%)

Figure 5 Initial page of CECAD (Access: October 5th of 2022).

Figure 6a proves an increase in the number of families registered since 2012, this is confirming the importance of the commitment of cities to identify low-income families in favor of social development.

Figure 6b, presents that since 2018 the program has been suffering challenges, mostly in 2020, when the COVID-19 pandemic started affecting Brazil and a side effect making harder the registration process.

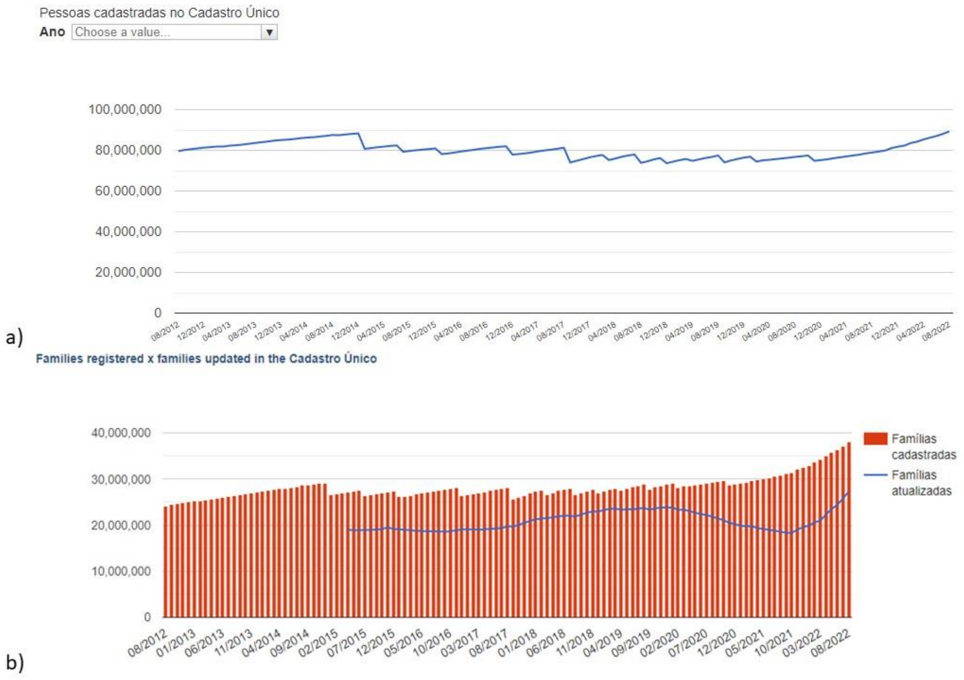


Figure 6 a) Presents the registrations in CECAD since its creation. b) Families Registered vs. Families Updated in Unified Registry from 2012 until 2020. (Access: October 21st of 2022).

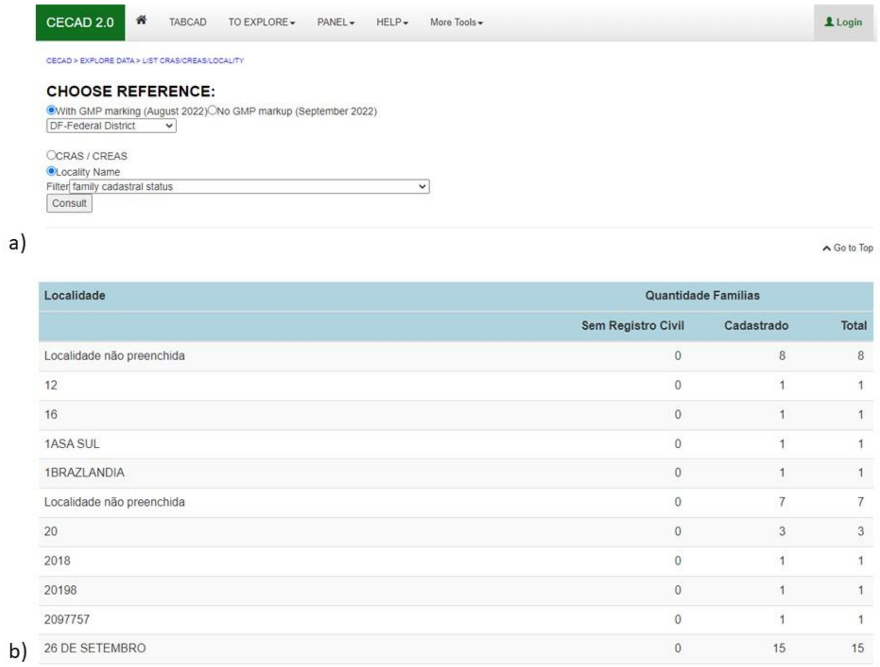
The tool CECAD is important to optimize the data access for the cities. São José do Rio Preto is a city that uses constantly the data for analysis and crosses with geo-referenced data. That analysis provided a relocation of CRAS (Reference Center of Social Assistance) and the creation of two flying teams. That relocation made CRAS more accessible to families in vulnerable places and unregularized. Besides this case, in the capital Rio de Janeiro, registered users of the Unified Registry were responsible for 54% of beneficiaries of the program *Tarifa Social*. This initiative represents 4 million reais of discount to families of low-income in the city.

Organizing the Unified Registry was important to standardize some concepts such as "what is family?". A family in Unified Registry is a nuclear unit composed of one or more people, eventually considered a few more, that support the income or have your expenses together and must be living in the same address". It is important to highlight that besides the program demands of a shared house, income, and expenses the people in the family do not necessarily must be blood-related.

The last version of the Unified Registry System is 100% online, which means that a configured computer can add an insert, update, transfer, or exclusion of people and families by internet on the national database independently of which city. This promotes the dynamic of the process and decreases the multiplicity and difference of data. (Sambiase et al., 2015)

Problem

As explained in section 2, the data gathering is made manually then grammatical, type, and interpretation errors have a high probability to happen. Figures 7a and 7b show the results of this data gathering in the CECAD. Figure 7a shows the fill-up of CECAD gaps to return aggregated information by the neighborhood of Unified Registry of Brasília, the same process can be done for other cities. Figure 7b shows a part of the output from the lookup of figure 7a in the platform CECAD.



CECAD > EXPLORE DATA > LIST CRAS/CREAS/LOCALITY

CHOOSE REFERENCE:

● With GMP marking (August 2022) ○ No GMP markup (September 2022)

DF-Federal District

○ CRAS / CREAS

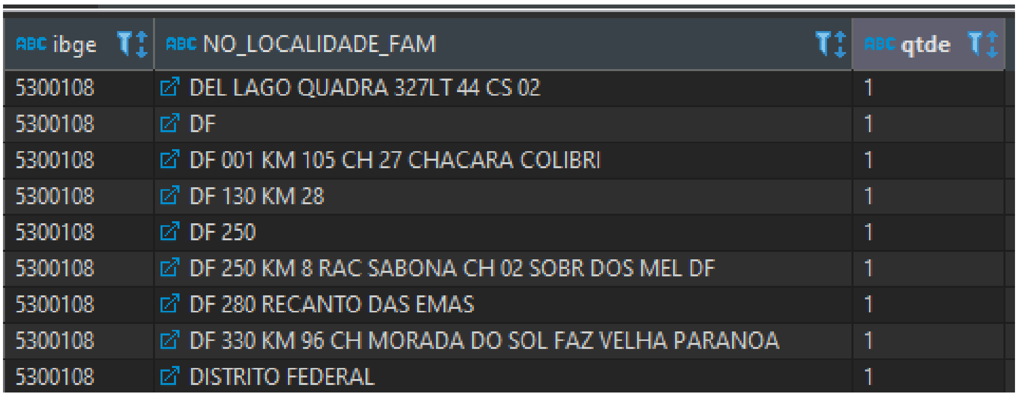
● Locality Name

Filter family cadastral status

Localidade	Quantidade Famílias		
	Sem Registro Civil	Cadastrado	Total
Localidade não preenchida	0	8	8
12	0	1	1
16	0	1	1
1ASA SUL	0	1	1
1BRAZLANDIA	0	1	1
Localidade não preenchida	0	7	7
20	0	3	3
2018	0	1	1
20198	0	1	1
2097757	0	1	1
26 DE SETEMBRO	0	15	15

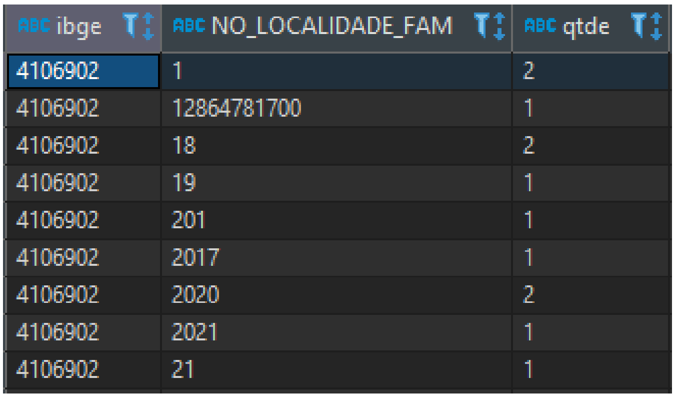
Figure 7 a) Look up in CECAD to return aggregated information by the neighborhood of the Unified Registry of Brasília. b) Part of the table returned from aggregated information by the neighborhood of the Unified Registry of Brasília.

Figures 8, 9, and 10 are examples of errors found on the lookup from CECAD by city: figure 8 shows a case where staff added the address instead of the neighborhood in the forms, and figure 9 is an example that considered a few neighborhoods' names as numbers and Figure 10 presents some grammatical errors related with a human error when typing. All these errors are ignored data (outliers) because staffs are not able to identify which neighborhood these families live and sometimes making it harder to identify the real needs of a community.



