|  |  |
| --- | --- |
|  | **UNIVERSIDADE DE SÃO PAULO**  **Instituto de Ciências Matemáticas e de Computação**  **Departamento de Sistemas de Computação** |

Mineração de Textos de Notas Fiscais Eletrônicas do Programa Nacional de Alimentação Escolar (PNAE)

***Nome do aluno***

***[Nome do Aluno]***

***Matheus Resende Miranda***

São Carlos - SP

Mineração de Textos de Notas Fiscais Eletrônicas do Programa Nacional de Alimentação Escolar (PNAE)

***Matheus Resende Miranda***

Orientador: Ricardo Marcondes Marcacini

|  |
| --- |
| Monografia referente ao projeto de conclusão de curso dentro do escopo da disciplina SSC0670 - Projeto de Formatura I do Departamento de Sistemas de Computação do Instituto de Ciências Matemáticas e de Computação – ICMC-USP para obtenção do título de Engenheiro de Computação. |
| Área de Concentração: Mineração de Textos |

**USP – São Carlos**

**2022**

*“A tecnologia move o mundo” – Steve Jobs*

Resumo

O Programa Nacional de Alimentação Escolar (PNAE) é uma política pública pautada na distribuição de recursos a entidades participantes a fim de promover a alimentação dos estudantes dessas instituições. Assim como diversas outras políticas, o PNAE é um programa de difícil fiscalização e monitoramento. O objetivo central deste trabalho, que estende e utiliza outros projetos, como a Tabela Brasileira de Composição de Alimentos (TACO), é mostrar a possibilidade de uso da mineração de textos como ferramenta de monitoramento para esse programa a partir das Notas Fiscais Eletrônicas emitidas pelos municípios em relação ao PNAE. O processo de mineração de textos envolve desde a coleta das notas fiscais eletrônicas de repasses do programa, a identificação dos produtos alimentares na descrição das notas, e o mapeamento automático desses produtos na tabela TACO para entender como os recursos desses repasses estão sendo administrados. Dessa forma foi construído um indicador baseado em mapas de calor de um estado exemplo, com a proposta de fiscalizar de maneira visual se os recursos do PNAE foram ou estão sendo bem gastos, de acordo com uma métrica de pontuação pré-estabelecida.

Índice

[Lista de Abreviaturas/Siglas v](#_Toc119424113)

[Lista de Tabelas vi](#_Toc119424114)

[Lista de Figuras vii](#_Toc119424115)

[CAPÍTULO 1: INTRODUÇÃO 1](#_Toc119424116)

[1.1. Contextualização e Motivação 1](#_Toc119424117)

[1.2. Objetivos 2](#_Toc119424118)

[1.3. Organização do Trabalho 2](#_Toc119424119)

[CAPÍTULO 2: REVISÃO BIBLIOGRÁFICA 4](#_Toc119424120)

[2.1. Considerações Iniciais 4](#_Toc119424121)

[2.2. Conceitos e Técnicas relevantes 4](#_Toc119424122)

[2.2.1. Pré-processamento de textos 4](#_Toc119424123)

[2.2.2. Similaridade com NFe 7](#_Toc119424124)

[2.2.3 Tabela Brasileira de Composição de Alimentos (TACO) 8](#_Toc119424125)

[2.4. Considerações Finais 9](#_Toc119424126)

[CAPÍTULO 3: DESENVOLVIMENTO DO TRABALHO 10](#_Toc119424127)

[3.1. Considerações Iniciais 10](#_Toc119424128)

[3.2. Base de NFe 10](#_Toc119424129)

[3.3. Mapeamento entre NFe e TACO 12](#_Toc119424130)

[3.4. Resultados Obtidos 14](#_Toc119424131)

[3.4.1 Informações nutricionais dos municípios 14](#_Toc119424132)

[3.4.2 Indicadores por meio de Mapas de Calor 16](#_Toc119424133)

[3.5. Dificuldades e Limitações 20](#_Toc119424134)

[3.6. Considerações Finais 21](#_Toc119424135)

[CAPÍTULO 4: CONCLUSÃO 22](#_Toc119424136)

[4.1. Contribuições 22](#_Toc119424137)

[4.2. Considerações sobre o Curso de Graduação 22](#_Toc119424138)

[4.3. Trabalhos Futuros 23](#_Toc119424139)

[REFERÊNCIAS 24](#_Toc119424140)

[APÊNDICE A – Exemplo de Transposição (*Ts,t*) 25](#_Toc119424141)

# Lista de Abreviaturas/Siglas

CBOW *Continuous Bag of Words*

NFes Notas Fiscais eletrônicas

PNAE Programa Nacional de Alimentação Escolar

TACO Tabela Brasileira de Composição de Alimentos

# Lista de Tabelas

[**Tabela 1 – NFes captadas por estado brasileiro** 11](#_Toc119424112)

# Lista de Figuras

[**Figura 1 – Conversão de sequências em multidimensional embedding** 6](#_Toc119424101)

[**Figura 2 – Modelo CBOW para entrada de contexto (X) e saída da distribuição de probabilidades das possíveis palavras centrais (Y)**. 7](#_Toc119424102)

[**Figura 3 - Exemplo de NFe** 11](#_Toc119424103)

[**Figura 4 – DataFrame da tabela TACO** 13](#_Toc119424104)

[**Figura 5 – DataFrame gerado a partir das NFes do estado de SP** 13](#_Toc119424105)

[**Figura 6 – DataFrame com itens das NFes relacionados com seus dados nutricionais** 14](#_Toc119424106)

[**Figura 7 - Quantidade de amostras das NFe acima de um valor mínimo de Jaro-Winkler para o estado de SP** 15](#_Toc119424107)

[**Figura 8 – Mapa focado no estado de SP com os municípios participantes do PNAE** 16](#_Toc119424108)

[**Figura 9 – Mapa de calor do estado de SP** 17](#_Toc119424109)

[**Figura 10 - Quantidade de amostras retiradas das NFes de acordo com o ano para o estado de SP** 18](#_Toc119424110)

[**Figura 11 - Mapas de calor do estado de SP de 2014 até 2016** 19](#_Toc119424111)

# CAPÍTULO 1: INTRODUÇÃO

## 1.1. Contextualização e Motivação

O Programa Nacional de Alimentação Escolar (PNAE) é uma iniciativa do governo federal que oferece alimentação escolar e ações de educação alimentar e nutricional a estudantes matriculados em escolas públicas, filantrópicas e entidades comunitárias conveniadas com o poder público, desde a educação infantil até o ensino médio, por meio de repasses a estados, municípios e instituições federais. O governo federal fornece recursos para a cobertura de 200 dias letivos, os quais variam de acordo com o número de pessoas matriculadas na rede de ensino [FNDE 2022].

Os impactos de políticas públicas na sociedade são julgados a partir do monitoramento e avaliação de como os recursos são distribuídos e utilizados, bem como quais foram as mudanças sociais e econômicas resultantes dessa ação pública [PEDONE 1986]. O PNAE, por sua vez, é acompanhado e fiscalizado pelo Conselhos de Alimentação Escolar, FNDE, Tribunal de Contas da União, Controladoria Geral da União e pelo Ministério Público [FNDE 2022]. Porém, o monitoramento e avaliação de uma política de desenvolvimento social como essa é um desafio, uma vez que um programa de nível nacional abrange uma quantidade massiva de instituições que realizam diversas movimentações com os recursos repassados. Uma alternativa criada pelo Ministério da Educação foi o Sistema de Gestão de Prestação de Contas (SiGPC) que faz a gestão de recursos repassados aos estados, municípios e Distrito Federal pelo PNAE de forma transparente [FNDE 2022]. Nesse sistema é disponibilizado ao usuário, por exemplo, documentos relacionados ao uso desses recursos, como as Notas Fiscais eletrônicas (NFe), que trazem detalhes da instituição e da compra de alimentos feita pela entidade, descrevendo os itens e os valores gastos.

A disponibilização de NFe é de fato uma proposta transparente para se prestar contas a respeito da organização orçamentária e da descrição dos alimentos adquiridos pelas entidades executoras, porém pode-se tornar ineficiente do ponto de vista da fiscalização do programa, uma vez que uma grande quantidade de documentos torna-se um imenso volume de informações, não necessariamente úteis, que na ausência de métodos computacionais eficazes não resulta em conhecimento aplicável, muito menos mostra um panorama geral de como a política pública está desempenhando. A mineração de textos, por sua vez, é uma área de pesquisa da ciência de computação que tenta resolver a crise de sobrecarga de informações existente nos tempos atuais por meio de técnicas e processos combinando mineração de dados, aprendizado de máquina e processamento de linguagem natural para a extração de informações e conhecimentos úteis a partir de coleções de documentos de texto [FELDMAN 2007].

Posto isso, este trabalho pretende aplicar técnicas de mineração de textos como uma forma de analisar NFe e mapear os municípios participantes do PNAE do ponto de vista nutricional. Isso propõe uma nova ferramenta para a fiscalização de políticas públicas, com o intuito de que essas técnicas possam ser cada vez mais refinadas e modificadas para serem aplicadas em demais programas sociais.

## 1.2. Objetivos

O objetivo principal deste trabalho é apresentar a mineração de textos como ferramenta para acompanhamento e fiscalização do PNAE, fazendo um mapeamento nutricional dos municípios que contenham instituições participantes do programa. Para isso será captado o maior número possível de NFe referente ao PNAE por meio do Sistema de Prestação de Contas OnLine implementado pelo Ministério da Educação, separadas por estado. Com as notas em mãos, será feito um pré-processamento dos dados (texto) presente em cada uma delas, sendo descartado os dados irrelevantes e estruturando os demais dados em *Dataframes*. Os dados organizados serão comparados com os dados de referência da Tabela Brasileira de Composição de Alimentos (TACO), por meio de medidas de similaridade. A partir desse cruzamento de informações, uma pontuação nutricional será definida possibilitando a geração de um mapa de calor que resultará em uma visualização de como os recursos repassados pelo programa PNAE estão sendo utilizados.

## 1.3. Organização do Trabalho

No Capítulo 2 é apresentado uma revisão bibliográfica referente à mineração de textos, como pré-processamento e técnicas de análise estrutural e semântica, e à medidas de similaridade, que comparam conjuntos de caracteres. Já o Capítulo 3 se refere ao desenvolvimento do projeto em si, retratando as atividades realizadas com as notas fiscais eletrônicas e com a tabela TACO, além dos resultados e mapeamentos obtidos. Por fim, no Capítulo 4 são realizadas as considerações finais, conclusões e trabalhos futuros.

# CAPÍTULO 2: REVISÃO BIBLIOGRÁFICA

## 2.1. Considerações Iniciais

Neste capítulo são apresentados alguns conceitos e técnicas sobre a mineração de textos, como o pré-processamento do texto bruto, seus modos de representação e a medida de similaridade de Jaro-Winkler, que compõem a essência do trabalho. É importante mencionar que alguns modelos matemáticos que eventualmente sejam citados não serão aprofundados, dado a alta complexidade e relevância perante os objetivos do projeto. Por fim, é abordada também a composição da tabela nutricional utilizada como referência para a análise do conteúdo das NFe.

## 2.2. Conceitos e Técnicas relevantes

### 2.2.1. Pré-processamento de textos

Em mineração de dados, o pré-processamento consiste em organizar o seu conjunto de dados de modo que a aplicação que irá utilizá-los, por exemplo, modelos de aprendizado de máquina, possa ter a melhor performance possível. Quando se trata de textos, alguns procedimentos são essenciais antes de processá-los [ZONG et al., 2021]. Primeiramente, é feita a extração de *tokens*, que são uma sequência de caracteres do texto os quais são tratados como uma unidade indivisível pelo processamento. Após a extração, é necessário fazer um processamento desses *tokens* consolidando palavras iguais e palavras presentes em diferentes tempos verbais em um único *token* e retirando a pontuação e as *stop words*, que sãopalavras como artigos, preposições, conjunções e pronomes, as quais ocorrem com muita frequência e acabam não sendo muito discriminativas para a mineração. Isso resulta em um conjunto de termos que podem ser dispostos como uma representação multidimensional, também chamada de *vector space representation*, contendo a frequência de cada termo no texto. Por fim, é interessante, ainda, fazer uma normalização dos valores das frequências a fim de equilibrar os pesos das palavras mais discriminativas [AGGARWAL 2018].

Para exemplificar o conceito, considere o seguinte texto:

“O gato corre atrás do rato. O cachorro corre atrás do gato”

Após o processo de tokenização e processamento, são obtidos os termos

{gato, corre, atrás, rato, cachorro}

Considerando a ordem mostrada dos termos como sendo a ordem do vetor, a representação multidimensional não normalizada do texto seria:

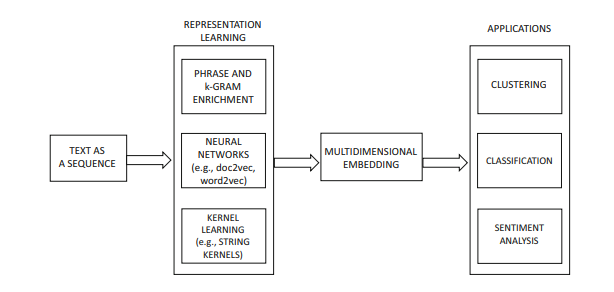
Visto como um texto pode ser representado, é possível tratar uma coleção de textos ou documentos, denominado *corpus*. Essas coleções de textos podem ser representadas como *bag-of-words*, que basicamente é uma representação esparsa e multidimensional com *n* documentos, ou como um conjunto de sequências de palavras por meio do qual é formada uma representação multidimensional da semântica das palavras no *corpus*, também chamada de *embeddings*.

O *corpus* como *bag-of-words* é a representação mais comum na mineração de textos. Nesse caso, a ordem que as palavras estão dispostas é ignorada e a coleção de textos é convertida em uma representação multidimensional e esparsa em que o universo dos termos corresponde às dimensões [AGGARWAL 2018]. Considerando um conjunto de *n* documentos que geram um universo de *d* termos, representados por {*w1, w2, ..., wd*} com suas respectivas frequências {*f11, f12, ..., f1n, ..., fdn*}, a sua representação como *bag-of-words* seria a matriz D de ordem *n* x *d*:

Essa representação na maior parte das vezes possui melhor rendimento no processamento quando normalizados, de maneira que as palavras com menores frequências ganhem mais peso na avaliação, pois geralmente são mais discriminativas. É possível, ainda alterar para uma representação binária, em que ao invés das frequências *f*, as posições na matriz seriam 1 para o caso de o termo estar presente no documento ou 0 caso não esteja.