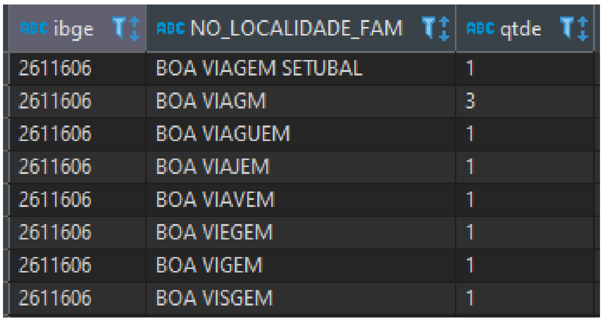
ibge	NO_LOCALIDADE_FAM	qtde
5300108	DEL LAGO QUADRA 327LT 44 CS 02	1
5300108	DF	1
5300108	DF 001 KM 105 CH 27 CHACARA COLIBRI	1
5300108	DF 130 KM 28	1
5300108	DF 250	1
5300108	DF 250 KM 8 RAC SABONA CH 02 SOBR DOS MEL DF	1
5300108	DF 280 RECANTO DAS EMAS	1
5300108	DF 330 KM 96 CH MORADA DO SOL FAZ VELHA PARANOA	1
5300108	DISTRITO FEDERAL	1

Figure 8 Example of wrong information typed as neighborhood information (NO_LOCALIDADE_FAM) in Brasília data. This example shows some addresses typed on neighborhood information.



ibge	NO_LOCALIDADE_FAM	qtde
4106902	1	2
4106902	12864781700	1
4106902	18	2
4106902	19	1
4106902	201	1
4106902	2017	1
4106902	2020	2
4106902	2021	1
4106902	21	1

Figure 9 Example of wrong information typed as neighborhood information (NO_LOCALIDADE_FAM) in Curitiba data. This example shows cases that are typed information that is not able to relate to neighborhood information, like numbers.



ibge	NO_LOCALIDADE_FAM	qtde
2611606	BOA VIAGEM SETUBAL	1
2611606	BOA VIAGM	3
2611606	BOA VIAGUEM	1
2611606	BOA VIAJEM	1
2611606	BOA VIAVEM	1
2611606	BOA VIEGEM	1
2611606	BOA VIGEM	1
2611606	BOA VISGEM	1

Figure 10 Example of wrong information typed as neighborhood information (NO_LOCALIDADE_FAM) in Recife data. This example shows the grammatical errors of manual typing that during a neighborhood analysis have a high probability to be considered an outlier even though is possible to identify which neighborhood those families are mapped.

A solution for the mentioned problems could be closed forms with a dropdown box with the list of neighborhoods of the city. This would optimize the CRAS' staff process and at least minimize some of the errors mentioned, although, as explained this is not able in Brazil because of the constant growth of cities and the challenge to map all the neighborhoods in 5570 cities.

Another solution would be to apply already-made algorithms used in Natural Process Language as SymSpell. Although in the problem of Unified Registry the words are names, which means that demands a specific dictionary, and this dictionary would be in constant update to correct new entries.

The complexity of the Unified Registry problem, this dissertation intends to develop an Automatic Spelling Correction (ASC) system to correct the name of wrongly typed neighborhoods in the Unified Registry based on statistics, mathematical and programming tools to increase the precision of CRAS' analysis.

4. METHODOLOGY

AUTOMATIC SPELLING CORRECTION

The goal of this project is to develop an Automatic Spelling Correction (ASC) system to correct the name of wrongly typed neighborhoods in the Unified Registry (Figure 11).

An ASC is a process to identify wrongly typed words and based on a dictionary, then correct each one for the most proper word for the context (Angell et al., 1982; Hládek et al., 2020). Hládek, 2020 describes the process as:

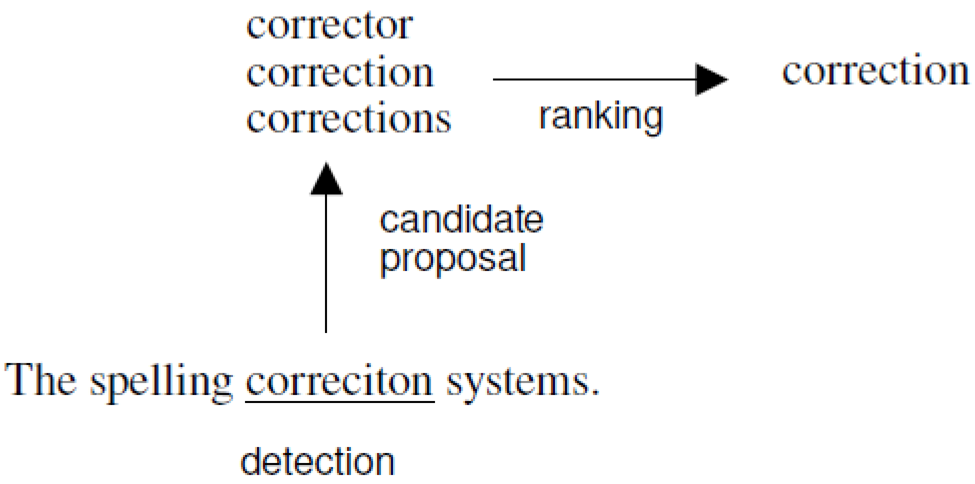


Figure 11 Process of Automatic Spelling Correction. (Hládek et al., 2020)

Composed by:

- Dictionary: The list of correct candidates that belongs to the set of all correct words. In this project, the dictionary was defined as the neighborhood per city which has at least 0.1% of all mapped families in the city of analysis registered there.
- Error model: is the probability of the string right happening over the wrong. A model which identifies string similarity. Previous works and references supported the conclusion to use the Edit Distance method (Damerau-Levenshtein) and vector space method (Jaccard Distance). Both of mentioned methods will be described in the next section.
- Context Model: candidates that best fit into the current context.

STRING SIMILARITY METHODS

Any human action has a high probability to generate errors, for example, typing usually has many spelling errors. To solve these problems were developed "String Correction" methods. String correction problems can be faced from two different perspectives: similarity and dissimilarity metrics (edit distance) or statistical similarities (vector space model). (Hussain, 2012)

Edit distance refers to the class of string similarity metrics, which focuses on finding a measure for either similarity or dissimilarity of two given strings. More specifically these algorithms find the distance between two strings, which is the number of operations required for one string to be transformed into another. The measure distance can be calculated into a similarity score between strings. Since the similarity measure is a distance, the algorithms will perform poorly for strings of different sizes. The magnitude of this poor performance is relative to the difference in size. (Hussain, 2012; Jacobs, 2004)

Meanwhile, the statistical similarity models refer to measures in a vector space of strings or queries. These algorithms find how much two strings have in common by weighting their commonality. The term is the measure for the similarity comparison. (Dubin, 2014; Hussain, 2012)

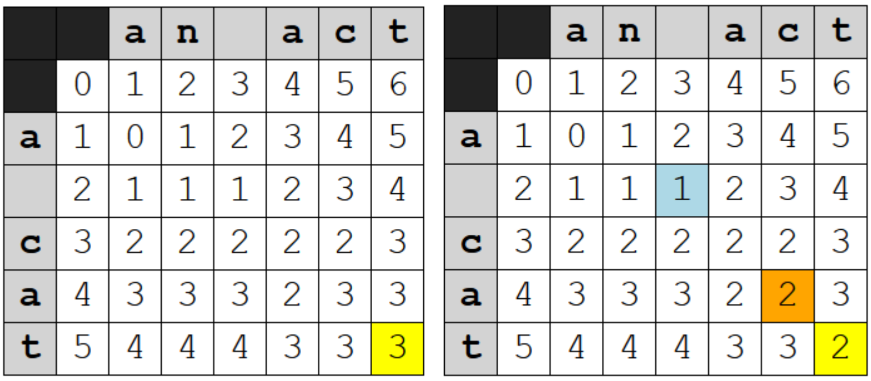
To apply the statistical similarity models, use N-grams, a Natural Language Processing (NLP) method, a sequence of words with N length. It is used to optimize many processes of text analysis, such as spelling error corrections, they can be useful to identify the probability of the high probability occurrence after some specific N-gram. For example, in the words {me, ed, di, it, ta, at, te} and {me, ed, di, it, tt, ta, at, ti, io, on}, the second string has a spelling error which by bigram frequency is possible to identify the probability of the bigram after "it" be right or wrong. (Hussain, 2012; Prachi Kumar, 2017)

This research will approach the edit distance methods Damerau-Levenshtein and Jaccard Distance. The criteria to choose the methods were based on references and previous works.

- Damerau-Levenshtein

Damerau-Levenshtein distance is an improvement of the Levenshtein edit distance. The Levenshtein Distance allows counts of each substitution, insertion, and deletion with the weight of one to convert a word to another. This is done by checking if each letter matches, if not an operation is executed until the complete string were checked. (Hussain, 2012)

The difference in Damerau-Levenshtein is the addition of transposition beyond the costs of the Levenshtein. Transposition is when the last two elements in each sequence are the same but swapped but consider two steps before the transposition and add one (Figure 12). (James M, 2022)



		a	n		a	c	t
	0	1	2	3	4	5	6
a	1	0	1	2	3	4	5
	2	1	1	1	2	3	4
c	3	2	2	2	2	2	3
a	4	3	3	3	2	3	3
t	5	4	4	4	3	3	3

		a	n		a	c	t
	0	1	2	3	4	5	6
a	1	0	1	2	3	4	5
	2	1	1	1	2	3	4
c	3	2	2	2	2	2	3
a	4	3	3	3	2	2	3
t	5	4	4	4	3	3	2

Figure 12 Calculation of edit distance between "na act" and "a cat". The left is the Levenshtein Distance and the right, the Damerau-Levenshtein Distance. (James M, 2022)

Figure 12 shows the difference between Levenshtein Distance and Damerau-Levenshtein Distance. In the Levenshtein distance, the edit distance is 3 between "an act" and "a cat". Meanwhile, in the Damerau-Levenshtein Distance, the distance becomes 2 because the transposition is a single action instead of two substitutions. (James M, 2022)

The addition of the transposition in Damerau-Levenshtein Distance guarantees that triangle inequality, the distance between two sequences can be greater than the sum of the distances between each to a common third sequence, does not hold. (James M, 2022)

- Jaccard Distance

Previous projects as Henrique et. al, 2020 used Dice's coefficient to correlate the Unified Registry with the Address list update service requested for Statistical Purposes – CNEFE (Cadastro Nacional de Endereços para Fins Estatísticos; National Registry of Address for Statistics Analysis) was the reference to work with statistical similarity methods. However, other research described that Dice's coefficient was very similar to Jaccard Similarity, as shown in the equations below:

$$Jaccard\ Index = \frac{|\,A \cap B\,|}{|\,A \cup B\,|} \quad \& \quad Dice's\ Coefficient = \frac{2\,|\,A \cap\ B\,|}{|A| + |B|}$$

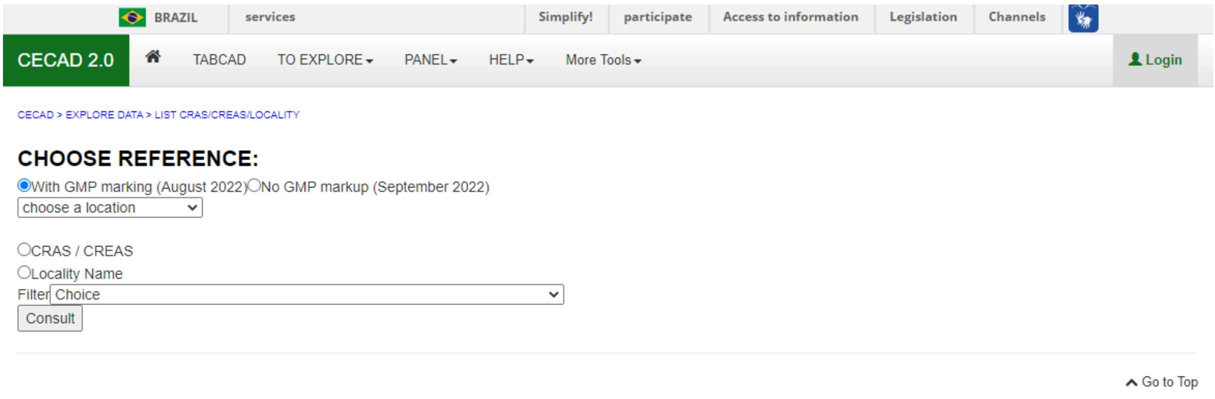
Both are not considered distance metrics, because they hold the triangle inequality (Gallagher, 1999). Due to this perspective, was decided that the best option would be "Jaccard Distance". In (Kosub, 2016) prove that it fulfills the properties of a metric, mostly the triangle inequality.

Jaccard Distance measures the similarity between two sets, mathematically defined, it is the calculation of the ratio of the difference between set union and set intersection over set union (Kosub, 2016):

$$Jaccard\ Distance = \ 1 - Jaccard\ Index$$

TOOLS FOR PROJECT DEVELOPMENT

The data were extracted from CECAD (https://cecad.cidadania.gov.br/). To extract the data the user must go to the "TO EXPLORE" tab and choose the CRAS/CREAS LOCALITY option. It will return to Figure 13, in this page the user can choose the city which he or she wants to analyze or by CRAS office. Then choose the filter, which in this project was chosen "family cadastral status". All the data used in this project were from 2021.



CHOOSE REFERENCE:

● With GMP marking (August 2022) ○ No GMP markup (September 2022)

choose a location

○ CRAS / CREAS

○ Locality Name

Filter Choice

Consult

Figura 13 CECAD interface to extract data by the city for analysis. (Ministério do Desenvolvimento Social, 2022)

After the data extraction, started the process of processing the data and analysis. All the data was processed using Python programming language on the IDE program PyCharm Community Edition

2021.3.2. The code is indexed in https://github.com/hnakashima96/Thesis_2022 and explained in its README. The results were analyzed in DBeaver, an open-source database administrator.

The development of the project will be described in the following section with the detailed methodology.

Figure 14 presents the flowchart to develop this project.

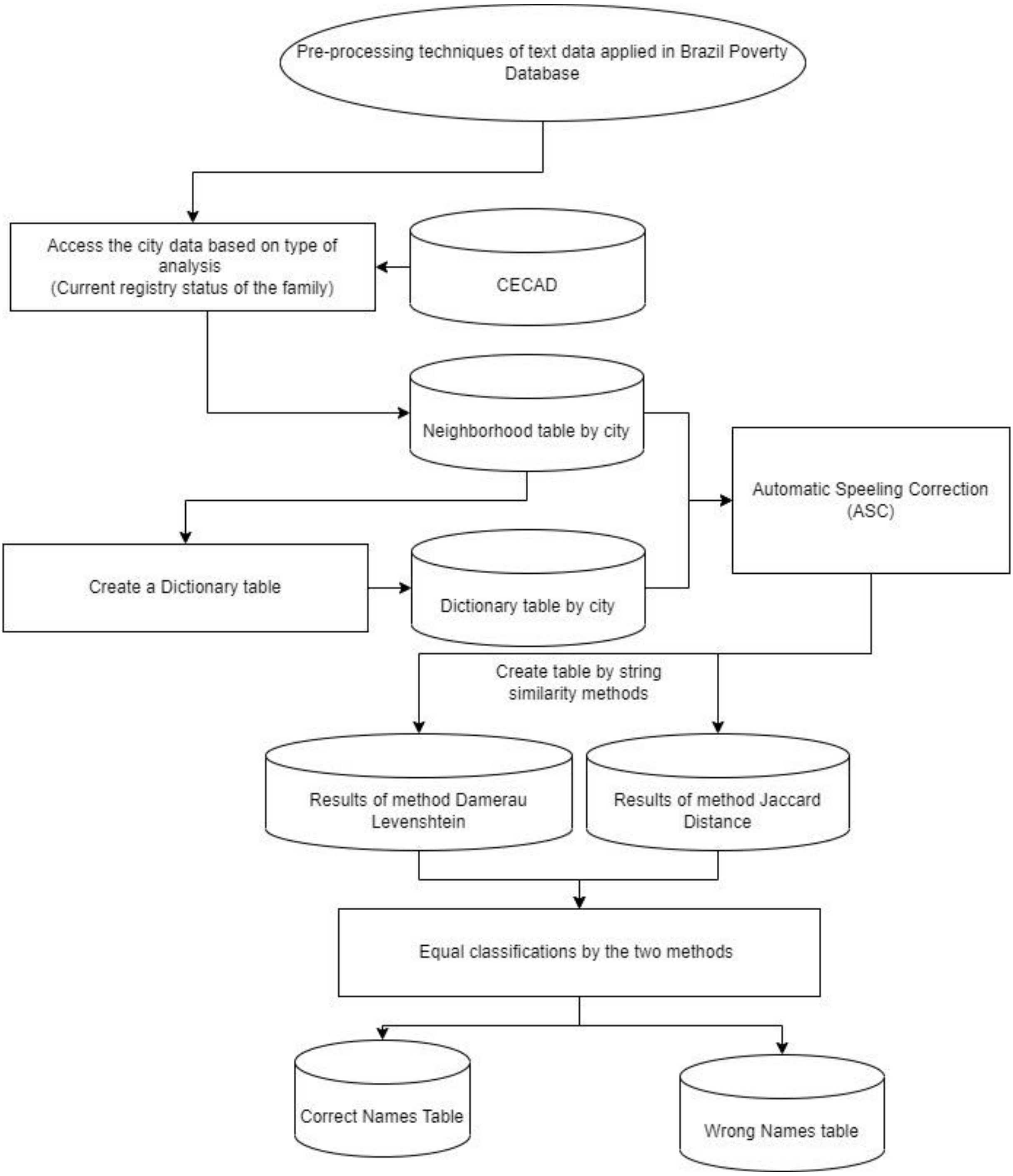

Pre-processing techniques of text data applied in Brazil Poverty Database
Access the city data based on type of analysis (Current registry status of the family)
CECAD
Neighborhood table by city
Automatic Speeling Correction (ASC)
Create a Dictionary table
Dictionary table by city
Create table by string similarity methods
Results of method Damerau Levenshtein
Results of method Jaccard Distance
Equal classifications by the two methods
Correct Names Table
Wrong Names table

Figure 14 Flowchart of the steps to develop this project.

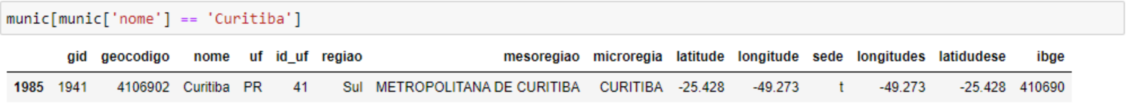
5. DISCUSSION

DICTIONARY DEVELOPMENT

The dictionary used in the ASC of this research was created by a statistical analysis although it can be improved by an agent of government support.

The statistical analysis considered the relevance of the neighborhood would define its name as right, therefore relevance was calculated based on the number of families registered in its name, and the neighborhood relevance above a defined threshold becomes the dictionary for the specific city.

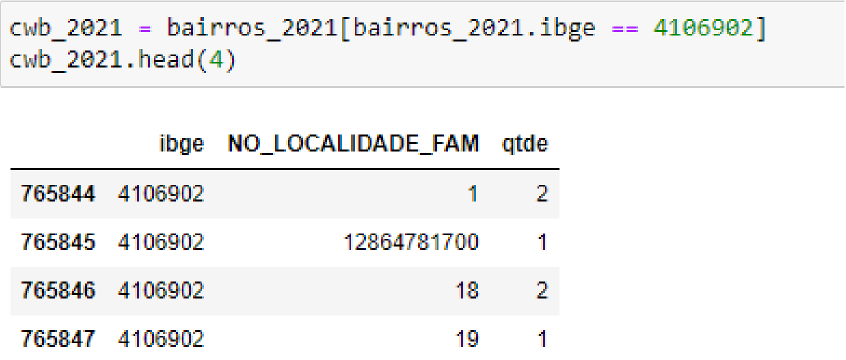
Firstly, were filtered the data of a city in the “bairros_2021” database based on its geocodigo founded on the “municípios” database, in the Curitiba example (Figure 15) is “4106902”. Figure 16 is the fourth first result of the Curitiba database.



```
munic[munic['nome'] == 'Curitiba']
```

	gid	geocodigo	nome	uf	id_uf	regiao	mesoregiao	microregia	latitude	longitude	sede	longitudes	latidudese	ibge
1985	1941	4106902	Curitiba	PR	41	Sul	METROPOLITANA DE CURITIBA	CURITIBA	-25.428	-49.273	t	-49.273	-25.428	410690

Figure 15 IBGE information from the city of Curitiba.