Para muitas aplicações como classificações, *topic-modeling* e sistemas de recomendação esse tipo de representação é suficiente, mas existem casos em que a ordem das palavras é necessária, pois revela a semântica do texto. Assim, o *corpus* pode ser tratado como um conjunto de sequências de termos. Uma forma comum de abordar esse conjunto é, também, transforma-lo em uma representação multidimensional, porém, de *embeddings*, que diferentemente do *bag-of-words*, incorpora a estrutura sequencial dos dados, ou seja, tem natureza semântica. Uma das maneiras de construir esse tipo de representação é por meio de métodos chamados *Language-independent methods*, que consistem em técnicas para utilizar análise estatística do conjunto de sequências para criar modelos de linguagem. Um modelo de linguagem estatístico é uma distribuição de probabilidade gerada a partir das sequências de palavras, aprendendo a probabilidade de uma palavra baseado no contexto. A Figura 1 representa o processo de construção da representação por meio de *embeddings* [AGGARWAL 2018]*.*

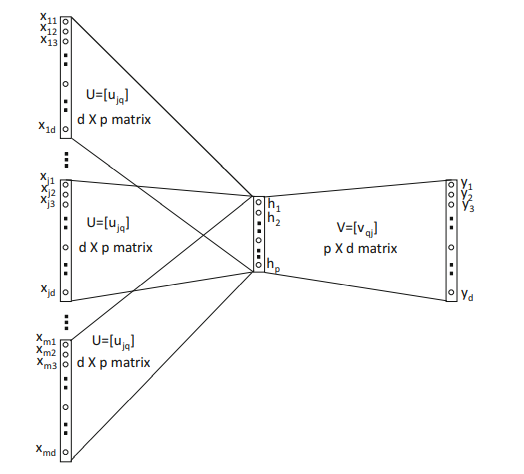
**Figura 1 – Conversão de sequências em multidimensional embedding**



Fonte: AGGARWAL, Charu C. **Machine learning for text**. 2018

O método que foi aplicado no projeto foi o *word2vec*, que é um modelo de linguagem neural, ou seja, utiliza redes neurais para criar a representação estrutural do texto por meio do conjunto de sequências. Basicamente, uma das abordagens é o *Continuous Bag of Words* (CBOW), que consiste em definir uma janela *t*, que corresponde ao número de palavras que irá compor uma sequência dentro do conjunto, e tentar prever a palavra central com base nas adjacentes, ou seja, dada uma sequência de palavras {*wi-t, ..., wi+t*}, a rede neural tenta prever a palavra *wi*. Para que isso seja possível, o modelo deve ser treinado com uma quantidade grande de sequências pertencentes ao *corpus* em questão, a fim de ajustar os pesos da rede da melhor forma possível [ZONG et al. 2021]. A Figura 2 ilustra esse processo fazendo uso dos *one hot vectors*, que é uma forma de representar as palavras de um texto. Por fim, é importante lembrar que a formalização matemática das redes neurais não será aprofundada pois não é o foco do projeto.

**Figura 2 – Modelo CBOW para entrada de contexto (X) e saída da distribuição de probabilidades das possíveis palavras centrais (Y)**.

****

Fonte: AGGARWAL, Charu C. **Machine learning for text**. 2018

### 2.2.2. Similaridade com NFe

A partir do momento em que temos os documentos do *corpus* organizados e pré-processados, é possível comparar termos de diferentes documentos. Para fazer a comparação de dois termos, ou *strings*, é utilizado medidas de similaridade, que calculam matematicamente o quanto duas *strings* são parecidas.

A medida colocada em prática no projeto é a medida de Jaro-Winkler, a qual é uma variação da distância de Jaro [JARO 1989]. Considerando duas *strings* s e t, s*’* corresponde aos caracteres de s que são comuns a t, e de forma análoga temos t’. De maneira simples, um caractere de s é comum se ele existe em t e está em uma posição próxima. Assumindo Ts,t como sendo o número de transposições dos caracteres de s’ relativos à t’ (exemplo disponível no Apêndice A), a similaridade de Jaro para as strings é

A variante Winkler faz uma pequena modificação que melhora os pesos de pares pouco similares, mas que compartilham um mesmo prefixo [COHEN, RAVIKUMAR, FIENBERG 2003]. Assim sendo, definindo um fator de escalonamento p,duas *strings* que compartilham os l primeiros caracteres nas mesmas posições possuiriam a seguinte medida de Jaro-Winkler:

Dessa forma, com as NFes pré-processadas, os termos delas poderão ser comparadas com os da tabela TACO utilizando a medida de Jaro-Winkler, podendo, assim, relacionar os alimentos da NFe com os nutrientes listados na tabela.

### 2.2.3 Tabela Brasileira de Composição de Alimentos (TACO)

A Tabela Brasileira de Composição de Alimentos foi a tabela escolhida para ser utilizada como referência no projeto com relação aos dados nutricionais dos alimentos a serem analisados nas NFes.

Por sua vez, a tabela TACO é um imenso conjunto de dados sobre a composição dos principais alimentos consumidos no Brasil. Os resultados trazidos neste documento têm grande confiabilidade, uma vez que o plano de amostragem garante valores representativos, com análises realizadas por laboratórios com capacidade analítica comprovada [TACO 2011]. O documento que será usado no projeto, ao invés da forma textual, será o formato .xls, que facilita na importação e no pré-processamento, caso necessário.

## 2.4. Considerações Finais

Nesse capítulo foi explicado como é feito o pré-processamento de um texto bruto e como eles são representados em uma visão matemática e computacional, a fim de aplicar técnicas e algoritmos. Além disso, ainda foi citada a tabela TACO e a medida de similaridade Jaro-Winkler, que serão os principais artefatos para realizar a análise de cada NFe. No Capítulo 3 será apresentado o desenvolvimento do projeto, desde a extração das NFes até o mapeamento dos municípios inspecionados.