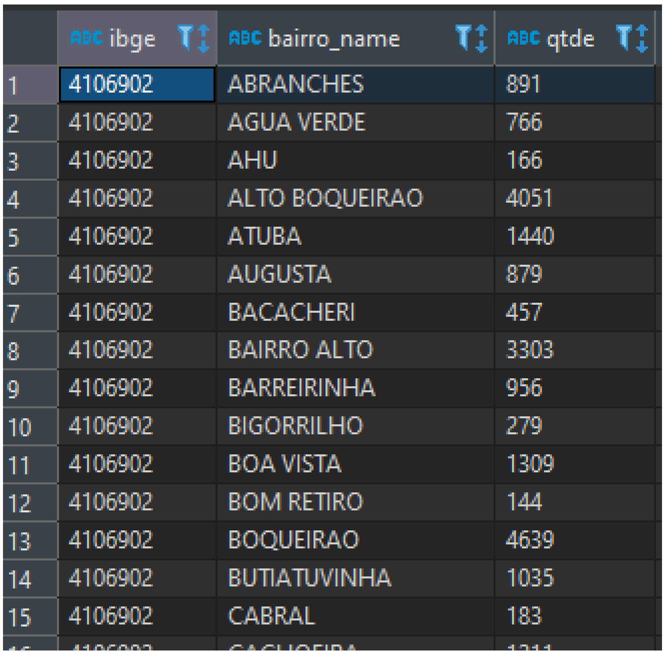
```
cwb_2021 = bairros_2021[bairros_2021.ibge == 4106902]
cwb_2021.head(4)
```

	ibge	NO_LOCALIDADE_FAM	qtde
765844	4106902	1	2
765845	4106902	12864781700	1
765846	4106902	18	2
765847	4106902	19	1

Figure 16 First fourth entries from the Curitiba neighborhood database.

Given the result city table (Figura 16), “ibge” represents the city identification, the “NO_LOCALIDADE_FAM” the neighborhood name registered on the Unified Registry, and “qtde” the number of families registered in that neighborhood name. Then I created a new column called “qtde_per” that calculates the ratio of “qtde” over the total families registered in the city of analysis. Qtde_per will represent the relevance of the neighborhood name because I understood that if a neighborhood has a relevant percentual of families registered in its name, it has a high probability of being right-typed.

Following the conclusions of the new column, were filtered all neighborhoods with qtde_per above 0.001, equivalent to 0.1% to create the final dictionary table of the city (Figura 17).



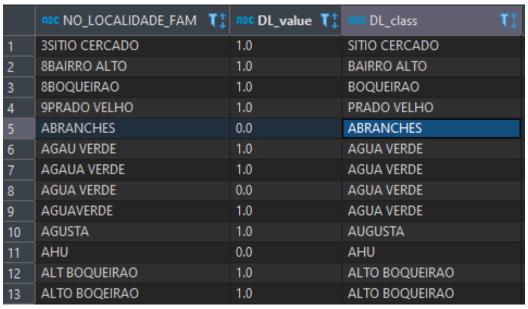
	ibge	bairro_name	qtde
1	4106902	ABRANCHES	891
2	4106902	AGUA VERDE	766
3	4106902	AHU	166
4	4106902	ALTO BOQUEIRAO	4051
5	4106902	ATUBA	1440
6	4106902	AUGUSTA	879
7	4106902	BACACHERI	457
8	4106902	BAIRRO ALTO	3303
9	4106902	BARREIRINHA	956
10	4106902	BIGORRILHO	279
11	4106902	BOA VISTA	1309
12	4106902	BOM RETIRO	144
13	4106902	BOQUEIRAO	4639
14	4106902	BUTIATUVINHA	1035
15	4106902	CABRAL	183

Figure 17 The first 15th neighborhood from the dictionary table with the number of low-income families is relevant to be considered right of Curitiba.

Error Model

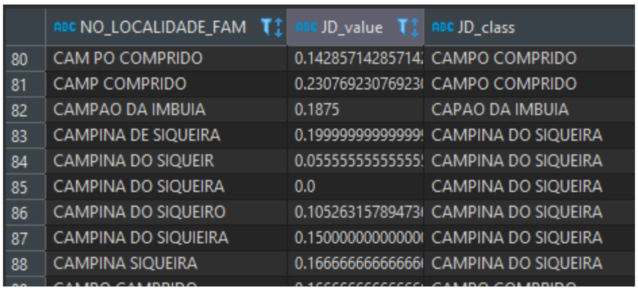
As mentioned in the methodology section, the error model used was two methods of string similarity: an edit distance (Damerau-Levenshtein) and statistical similarity (Jaccard Distance).

Both methods were applied to compare each neighborhood name on the city database to all neighborhood names on the dictionary created for the respective city. For the Damerau-Levenshtein method, the closer to the Damerau-Levenshtein coefficient more similar the names would be (Figure 18). If the coefficient is equal to zero, it means that the words are equal, which means that match one of the words in the dictionary. This was expected because the database to create the dictionary and do the comparison are the same. In the Jaccard Distance method, closer to zero, the Jaccard Distance index is more like the names would be (Figure 19).



	NO_LOCALIDADE_FAM	DL_value	DL_class
1	3SITIO CERCADO	1.0	SITIO CERCADO
2	8BAIRRO ALTO	1.0	BAIRRO ALTO
3	8BOQUEIRAO	1.0	BOQUEIRAO
4	9PRADO VELHO	1.0	PRADO VELHO
5	ABRANCHES	0.0	ABRANCHES
6	AGAU VERDE	1.0	AGUA VERDE
7	AGAUA VERDE	1.0	AGUA VERDE
8	AGUA VERDE	0.0	AGUA VERDE
9	AGUAVERDE	1.0	AGUA VERDE
10	AGUSTA	1.0	AUGUSTA
11	AHU	0.0	AHU
12	ALT BOQUEIRAO	1.0	ALTO BOQUEIRAO
13	ALTO BOQEIRAO	1.0	ALTO BOQUEIRAO

Figure 18 Example of Damerau-Levenshtein string similarity in data of Curitiba. DL_value represents the Damerau-Levenshtein coefficient. DL_class represents the right neighborhood name to classify the wrong name.



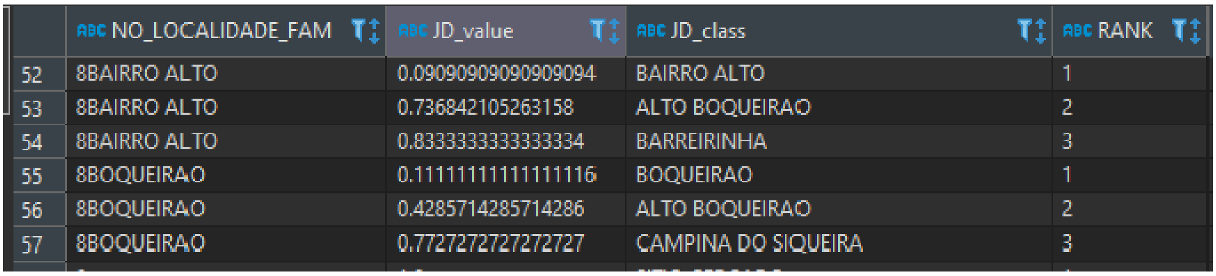
	NO_LOCALIDADE_FAM	JD_value	JD_class
80	CAM PO COMPRIDO	0.14285714285714	CAMPO COMPRIDO
81	CAMP COMPRIDO	0.23076923076923	CAMPO COMPRIDO
82	CAMPAO DA IMBUIA	0.1875	CAPAO DA IMBUIA
83	CAMPINA DE SIQUEIRA	0.19999999999999	CAMPINA DO SIQUEIRA
84	CAMPINA DO SIQUEIR	0.05555555555555	CAMPINA DO SIQUEIRA
85	CAMPINA DO SIQUEIRA	0.0	CAMPINA DO SIQUEIRA
86	CAMPINA DO SIQUEIRO	0.10526315789473	CAMPINA DO SIQUEIRA
87	CAMPINA DO SIQUIEIRA	0.15000000000000	CAMPINA DO SIQUEIRA
88	CAMPINA SIQUEIRA	0.16666666666666	CAMPINA DO SIQUEIRA

Figure 19 Example of Jaccard Distance string similarity in data of Curitiba. JD_value represents Jaccard Distance Index. JD_value represents the right neighborhood name to classify the wrong name.

Context Model

The context model in this problem consists of the number of families each live in the neighborhood. Therefore, the tiebreaker for names with the same similarity coefficients would be the number of families in the reference neighborhood. For example, if the algorithm reaches that a wrong neighborhood name can be classified between two neighborhoods of the dictionary the algorithm will choose the one which has more families registered.

Also, the final database of analysis was created with the first three dictionary neighborhood name chosen for a single wrong neighborhood name (Figure 20). This increase in the database makes it possible to identify the "fuzzy" classification between the right neighborhood names and better analyze how the neighborhoods can be classified wrong between the right neighborhood names which improves the learning of how a correct neighborhood name can be written wrong.



	NO_LOCALIDADE_FAM	JD_value	JD_class	RANK
52	8BAIRRO ALTO	0.09090909090909094	BAIRRO ALTO	1
53	8BAIRRO ALTO	0.736842105263158	ALTO BOQUEIRAO	2
54	8BAIRRO ALTO	0.8333333333333334	BARREIRINHA	3
55	8BOQUEIRAO	0.11111111111111116	BOQUEIRAO	1
56	8BOQUEIRAO	0.4285714285714286	ALTO BOQUEIRAO	2
57	8BOQUEIRAO	0.7727272727272727	CAMPINA DO SIQUEIRA	3

Figure 20 This image examples the rank classification of the three first dictionary choices for a name correction.

6. EVALUATION

Following the perspective from Hlàdek (2020) and Carneiro De Araújo et al., n.d., were found define three main evaluation methods:

- Accuracy, precision, recall
- BLEU Score

To use accuracy, precision and recall demanded to be clear which would be correct or wrong answers for each neighborhood name. The database provided an overall and do not have a classification of which would be the correct name for each wrongly typed name. For that reason, I decided to start the analysis with Damerau-Levenshtein, an edit distance, which can create a criterion to classify names through actions (insert, delete, replace and transpose) between two strings and increasing the change of error and followed by the Jaccard Distance, vector space model, which decreases the sample space of the possible words.

Therefore, the Damerau-Levenshtein returns more possible ways to type wrongly a neighborhood name and the Jaccard Distance shrinks the possible classification. Due to it the evaluation of the first one will be done only a manual accuracy, where will be manually counted if the algorithm returned the name right or wrong. Then the evaluation of the Jaccard Distance also will be done manually but will compare the results to the Damerau-Levenshtein classification to consider the precision, and recall. The BLEU Score will be applied in both methods separately. Finally, to create the final product of this project will be considered the evaluation of the combination of both methods.

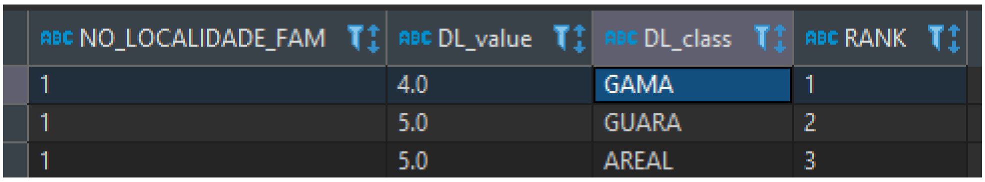
Based on Angell R, 1983 research, the analysis will be made in only three cities: Brasília (BSB), Curitiba (CWB), and Recife (REC). Brasília and Curitiba due to the knowledge of the research of this study about the city, identified extra errors in the autocorrection and Recife due to the recent flood which uses the methodology developed here in their database to define future social solutions.

The classification of each neighborhood was done by the best similarity score for each method, the lowest value higher is the similarity between words, followed by the weight of poverty families in the neighborhood.

Damerau-Levenshtein – Accuracy

Since is not possible to define what was the exact neighborhood expected to be written in each error, the sample space is not clearly defined. Therefore, some rules were defined to guarantee the pattern of similarity between words.

- Names typed as numbers they were considered all wrong (Figure 21).

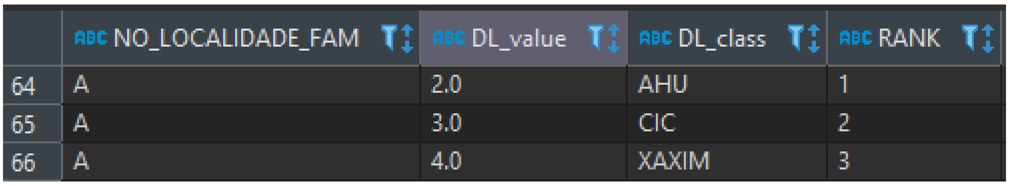


NO_LOCALIDADE_FAM	DL_value	DL_class	RANK
1	4.0	GAMA	1
1	5.0	GUARA	2
1	5.0	AREAL	3

Figure 21 Names of neighborhoods typed as numbers, an example from Brasília.

They were not filtered in a pre-processing step because this project is analyzing only three cities of 5570. It is not possible to prevent if one of the cities has a neighborhood written as a number and since this project has the goal to be applied in all cities made more sense to keep them.

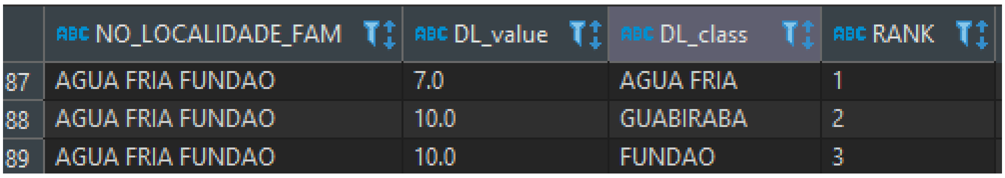
- Names until two characters were considered wrong (Figure 22)



	NO_LOCALIDADE_FAM	DL_value	DL_class	RANK
64	A	2.0	AHU	1
65	A	3.0	CIC	2
66	A	4.0	XAXIM	3

Figure 22 Neighborhood names with less than two characters, an example from Curitiba.

- Names that were written two neighborhood names were considered wrong because it is not possible to define which of them would be right (Figure 23)

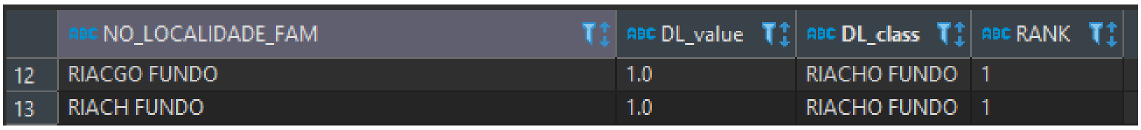


	NO_LOCALIDADE_FAM	DL_value	DL_class	RANK
87	AGUA FRIA FUNDAO	7.0	AGUA FRIA	1
88	AGUA FRIA FUNDAO	10.0	GUABIRABA	2
89	AGUA FRIA FUNDAO	10.0	FUNDAO	3

Figure 23 Neighborhood names written with two neighborhood names, an example from Recife.

- In the case of the reference table from the city a neighborhood can have similar names, but different extensions becoming a composed name they had more specific analysis, for example, Riacho Fundo, Riacho Fundo I, and Riacho Fundo II, Brasilia's neighborhoods.

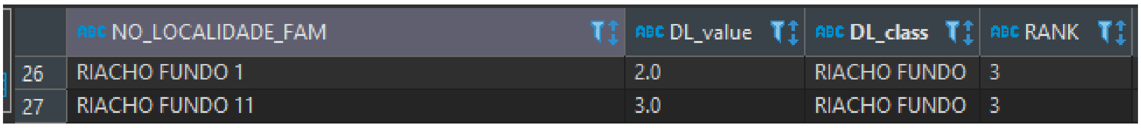
The first case would only consider right if were only typed two words and the words were in a sentence not followed by a similar meaning to "I" and "II", or the words suffered any type of action such as delete, insertion, replace or transposition (Figure 24).



	NO_LOCALIDADE_FAM	DL_value	DL_class	RANK
12	RIACGO FUNDO	1.0	RIACHO FUNDO	1
13	RIACH FUNDO	1.0	RIACHO FUNDO	1

Figure 24 Examples of misspelled neighborhood name "Riacho Fundo" and the equivalent correction to Riacho Fundo with the Damerau-Levenshtein method.

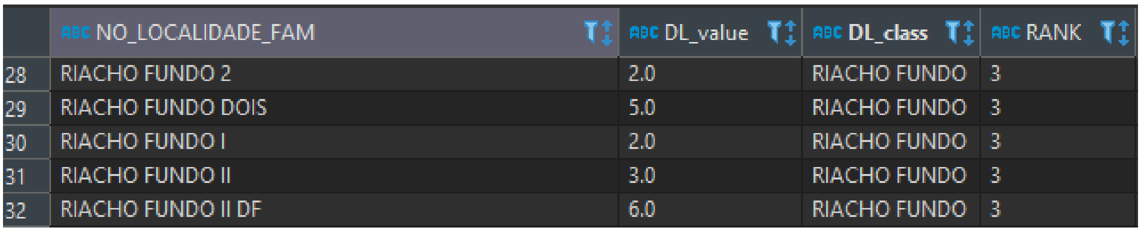
In the second case, the right correction was only the case where it was 3 words but the last one was "I", "1" or related to the number 1. And still considering the actions of delete, insertion, replace, or transposition in the first two words (Figure 25).



	NO_LOCALIDADE_FAM	DL_value	DL_class	RANK
26	RIACHO FUNDO 1	2.0	RIACHO FUNDO	3
27	RIACHO FUNDO 11	3.0	RIACHO FUNDO	3

Figure 25 Examples of misspelled neighborhood name "Riacho Fundo I" and the equivalent correction of Riacho Fundo I with the Damerau-Levenshtein method.

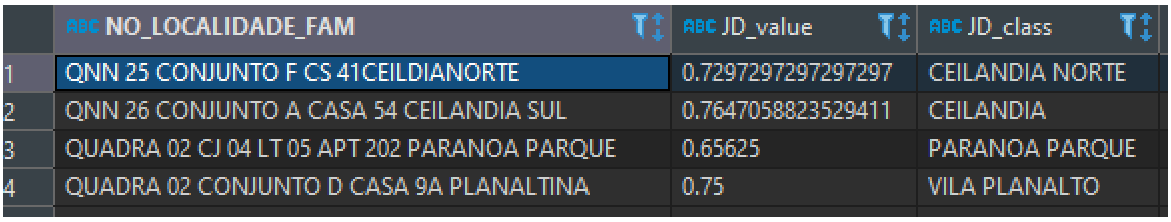
The third case was like the second case, but the third word was “II”, “2” or related to the number two. Keep considering the actions of delete, insertion, replace, or transposition in the first two words (Figure 26).



	NO_LOCALIDADE_FAM	DL_value	DL_class	RANK
28	RIACHO FUNDO 2	2.0	RIACHO FUNDO	3
29	RIACHO FUNDO DOIS	5.0	RIACHO FUNDO	3
30	RIACHO FUNDO I	2.0	RIACHO FUNDO	3
31	RIACHO FUNDO II	3.0	RIACHO FUNDO	3
32	RIACHO FUNDO II DF	6.0	RIACHO FUNDO	3

Figure 26 Examples of misspelled neighborhood name “Riacho Fundo II” and the equivalent correction of Riacho Fundo II with the Damerau-Levenshtein method.

- Addresses written on the neighborhood space were considered if it has the neighborhood name typed (Figure 27).