# CAPÍTULO 3: DESENVOLVIMENTO DO TRABALHO

## 3.1. Considerações Iniciais

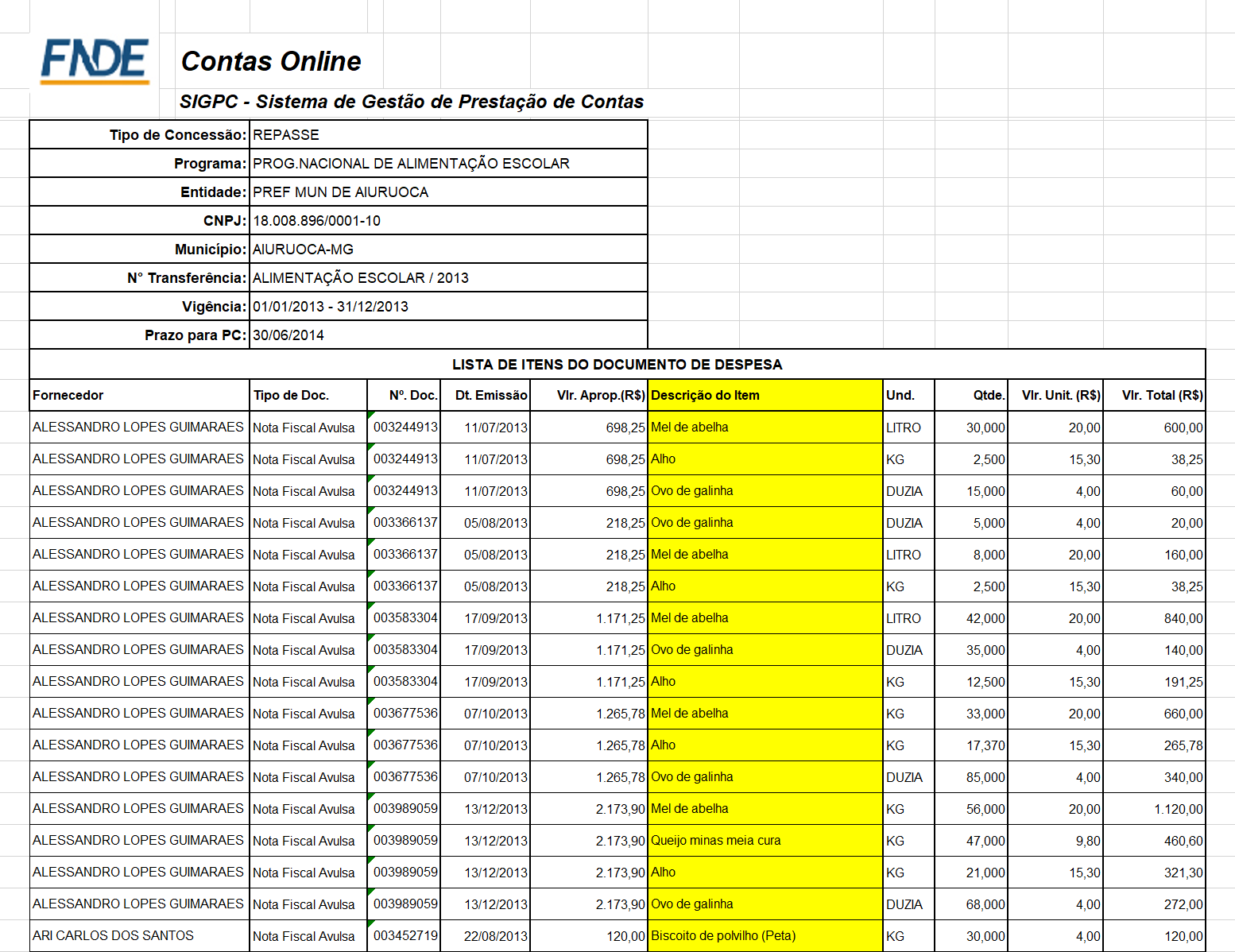
Neste capítulo será descrito de forma sucinta como o projeto foi desenvolvido. O trabalho começa com uma breve análise do Sistema de Gestão de Prestação de Contas (SiGPC) e como as notas fiscais eletrônicas (NFes) referentes ao PNAE estão disponíveis para serem acessadas. Com as notas em mãos será feito um breve pré-processamento e, logo em seguida, uma comparação dos itens presentes na NFe com a tabela TACO por meio da medida de similaridade de Jaro-Winkler. Para finalizar, serão apresentados os resultados do trabalho com as informações nutricionais dos municípios e indicadores por meio de mapas de calor. Vale ressaltar que todos os algoritmos e métodos foram feitos com a linguagem de programação Python.

## 3.2. Base de NFe

O sistema que disponibiliza as notas fiscais referentes ao Programa Nacional de Alimentação Escolar é o Sistema de Gestão de Prestação de Contas (SiGPC) que pode ser acessado por meio do site do Fundo Nacional de Desenvolvimento [FNDE 2022]. Uma vez acessado, é possível pesquisar um tipo de concessão referente a um programa governamental específico filtrado por estado. Neste trabalho foram captados registros de repasse disponíveis do PNAE de todos os estados brasileiros a partir da criação de um algoritmo em Python usando a biblioteca Selenium.

A Figura 3 ilustra como são os arquivos das notas fiscais baixadas. É possível notar que o arquivo contém os dados da entidade e as transações realizadas, além do fato das informações estarem dispostas em formato de tabela. Esse formato facilita o pré-processamento e a captação dos dados da instituição, do município e dos alimentos. A coluna marcada em amarelo pelo autor deixa em evidência o principal atributo da tabela que será comparado com a tabela TACO. Nota-se que, salvo algumas exceções, a coluna já é composta por termos, os quais estão praticamente prontos para a análise de similaridade.

**Figura 3 - Exemplo de NFe**



Fonte: https://www.fnde.gov.br/sigpcadm/sistema.pu?operation=localizar

Finalmente, a Tabela 1 registra a quantidade de repasses captados de cada estado brasileiro.

**Tabela 1 – NFes captadas por estado brasileiro**

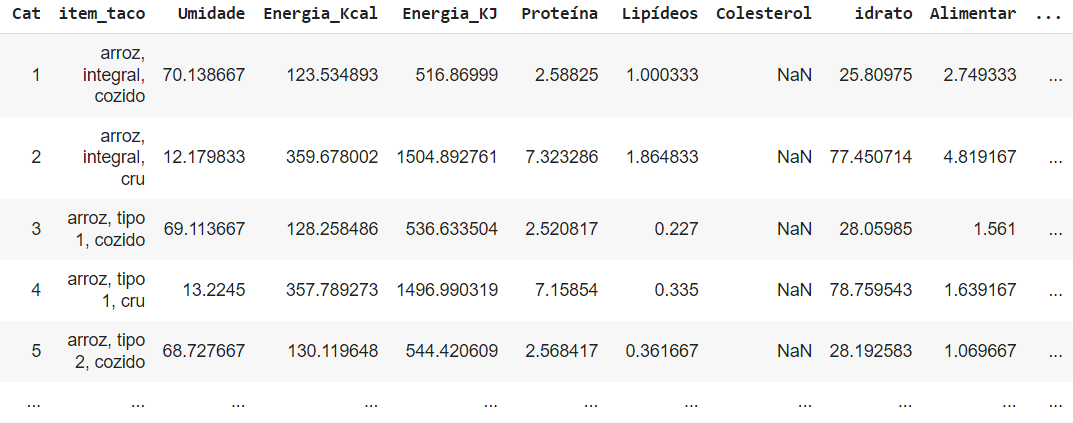
|  |  |
| --- | --- |
| **UF** | **Número de repasses (NFe)** |
| AC | 206 |
| AL | 913 |
| AM | 530 |
| AP | 121 |
| BA | 3.742 |
| CE | 1.656 |
| DF | 9 |
| ES | 704 |
| GO | 2.153 |
| MA | 1.921 |
| MG | 7.543 |
| MS | 703 |
| MT | 1.248 |
| PA | 1.285 |
| PB | 1.997 |
| PE | 1.654 |
| PI | 1.988 |
| PR | 3.577 |
| RJ | 798 |
| RN | 1.478 |
| RO | 457 |
| RR | 123 |
| RS | 4.431 |
| SC | 2.628 |
| SE | 679 |
| SP | 4.068 |
| TO | 1.220 |

Fonte: Desenvolvido pelo autor.

## 3.3. Mapeamento entre NFe e TACO

Já foi ressaltado anteriormente que tanto a tabela TACO quanto as notas fiscais eletrônicas estão em formato de tabelas e não de texto livre, o que facilita muito a importação delas para o formato de *Dataframe* por meio do uso da biblioteca Pandas. A tabela TACO, primeiramente, após ser importada foi pré-processada e transformada no *DataFrame* ilustrado na Figura 4, que possui os alimentos registrados na tabela concatenados com suas informações nutricionais.

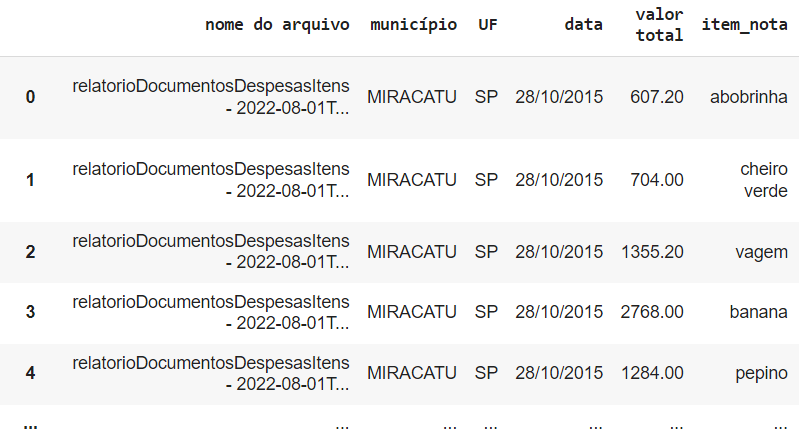
**Figura 4 – DataFrame da tabela TACO**



Fonte: Desenvolvido pelo autor

A título de exemplo, serão utilizadas as notas do estado de São Paulo, porém o mesmo processo pode ser feito para os demais estados sem perda de generalidade. Dito isso, a partir de cada NFe, com o mesmo formato daquela mostrada na Figura 3, é possível obter um *Dataframe* das NFes, mostrado na Figura 5, o qual apresenta os dados das notas que são interessantes: nome do arquivo, município, UF, data, valor total gasto e item comprado.

**Figura 5 – DataFrame gerado a partir das NFes do estado de SP**



Fonte: Desenvolvido pelo autor

Percebe-se que com o processamento das notas fiscais já foi possível relacionar os municípios com os alimentos comprados por suas escolas residentes. Com isso, basta que os dois *DataFrames* sejam relacionados para que cada municípios seja mapeado não só pelos alimentos comprados, mas também por suas informações nutricionais, a fim de mapea-los nutricionalmente.

Para relacionar os dois *Dataframes*, cada item das NFes (coluna “item\_nota”) deve ser comparado aos itens da tabela TACO (coluna “item\_taco”), medindo a similaridade por meio da medida Jaro-Winkler e selecionando o maior valor. Assim, existe uma grande probabilidade de que cada item da nota seja relacionado com suas verdadeiras informações nutricionais. Esta probabilidade é maior se o limiar mínimo de similaridade obtida pela Jaro-Winkler se aproximar de 1.

## 3.4. Resultados Obtidos

### 3.4.1 Informações nutricionais dos municípios

O primeiro resultado é o *Dataframe* antes desejado, que relaciona os municípios com os dados nutricionais adquiridos com os recursos do PNAE, mostrado na Figura 6.

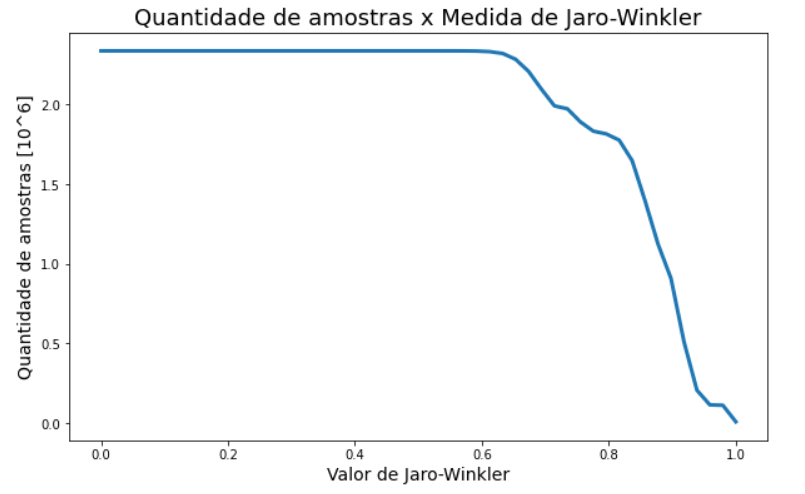
**Figura 6 – DataFrame com itens das NFes relacionados com seus dados nutricionais**



Fonte: Desenvolvido pelo autor

É possível perceber que na linha de número 1 a medida de Jaro-Winkler é de 0,88, o que faz com que ela relacione de forma correta o item “banana nanica climatizada” da NFe com o item “banana, nanica, crua” da tabela TACO, enquanto que na linha 0, é relacionado “ovos brancos” e “repolho, branco, cru” com medida de aproximadamente 0,66. Isso mostra que alguns produtos presentes nas NFes podem não existir na tabela TACO e o algoritmo irá buscar o mais próximo possível, o que nem sempre é satisfatório. Assim sendo, para fazer um mapeamento correto entre os municípios e os dados nutricionais, relacionando de maneira correta os alimentos, é necessário discutir a cobertura dessa medida de similaridade para estabelecer um valor mínimo. Dessa forma, foi construído um gráfico que relaciona a quantidade de notas fiscais com o valor da medida de similaridade, apresentado na Figura 7. É importante ressaltar que quanto maior o limiar maior a perda de amostras das NFes, como indica o gráfico.