	NO_LOCALIDADE_FAM	JD_value	JD_class
1	QNN 25 CONJUNTO F CS 41CEILDIANORTE	0.7297297297297297	CEILANDIA NORTE
2	QNN 26 CONJUNTO A CASA 54 CEILANDIA SUL	0.7647058823529411	CEILANDIA
3	QUADRA 02 CJ 04 LT 05 APT 202 PARANOA PARQUE	0.65625	PARANOA PARQUE
4	QUADRA 02 CONJUNTO D CASA 9A PLANALTINA	0.75	VILA PLANALTO

Figure 27 Examples of addresses written on the neighborhood namespace.

These cases were considered due to the possibility to identify the neighborhood and understanding the behavior of the algorithm to find and correct the error.

These considerations were relevant in all city’s neighborhoods, thereby were defined the sample of each Damerau-Levenshtein correction and applying accuracy evaluation. Table 1 presents the results of the accuracy analysis on cities of analysis.

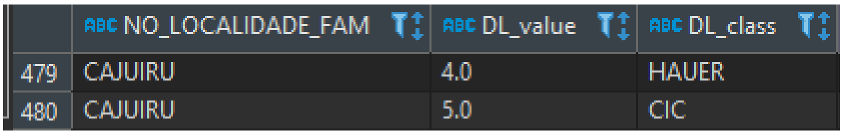
Table 1 Accuracy results of the Damerau-Levenshtein method detailed by city and the average result of all.

City	Accuracy Score
Brasília	0.2337
Curitiba	0.2932
Recife	0.3139
Average	0.2802

Through these results, it was possible to come up with some conclusions:

- Is not possible to predict which right name is expected to be typed in each wrong spelling

- Edit Distance method counts the actions to make one word become another, depending on the error a wrongly typed name can become part of a sample of another neighborhood even though it is close to another group (Figure 28).

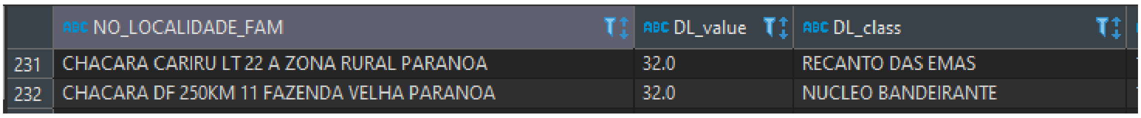


	NO_LOCALIDADE_FAM	DL_value	DL_class
479	CAJUIRU	4.0	HAUER
480	CAJUIRU	5.0	CIC

Figure 28 Examples of the wrong classification of names by the Damerau-Levenshtein method, which even though the Damerau-Levenshtein coefficient (DL_value) is low, classifies wrong. These examples were supposed to return “Cajuru” from the dictionary table.

These two examples were classified into totally different neighborhoods besides it was clear, the supposed name from the Dictionary table was “CAJURU”.

- Address typed with the neighborhood name in it, which were not able to identify (Figure 29).



	NO_LOCALIDADE_FAM	DL_value	DL_class
231	CHACARA CARIRU LT 22 A ZONA RURAL PARANOA	32.0	RECANTO DAS EMAS
232	CHACARA DF 250KM 11 FAZENDA VELHA PARANOA	32.0	NUCLEO BANDEIRANTE

Figure 29 Examples of the wrong classification of names by the Damerau-Levenshtein method, which is not able to identify the neighborhood name in the address even though it is written. These examples were supposed to return “Paranoa” from the dictionary table.

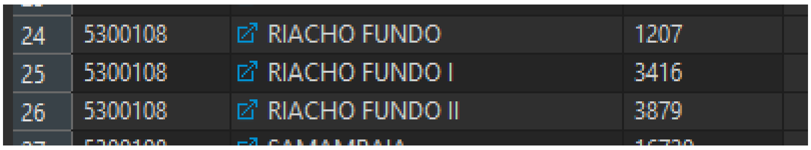
These examples they supposed to be classified as “Paranoá”, but due to the long string, they were mixed to compose names.

Given the logic and results of Damerau-Levenshtein, were decided to follow the research using the results to create a sample space for Jaccard Distance. For each classification of the neighborhood by Damerau-Levenshtein, were analyzed the results of Jaccard Distance classification and calculated the accuracy, precision, and recall of the vector space method.

Jaccard Distance – Accuracy, Precision, Recall

Jaccard Distance is a statistical similarity method. Due to it, to identify the True Positive, False Positive and False Negative classifications were defined a few criteria:

- Jaccard Distance was applied in bigrams of the words, so the similarity is given by the number of bigrams between words
- If exists similar neighborhood names in the reference table, the similarity was considered by the number of the same letters or words in each neighborhood name (Figure 30).



24	5300108	RIACHO FUNDO	1207
25	5300108	RIACHO FUNDO I	3416
26	5300108	RIACHO FUNDO II	3879

Figure 30 Examples of similar neighborhood names of the same city, an example of Brasília.

In Brasília exists three neighborhoods with almost the same name. So given the logic from Jaccard Distance, the results like “RIACHU FUNDI I” can be corrected by “RIACHO FUNDO”, but “RIACHU

FUNDI” cannot be corrected by “RIACHO FUNDO I”. The Vann diagram (Figure 31) represents part sample space occupied by the name in the example.

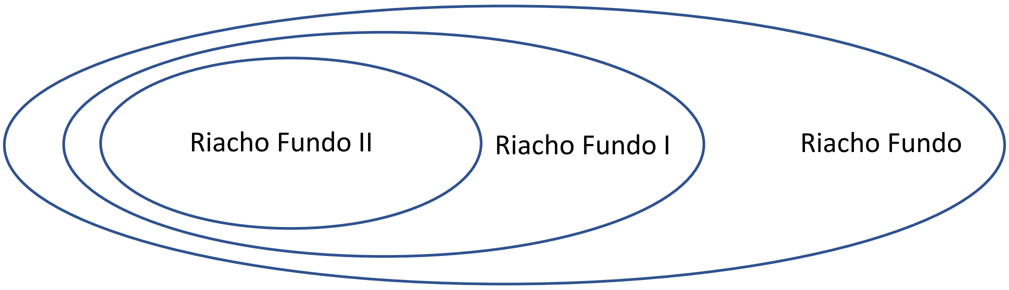
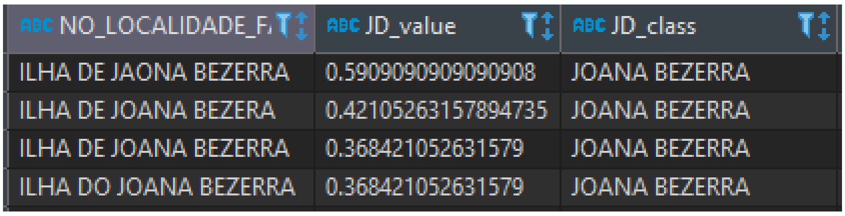


Figure 31 Vann Diagram of Riacho Fundo, Riacho Fundo I, and Riacho Fundo II neighborhood help the comprehension of how their name is related to one another.

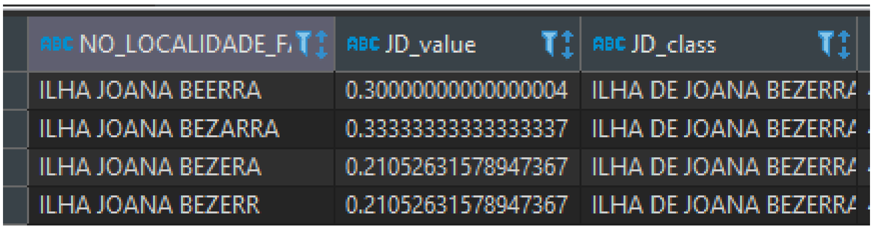
- Neighborhood names that create other ones were considered correct classified, another way around wrong classified (Figure 32).



NO_LOCALIDADE_F,	JD_value	JD_class
ILHA DE JAONA BEZERRA	0.5909090909090908	JOANA BEZERRA
ILHA DE JOANA BEZERA	0.42105263157894735	JOANA BEZERRA
ILHA DE JOANA BEZERRA	0.368421052631579	JOANA BEZERRA
ILHA DO JOANA BEZERRA	0.368421052631579	JOANA BEZERRA

Figure 32 Example of composed names in which one of the names has another neighborhood name from the dictionary table of the city, an example of Recife.

Here the classification is right, because “JOANA BEZERRA” is part of “ILHA DE JOANA BEZERRA”.



NO_LOCALIDADE_F,	JD_value	JD_class
ILHA JOANA BEERRA	0.30000000000000004	ILHA DE JOANA BEZERRA
ILHA JOANA BEZARRA	0.33333333333333337	ILHA DE JOANA BEZERRA
ILHA JOANA BEZERA	0.21052631578947367	ILHA DE JOANA BEZERRA
ILHA JOANA BEZERR	0.21052631578947367	ILHA DE JOANA BEZERRA

Figure 33 Example of composed names for which the classification was considered wrong even though they seem similar, because in the dictionary table of Recife “ILHA JOANA BEZERRA” and “ILHA DE JOANA BEZERRA” were two different neighborhoods.

Here the classification is wrong, because “ILHA DE JOANA BEZERRA” need to have a “DE”, which does not present in the neighborhood type.

- All the neighborhoods were considered unique and independent about the table, even though in real life they represent the same neighborhood.

The results found in Jaccard Distance are presented in table 2.

Comparing the results with Damerau-Levenshtein, the accuracy of the method was lower. However, due to the sample space defined was possible to calculate the Precision and Recall. The precision shows that the algorithm does not tend to misclassify the neighborhood due to the high similarity between names as explained above. Meanwhile, the Recall had a high score, showing that the method has not missed many names that were supposed to be classified and were not (false negatives).

Table 2 Accuracy, Precision, Recall results of Jaccard Distance method detailed by city and the average result of all.