**Figura 7 - Quantidade de amostras das NFe acima de um valor mínimo de Jaro-Winkler para o estado de SP**



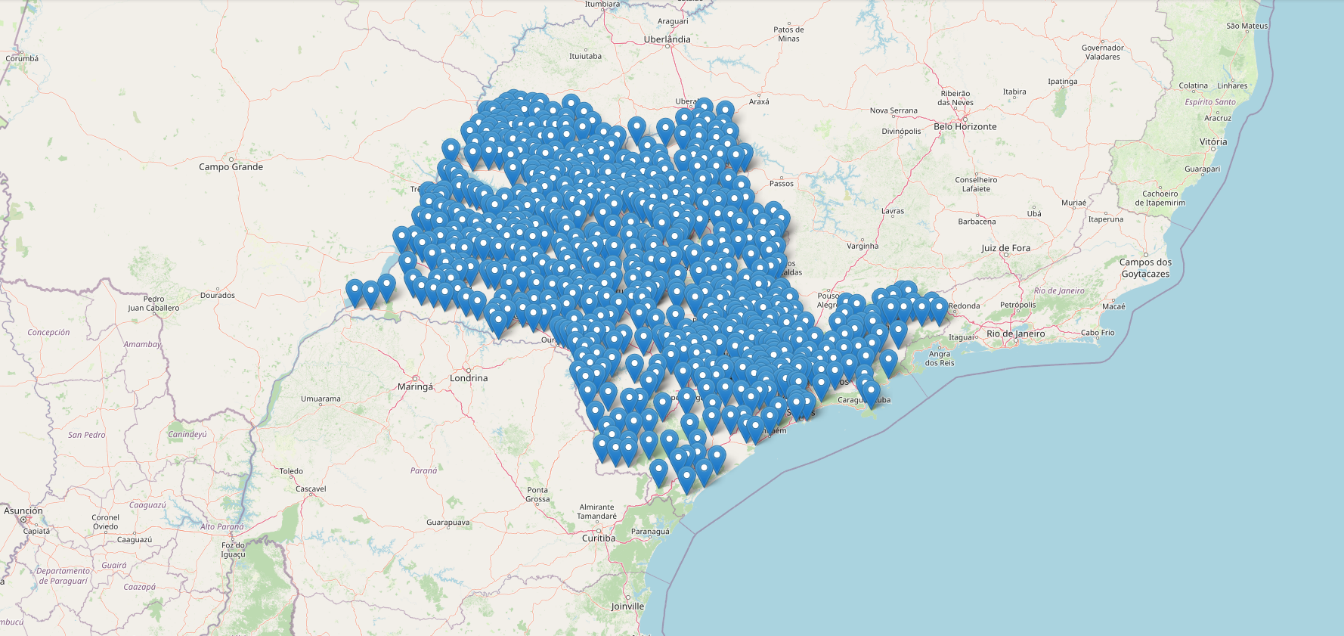
Fonte: Desenvolvido pelo autor

A partir das Figuras 6 e 7 entende-se que no patamar de medida de 0,76 existe uma quantidade interessante de notas, porém apresenta erros de correlação, enquanto que de 0,85 até aproximadamente 1 existem ótimos resultados, mas poucas amostras. Assim, um valor de limiar satisfatório para ser analisado foi fixado em 0,85, com 1481814 registros amostrados.

### 3.4.2 Indicadores por meio de Mapas de Calor

Após ser obtido o *Dataframe* final, apenas com itens que se relacionaram corretamente, por conta do limiar de similaridade estabelecido, foi possível captar as coordenadas de latitude e longitude dos municípios, mostrado na Figura 8, e estabelecer uma pontuação para criação de mapas de calor. A pontuação escolhida neste projeto foi baseada apenas no atributo energia em kcal dos alimentos registrados no *DataFrame* da Figura 6, normalizada com o valor total gasto pelo município, para que cidades com mais entidades não sobreponham às demais, uma vez que recebem mais recursos e, logo, teoricamente, compram mais alimentos. Esse indicador mostra a compra de calorias por valor repassado, ou seja, como estão sendo geridos os recursos financeiros.­

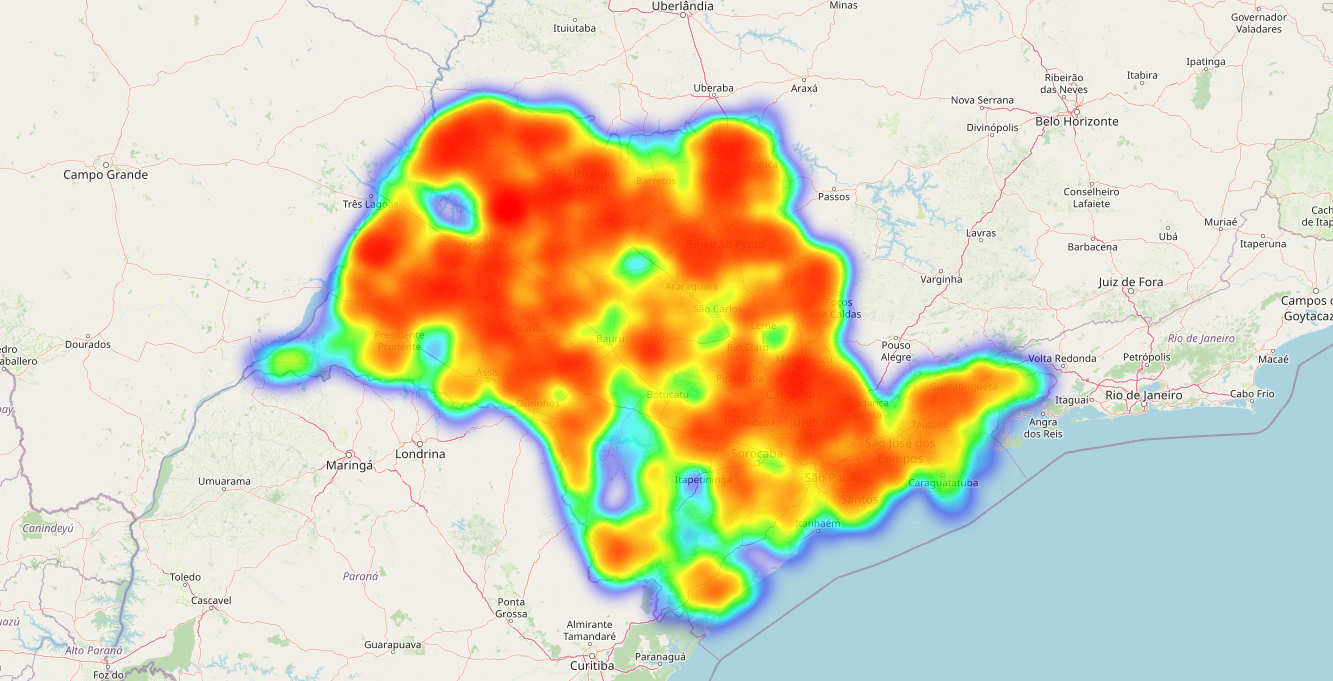
**Figura 8 – Mapa focado no estado de SP com os municípios participantes do PNAE**



Fonte: Desenvolvido pelo autor

A partir das coordenadas dos municípios, foi gerado o mapa de calor, presente na Figura 9, para o estado de São Paulo, mapeando os municípios participantes do PNAE de acordo com a pontuação nutricional definida.

**Figura 9 – Mapa de calor do estado de SP**

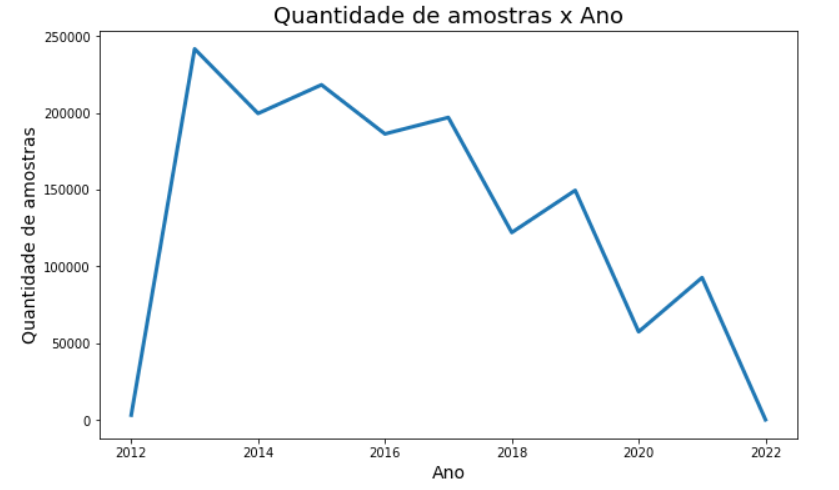


Fonte: Desenvolvido pelo autor

Esse primeiro mapa engloba todas as transações recolhidas por meio das NFes de repasse para o estado de São Paulo. A partir dele é tirado conclusões gerais, por exemplo, regiões com maior pontuação apresentam uma melhor administração dos recursos de maneira geral, enquanto que as regiões com pontuação baixa ou sem nenhuma pontuação têm má administração dos recursos ou não estão mapeadas no mapa por não possuírem instituições participantes do PNAE. Isso resulta em uma informação visual de quais lugares precisam ser monitorados com mais atenção, além de alertar municípios que não possuem entidades capazes de participarem do programa e aproveitar de seus auxílios.

Por meio da estratégia desenvolvida no trabalho, também é possível fazer a análise do estado anualmente. Para isso, primeiramente, foi feito um gráfico, presente na Figura 10, que relaciona as amostras retiradas das notas entre 2012 e 2022. Nota-se que a quantidade varia bruscamente, pois foi constatado, durante a construção, que várias das datas apresentavam o ano com valores errados, sendo, assim, elas foram retiradas. Isso não afeta a análise anterior, que computa todas as amostras independente da data.

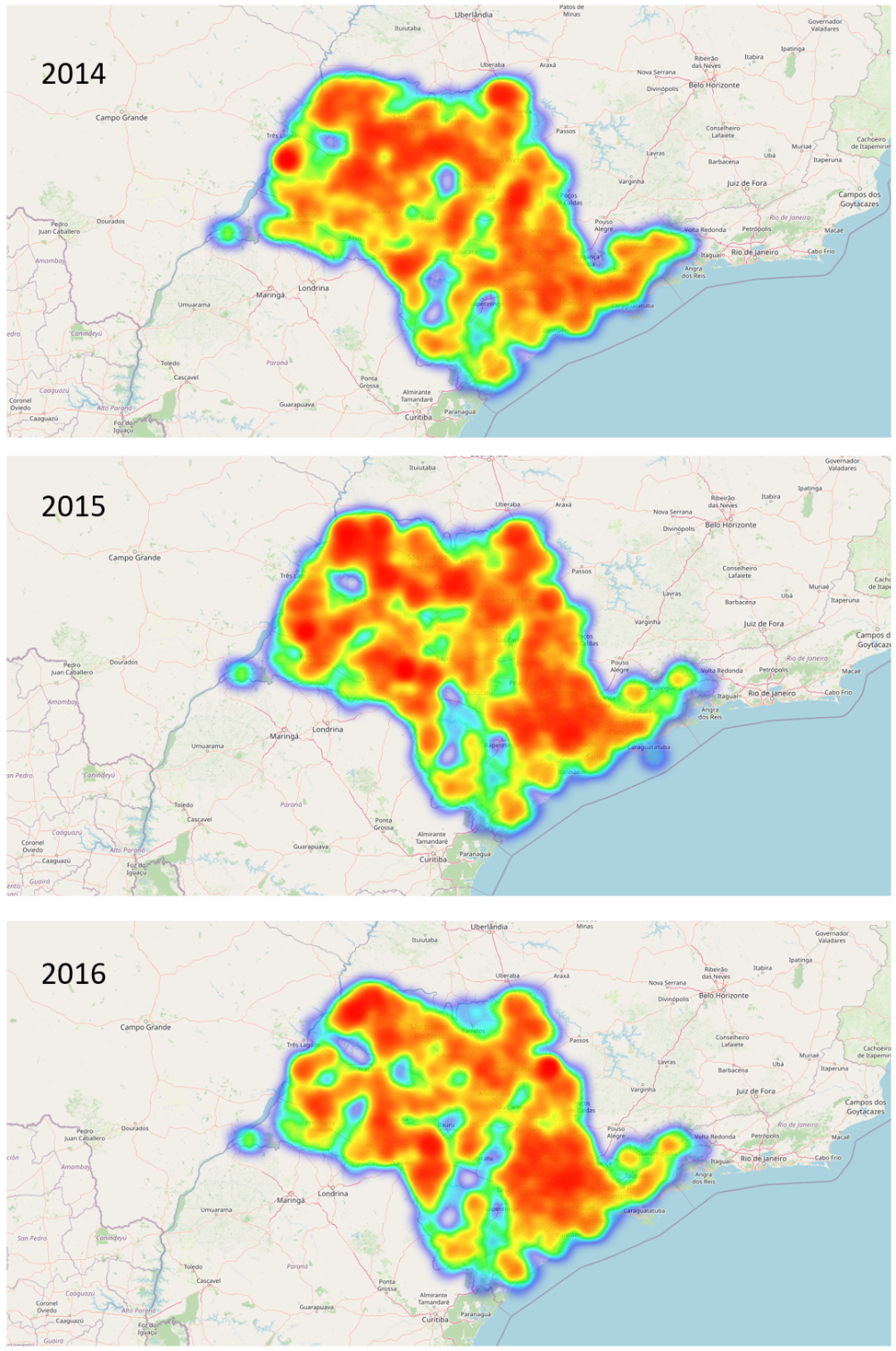
**Figura 10 - Quantidade de amostras retiradas das NFes de acordo com o ano para o estado de SP**



Fonte: Desenvolvido pelo autor

Considerando que a quantidade de amostras reflete o número de entidades participantes do programa, é sensato que a comparação seja feita entre anos com amostragem parecidos. Dessa forma, pode-se gerar os mapas de calor para os anos de 2014 até 2016, os quais estão presentes na Figura 11.

**Figura 11 - Mapas de calor do estado de SP de 2014 até 2016**



Fonte: Desenvolvido pelo autor

Observando a Figura 11, pode-se afirmar que embora 2015 possua maior quantidade de amostras, elas estão mais concentradas, o que pode indicar piora na administração de recursos de alguns municípios e melhora em outras regiões. Isso pode ser visto na região leste do estado, com perda de pontuação no extremo e melhora em regiões próximas a capital. Já em 2016 o mapa está menos preenchido devido a diminuição de amostras, porém com formato muito parecido com os anteriores. Se comparado à 2014, que tem uma amostragem parecida, é possível ver que também está mais concentrado e menos distribuído, com piora na pontuação do centro-sul do estado.

Finalmente, vale lembrar que essas análises são de caráter geral para ilustrar o poder de monitoramento dessa ideia, porém é possível ser tão detalhada quanto o usuário queira, por exemplo, realizar observações anuais sobre um único município, de maneira singular.