City	Accuracy	Precision	Recall
Brasília (BSB)	0.22	0.62	0.92
Curitiba (CWB)	0.27	0.73	0.98
Recife (REC)	0.27	0.66	0.94
Average	0.2574	0.6729	0.9488

DAMERAU-LEVENSHTEIN AND JACCARD DISTANCE – BLEU SCORE

Following the basic evaluation of classifications, it was also done the BLEU Score. This evaluation computes the precision of tokens from the candidate chosen in the reference and penalizes the candidates that appear more times than expected.

This analysis will be compared the typed neighborhood with the name classified by the algorithms. The names will be split in bigrams, but some adjustments were considered for the research case:

- The result will be the fraction of candidates' bigrams in the typed neighborhood over the bigrams in the name classified in the reference table;
- If a bigram appears more time in the typed neighborhood than in the name classified, it will suffer a penalty;
- If a bigram does not exist in the name classified, it will suffer a penalty.

Jaccard Distance presented a better result for classification in Brasília than Damerau-Levenshtein (Figures 34 and 35).

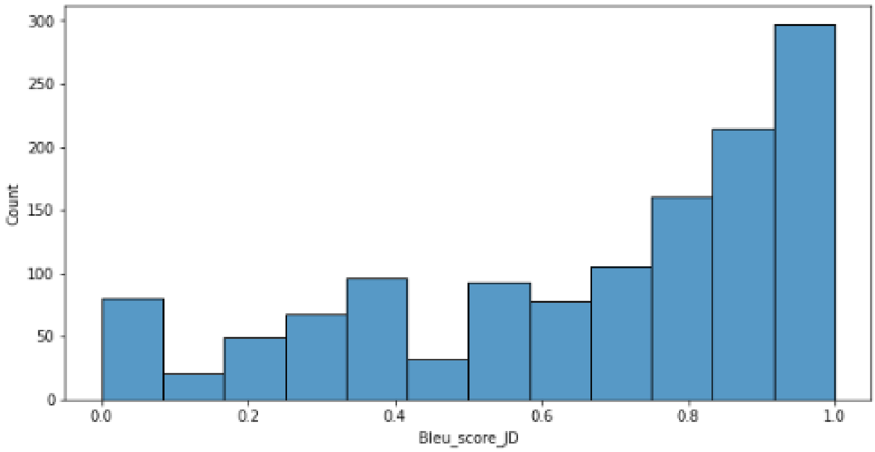


Figure 34 BLEU Score of Jaccard Distance method.

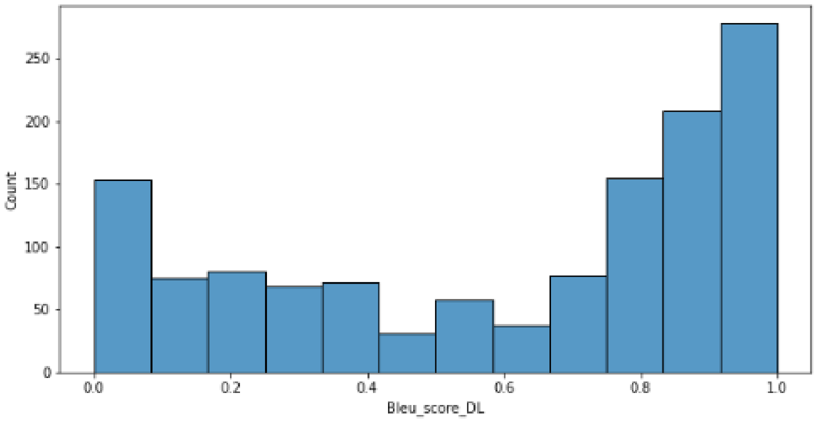
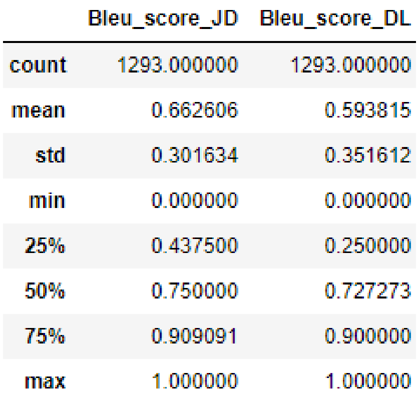


Figure 35 BLEU Score of Damerau-Levenshtein method.

The first plot represents the results of Jaccard Distance, in which most of the comparisons are above 60% of similarity with the reference. Meanwhile, the Damerau-Levenshtein had many comparisons that were below 20% of similarity with the reference.

Figure 36 shows that the similarity of Jaccard Distance was higher and more precise than Damerau-Levenshtein.



	Bleu_score_JD	Bleu_score_DL
count	1293.000000	1293.000000
mean	0.662606	0.593815
std	0.301634	0.351612
min	0.000000	0.000000
25%	0.437500	0.250000
50%	0.750000	0.727273
75%	0.909091	0.900000
max	1.000000	1.000000

Figure 36 Statistic comparison of BLEU Score between the methods Jaccard Distance (Bleu_score_JD) and Damerau-Levenshtein (Bleu_score_DL).

DAMERAU-LEVENSHTEIN AND JACCARD DISTANCE – ACCURACY

After all, results, using both methods would be the best way to reach the right correction. Then it was decided to calculate the accuracy of this combination.

Table 3 Accuracy results of the combination of both methods.

City	Accuracy Score
Brasília	0.64
Curitiba	0.76

Recife	0.66
Average	0.67

Table 3 shows that considering both methods to classify the name, the accuracy increased by almost 60%.

7. RESULTS

This research developed an Automatic Dictionary Creator, presented in section 5. The algorithm found the proper neighborhood names of a city by the methodology proposed.

Table 4 presents a precise number of neighborhoods per city after the application of the algorithm. Given the cities in the analysis, at least 90% of the neighborhood names are wrongly typed.

Table 4 Comparison of the number of unique neighborhoods before and after the algorithm application.

City	Number of unique neighborhoods in CECAD	Number of unique neighborhoods after the algorithm
Brasília	1293	38
Curitiba	678	67
Recife	1560	80

The adjustments of the ASC system and its respective evaluation presented in section 6, conclude that the best string similarity method would be the combination of Damerau-Levenshtein and Jaccard Distance for the present problem. Due to this, Table 5 shows the results of the ASC system in Brasília, Curitiba, and Recife.

Table 5 The general and detailed data of the neighborhood families before and after the application of the algorithm.

City	Previous Families Mapped	After processing Families Mapped	Difference (families)
Brasília	166483	169274	2791
Curitiba	129309	130860	1551
Recife	212237	216754	4517
General	508029	516888	8859

It was possible to add almost nine thousand new families as shown below in the analysis of the three cities, considering only the analysis of 3 cities out of 5570. Curitiba was the city in which the algorithm had less impact, probably related to social education development from the region. However, Recife showed relevant growth, more families were mapped.

The result represents a considerable number of families included, which will impact the development of social projects. Also, it is important to highlight that family is a group of people in low-income situations living together, so this project has a bigger impact than only one perspective.

8. CONCLUSION

Brazil is a developing country that currently is suffering social impacts, one of them being the reality of low-income families. Many families do not have income for basic needs such as education, health security, or food also these same families usually do not live in legal places and start neighborhoods called “Subnormal agglomerates”, known as “*favelas*”.

Therefore, government programs such as the Unified Registry help to find families in critical social reality and promote social solutions for them. However, with their sensible access to technology and “society”, the registration of these families keeps being a challenge because part of the process is still manual. Because of this many addresses are registered wrongly and delivered wrong analysis.

A solution was presented in this dissertation; the Automatic Spelling Corrector system can optimize solving this problem. The Brazilian Government is developing a cash transfer program called “calamity help for needy families”, so the current project has the power to support the analysis of the implementation of the social program. With the ASC system created the results above lead to a higher precision of the neighborhood data besides redirecting the social money wisely.

From a technical perspective, the Automatic Spelling Correction algorithm can be a solution to correct wrongly typed neighborhood names. However, it is essential to point out a few necessary improvements to reach an optimal result:

- The dictionary table can be improved with external help from CRAS staff. If the user defines and validates all the neighborhoods inserted on the dictionary table along with the possibility to manually add neighborhoods that were not considered relevant by the mathematical perspective, these could improve the accuracy of the algorithm.
- Define which data can be considered outliers for this analysis.

Since the Brazilian Government is developing the program “calamity help for needy families”, the dissertation has the power to support the results. Through this project, the results above promote a higher precision data accuracy with the help of the algorithm automatizing part of the process. Moreover, the manual support of a user can achieve greater results in further projects.

Possible further developments for this project are:

- Relate the neighborhoods with georeferenced data to validate the location of each neighborhood and analyze the real needs based on mapped basic services.
- Create an incremental process to facilitate the application of the algorithm in future data as of 2022 and the future as shown in figure 37.