## 3.5. Dificuldades e Limitações

As dificuldades e limitações começam na captação das NFes no SiGPC, pois as notas não são agrupadas, ou seja, devem ser pegas uma a uma. Assim, é difícil criar um algoritmo que pegue as notas de forma eficiente. Ainda sobre o SiGPC, existem repasses que não são possíveis de serem baixados, o que causou perda de informações e redução no desempenho no algoritmo de captação.

É importante mencionar que esse mapeamento pode ser feito para os municípios de todo o Brasil que tem entidades participantes do PNAE, e isso pode incluir regionalismo no nome dos produtos. Como o uso da medida de Jaro-Winkler é uma comparação de caracteres, ou seja, é uma análise estrutural, que ignora a semântica, outra grande limitação é que não é possível relacionar, por exemplo, “mandioca” com “aipim” ou “macaxeira”.

Uma outra grande limitação pode ser vista nos erros que foram encontradas em diversas notas, sejam por datas erradas ou por informações faltantes. Embora esse projeto aponte uma estratégia interessante de se obter indicadores da nutrição das escolas, ele fica, ainda, à mercê dos dados que são gerados pelas entidades, muitas vezes defeituosos.

Finalmente, embora não fosse o foco do projeto, uma dificuldade é encontrar uma forma de pontuação mais eficiente para geração dos mapas de calor, que pudesse representar de fato se os municípios estão utilizando os recursos passados pelo PNAE para comprar alimentos nutritivos, o que uma simples medida da energia em Kcal de cada alimento não aborda satisfatoriamente. Ainda nesse quesito, a normalização da pontuação feita com os valores totais também não é a melhor, visto que uma normalização feita com a quantidade de alunos nas entidades seria mais eficiente.

## 3.6. Considerações Finais

Neste capítulo foi apresentado todo o desenvolvimento prático realizado no projeto, desde a captação das notas fiscais eletrônicas referentes a repasses feitos pelo PNAE até os mapas de calor gerados atrelando os municípios aos alimentos comprados e suas informações nutricionais. Foi possível, assim, observar como é viável monitorar os estados e municípios, desde que seja utilizada uma pontuação discriminativa. O Capítulo 4 finaliza o projeto, mostrando as contribuições e considerações para possíveis trabalhos futuros, além de concluir se os objetivos propostos anteriormente foram atingidos.

# CAPÍTULO 4: CONCLUSÃO

## 4.1. Contribuições

O projeto contribuiu, principalmente, para mostrar como a mineração, não apenas de textos, mas de dados em geral, é uma ferramenta poderosa, assim como serviu para mostrar, também, como políticas governamentais são difíceis de fiscalizar, por mais que seja necessário. O objetivo central do projeto que constava em utilizar NFes como documento a ser analisado por medidas de similaridade e algoritmos de mapeamento, a fim de colocar a mineração de textos como instrumento de monitoramento, foi bem-sucedida. Isso contribui para mostrar como as políticas públicas estão progredindo dentro do escopo pelo qual elas foram criadas. A utilização, desse programa de alimentação em específico coloca ainda a mineração de textos como uma possível ferramenta importante no combate à fome, uma vez que esse e demais programas similares têm esse combate como principal foco.

A ferramenta (https://github.com/mathrmiranda/NFE\_JaroWinkler) construída neste projeto será utilizada futuramente dentro de um projeto maior associado ao Centro de Inteligência Artificial (C4AI), na temática de combate à Segurança Alimentar (https://c4ai.inova.usp.br/research/#AgriBio).

## 4.2. Considerações sobre o Curso de Graduação

O curso de graduação traz muitas matérias obrigatórias sobre estruturas de dados e programação, que apresentam introduções sobre como fazer algoritmos em diversas linguagens, e ainda existem matérias opcionais que trazem como especialidade a ciência de dados. Assim, a linguagem Python, além das técnicas de mineração de dados foram essenciais para o desenvolvimento do projeto.

Com relação a críticas, um ponto negativo no curso, como, por exemplo, a incrivelmente extensa grade obrigatória, que possui matérias ultrapassadas ou que não tem muita aplicabilidade atual, o que torna a evasão do curso alta. Porém, um ponto positivo é que o curso prepara, mesmo que de maneira básica, tanto para trabalhar em áreas relacionadas a engenharia elétrica quanto em áreas da ciência da computação.

Uma sugestão aos gestores do curso de engenharia de computação é diminuir a grade obrigatória, retirando assuntos muito específicos e dando a liberdade ao estudante de escolher qual das duas áreas citadas anteriormente ele deseja dar ênfase, tornando o curso mais maleável e proveitoso.

## 4.3. Trabalhos Futuros

Um possível trabalho que pode tomar esse projeto como referência pode ter como ponto de partida a utilização de análises semânticas, como o *word2vec*, ao invés da análise estrutural de medidas de similaridade, a fim de resolver limitações, como o regionalismo, que foi citado. Também pode ser interessante melhorar a pontuação estabelecida no mapeamento dos municípios, beneficiando-se de parcerias com grupos especializados em nutrição, para que os resultados reflitam a nutrição dos alunos da melhor maneira possível. Por fim, as estratégias desenvolvidas no trabalho podem ser aplicadas a diferentes programas com diferentes documentações, com intuito de monitorar outras políticas e talvez estabelecer um documento que seja mais consistente e efetivo que as notas fiscais eletrônicas.

# REFERÊNCIAS

AGGARWAL, Charu C. **Machine learning for text**. Cham: Springer, 2018.

COHEN, William; RAVIKUMAR, Pradeep; FIENBERG, Stephen. A comparison of string metrics for matching names and records. In: **Kdd workshop on data cleaning and object consolidation**. 2003.

FELDMAN, Ronen et al. **The text mining handbook: advanced approaches in analyzing unstructured data**. Cambridge university press, 2007.

FNDE. **Site do FNDE**, 2022. Plataforma do Fundo Nacional de Desenvolvimento da Educação do Ministério da Educação. Disponível em: https://www.fnde.gov.br. Acesso em: 20 out. 2022.

JARO, Matthew A. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. **Journal of the American Statistical Association**, v. 84, n. 406, p. 414-420, 1989.

PEDONE, Luiz. **Formulação, implementação e avaliação de políticas públicas**. 1986.

TACO, **Tabela Brasileira de Composição de Alimentos**. 4ª edição revisada e ampliada, 2011.

ZONG, C. Xia R, Zhang J. **Text Data Mining**. Singapore: Springer; 2021.

# APÊNDICE A – Exemplo de Transposição (*Ts,t*)

Considerando duas *strings*:

s = abcdefg

t = abcjkgf

Os dois conjuntos s­­*’* e t*’* são:

*s’* = {a, b, c, f, g}

*t’* = {a, b, c, g, f}

Dessa forma, a transposição Ts,t ocorre apenas uma vez, já que os únicos caracteres pertencentes a s­­*’* e t*’* que estão trocados são “f” e “g”. Assim, a distância de Jaro seria

Porém, assumindo p = 0,1 e considerando que o número de caracteres iguais no prefixo das *strings* é l = 3, a medida de similaridade de Jaro-Winkler para as duas palavras é