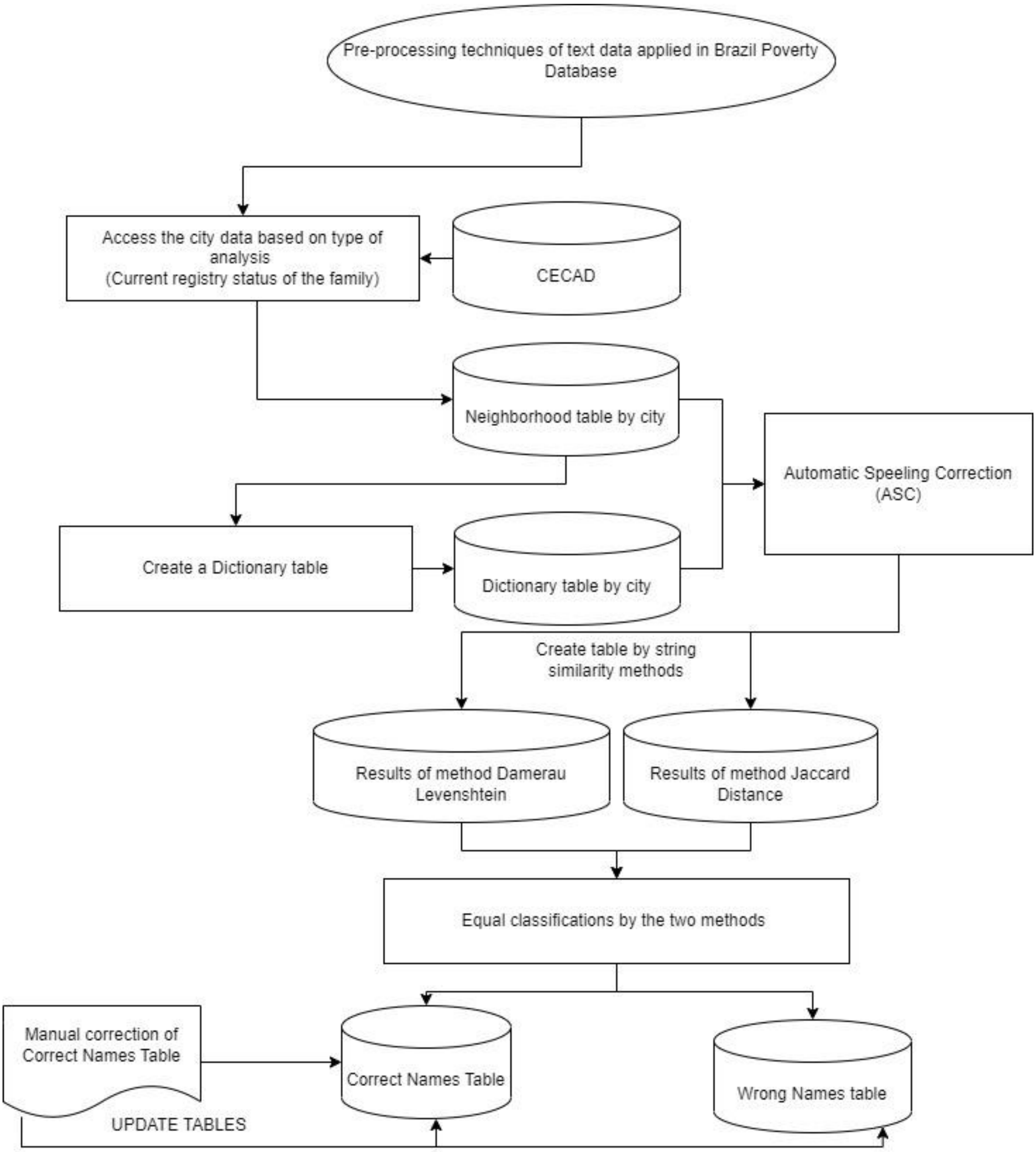

Pre-processing techniques of text data applied in Brazil Poverty Database
Access the city data based on type of analysis (Current registry status of the family)
CECAD
Neighborhood table by city
Automatic Speeling Correction (ASC)
Create a Dictionary table
Dictionary table by city
Create table by string similarity methods
Results of method Damerau Levenshtein
Results of method Jaccard Distance
Equal classifications by the two methods
Manual correction of Correct Names Table
Correct Names Table
Wrong Names table
UPDATE TABLES

Figure 37 Flowchart with the incremental process.

9. REFERENCES

Angell, R. C., Freund, G. E., & Willett, P. (1982). *AUTOMATIC SPELLING CORRECTION USING A TRIGRAM SIMILARITY MEASURE*.

Barca, V. (2017). *Integrating Data and Information Management for Social Protection: Social Registries and Integrated Beneficiary Registries*. www.itsanhonour.gov.au/coat-arms/index.cfm.

Carneiro De Araújo, L., de Lima Benevides, A., & Sansão, J. P. (n.d.). *Linguagem e Tecnologia Seção: Linguística e Tecnologia Desenvolvimento de um corretor ortográfico Developing a spell checker*. https://doi.org/10.35699/1983

Carolina Mesquita. (2021, October 25). *Mercado da miséria: frigoríficos vendem ossos de primeira e de segunda na periferia de Fortaleza*. https://diariodonordeste.verdesmares.com.br/negocios/mercado-da-miseria-frigorificos-vendem-ossos-de-primeira-e-de-segunda-na-periferia-de-fortaleza-1.3151320

Castro, J., & Modesto, L. (2010). *Bolsa Família 2003 – 2010: avanços e desafios* (2 v.). IPEA.

Chaves, J. (2021). *A identificação de público-alvo para as políticas de combate à pobreza: o caso do cadastro único e seu uso pelas coalizões de defesa* [Doctoral Thesis]. Universidade de Brasília .

Dubin, D. (2014). *The Most Influential Paper Gerard Salton Never Wrote*. https://www.researchgate.net/publication/32956274

Feitosa, P. (2010). *O Cidadão Codificado: A Digitalização da Cidadania em Banco de Dados de Interesse Público* [Master's Thesis]. Universidade Federal do Rio de Janeiro/COPPE.

Ferreira, R. (2009). SUAS - Sistema Único de Assistência Social. In *Ministério do Desenvolvimento Social e Combate à Fome*.

Gallagher, E. D. (1999). *COMPAH DOCUMENTATION*. The University of Massachusetts at Boston.

Gioras Xerez. (2021, October 18). *Moradores coletam comida em caminhão de lixo em Fortaleza; vídeo*. https://g1.globo.com/ce/ceara/noticia/2021/10/18/moradores-coletam-comida-em-caminhao-de-lixo-em-fortaleza.ghtml

Governo Brasileiro. (2022, July 9). *Participar de Serviços da Proteção Social Básica - programas e benefícios assistenciais*. Participar de Serviços Da Proteção Social Básica - Programas e Benefícios Assistenciais. https://www.gov.br/pt-br/servicos/participar-de-servicos-da-protecao-social-basica-programas-e-beneficios-assistenciais

Governo do Estado Paraná. (2022, October 12). *Serviços de Proteção Social Básica*. Serviços de Proteção Social Básica. https://www.justica.pr.gov.br/Pagina/Servicos-de-Protecao-Social-Basica

Henrique, G. (2020). *Ajuntamds*. Ajuntamds

Hládek, D., Staš, J., & Pleva, M. (2020). Survey of automatic spelling correction. *Electronics (Switzerland)*, *9*(10), 1–29. https://doi.org/10.3390/electronics9101670

Hussain, A. (2012). *Textual Similarity* [Technical University of Denmark]. www.imm.dtu.dk

Insituto Brasileiro de Geografia e Estatistica (IBGE). (2022, October 12). *Aglomerado Subnormal* . Aglomerado Subnormal. https://www.ibge.gov.br/geociencias/organizacao-do-territorio/tipologias-do-territorio/15788-aglomerados-subnormais.html?=&t=o-que-e

Jacobs, N. (2004). *Relational Sequence Learnind and User Modelling* [Doctoral Thesis]. Katholienke Universiteit Leuven.

James M, J. I. (2022, June 22). *Damerau-Levenshtein Edit Distance Explained*. https://www.lemoda.net/text-fuzzy/damerau-levenshtein/

Jornal Nacional. (2022, June 8). *Mais de 33 milhões de brasileiros passam fome todo dia, revela pesquisa*. https://g1.globo.com/jornal-nacional/noticia/2022/06/08/mais-de-33-milhoes-de-brasileiros-passam-fome-todo-dia-revela-pesquisa.ghtml

Koga, D. (2015). *Território de vivência em um país continental* (Vol. 14, Issue 1).

Kosub, S. (2016). *A note on the triangle inequality for the Jaccard distance*. http://arxiv.org/abs/1612.02696

Minisitério da Cidadania. (2015, June 22). *Centro de Referência de Assistência Social - Cras*. Centro de Referência de Assistência Social - Cras. http://mds.gov.br/assuntos/assistencia-social/unidades-de-atendimento/cras

Ministério do Desenvolvimento Social. (2022). *CECAD*. https://cecad.cidadania.gov.br/

Ministério do Desenvolvimento Social e Combate à Fome. (2014, July). *Relatório indica que Brasil saiu do Mapa Mundial da Fome em 2014*. https://www.gov.br/casacivil/pt-br/assuntos/noticias/2014/setembro/relatorio-indica-que-brasil-saiu-do-mapa-mundial-da-fome-em-2014

Nakashima, C. (2012). *Disponibilizando Informação para Desenho e Fomento de Políticas Sociais no Brasil: o Caso do CECAD*.

Paes De Barros, R., de Carvalho, M., & Mendonça, R. (2009). *Sobre as utilidades do Cadastro Único*.

Pedro Rafael Vilela. (2022, June 15). *Entenda por que a assistência social no Distrito Federal está tão caótica*. https://www.brasildefato.com.br/2022/06/15/entenda-por-que-a-assistencia-social-no-distrito-federal-esta-tao-caotica

Peret, E., Neto, J., Loschi, M., Lima, A., Guimarães, A., Perissé, C., Tallmann, H., Zasso, J., Grizoli, L., Santos, L., & Carlôto, P. (2018, February 20). *Favelas resistem e propõem desafios para urbanização*. https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/20080-favelas-resistem-e-propoem-desafios-para-urbanizacao

Petini, S., das Chagas, M., Ferreira, J. D., & Tramarin, Â. M. (2021). *A importância do Cadastro Único na vida das famílias cadastradas no município de Colíder-MT no período de 2010 a 2018 uma perspectiva de inclusão social*.

Prachi Kumar. (2017, October 21). *An Introduction to N-grams: What Are They and Why Do We Need Them?* https://blog.xrds.acm.org/2017/10/introduction-n-grams-need/

Sambiase, A., Oliveira, B., Bastos, B., & Curralero, C. (2015). *Manual de Gestão do Cadastro Único para programas Sociais do Governo Federal* (D. Tavares, M. Nogueira, & G. Sousa, Eds.; 2nd ed.). Ministério do Desenvolvimento Social e Combate a fome .

Secretaria de Estado de Governo do Distrito Federal. (2022, November 3). *Administrações Regionais*. Administrações Regionais. https://segov.df.gov.br/category/administracoes-regionais/

Torres, J. (2010). *O CadÚnico na Identificação e Classificação Social de Quem são os Pobres do Brasil* [Master's Thesis]. Universidade Federal da Bahia.

World Without Poverty. (2022, October 4). *Unified Registry*. http://wwp.org.br/en/social-policy/unified-registry/

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa