UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I - CAMPINA GRANDE
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE COMPUTAÇÃO
CURSO DE GRADUAÇÃO EM COMPUTAÇÃO

ÂNGELO GABRIEL PAZ DA SILVA

O USO DE TÉCNICAS DE SIMILARIDADE E EXPRESSÕES REGULARES PARA CLASSIFICAÇÃO DE PAUTAS FISCAIS DE ÁGUA

CAMPINA GRANDE

ÂNGELO GABRIEL PAZ DA SILVA

O USO DE TÉCNICAS DE SIMILARIDADE E EXPRESSÕES REGULARES PARA CLASSIFICAÇÃO DE PAUTAS FISCAIS DE ÁGUA

Trabalho de Conclusão de Curso de Graduação apresentado ao Curso de Computação do Centro de Ciência e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de Bacharel em Computação.

Área de concentração: Inteligência artificial e processamento de linguagem natural.

Orientador: Profª. Drª. Kezia de Vasconcelos Oliveira Dantas.

CAMPINA GRANDE
2023

S586u Silva, Angelo Gabriel Paz da.
O uso de técnicas de similaridade e expressões regulares para classificação de pautas fiscais de água [manuscrito] / Angelo Gabriel Paz da Silva. - 2023.
40 p. : il. colorido.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Computação) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2023.
"Orientação : Profa. Dra. Kezia de Vasconcelos Oliveira Dantas, Coordenação do Curso de Computação - CCT. "

1. Processamento de linguagem natural. 2. Classificadores. 3. Linguagem Python. 4. Pautas fiscais. I. Título

21. ed. CDD 005.3

Elaborada por Talita M. A. Tavares - CRB - CRB 15/971 BC/UEPB

ÂNGELO GABRIEL PAZ DA SILVA

O USO DE TÉCNICAS DE SIMILARIDADE E EXPRESSÕES REGULARES PARA CLASSIFICAÇÃO DE PAUTAS FISCAIS DE ÁGUA

Trabalho de Conclusão de Curso apresentado ao Curso de Computação do Centro de Ciência e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de Bacharel em Computação.

Área de concentração: Inteligência artificial e processamento de linguagem natural.

Aprovada em: 03/07/2021.

BANCA EXAMINADORA

Kézia de Vasconcelos Oliveira Dantas
Profa. Dra. Kezia de Vasconcelos O. Dantas (CCT/UEPB)
Orientador(a)

Sabrina de F. Souto.
Profa. Dra. Sabrina de Figueiredo Souto (CCT/UEPB)
Examinador(a)

Paulo Bh
Prof. Dr. Paulo Eduardo e Silva Barbosa (CCT/UEPB)
Examinador(a)

Dedico este TCC a minha família, meus pais e minha irmã que sempre foram e continuam sendo o pilar que me sustenta diante das adversidades, me lembrando que sou capaz de qualquer coisa que meu esforço for capaz de alcançar .

AGRADECIMENTOS

Primeiramente a Deus por me abençoar e me dar forças para trilhar meu caminho e fazer sua obra com felicidade, por ter me encontrado na área de ciências da computação.

À meus pais Almy Geraldo e Eliete Maria e minha irmã Gisela Cristiny, por me inspirarem todos os dias e darem todo suporte necessário me ajudando a reconhecer meu potencial.

Aos meus amigos e colegas que dividiram tantos momentos de superação e alegria comigo durante essa jornada: Jefferson Gomes, Klayton Marcos, Kennedy Johnson, Natália Maria, Rafaela Candido, Mikaelly Santos, Rodolfo Pereira e Renan Rey.

Ao meu amigo Bruno Queiroz sempre disposto a me ouvir me mostrando meus erros e acertos, me ajudando a me tornar uma versão melhor de mim mesmo.

Aos professores do Curso de Graduação da UEPB, em especial, a minha orientadora Profª. Drª. Kezia de Vasconcelos Oliveira Dantas, a Profª. Drª. Sabrina de Figueiredo Souto e ao Prof. Dr. Paulo Eduardo e Silva Barbosa, que me deram a oportunidade de demonstrar minha competência e comprometimento ao participar de projetos da UEPB/NUTES.

“Se eu não sei mais em um dia do que eu sabia no dia anterior, pra mim esse foi um dia desperdiçado”
(Neil De Grasse Tyson)

RESUMO

Diante do grande volume de dados gerados a partir das notas fiscais eletrônicas da SEFAZ-PB, a tributação manual torna-se demorada e custosa. Por isso, a Secretaria de Fazenda do estado(SEFAZ-PB) fez parceria com o NUTES/UEPB para desenvolver um classificador de produtos integrado ao sistema de faturamento automático. Objetivando classificar as descrições de contribuintes de produtos referentes à categoria de águas minerais, nas classes definidas pelas pautas fiscais, esse trabalho visa o desenvolvimento de um módulo de classificação de produtos referentes a águas minerais, nas classes definidas pelas pautas fiscais, disponibilizadas pela equipe técnica da SEFAZ-PB, com o fim de auxiliar os auditores fiscais no processo de tributação. Como resultados para esse trabalho, foi produzida uma versão desse componente com a linguagem Python utilizando bibliotecas de manipulação de dados e processamento de linguagem natural que já está integrado ao classificador e conta com 90% de acurácia na classificação.

Palavras-Chave:classificador; processamento de linguagem natural ;pautas fiscais.

ABSTRACT

Given the large volume of data generated from SEFAZ-PB electronic invoices, manual taxation becomes time-consuming and costly. For this reason, the State Finance Department (SEFAZ-PB) partnered with NUTES/UEPB to develop a product classifier integrated into the automatic billing system. Aiming to classify the taxpayer descriptions of products referring to the category of mineral waters, in the classes defined by the fiscal tariffs, this work aims at the development of a module of classification of products referring to mineral waters, in the classes defined by the fiscal tariffs, made available by the technical team of SEFAZ-PB, in order to assist tax auditors in the taxation process. As a result of this work, a version of this component was produced with the Python language using data manipulation libraries and natural language processing that is already integrated into the classifier and has 90% classification accuracy.

Keywords: classifier; natural language processing;tax tariffs.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

NF-e	Notas fiscais eletrônicas
ML	Machine Learning
PLN	Processamento de Linguagem Natural
NCM	Nomenclatura comum do Mercosul
MVA	Margem de valor agregado
SQ	Valor de identificação da pauta

SUMÁRIO

1 INTRODUÇÃO

As notas fiscais eletrônicas (NF-e) ajudaram no processo de obtenção de documentação para a tributação dos produtos por parte da SEFAZ-PB. Agora, as NF-e são geradas e enviadas em tempo real para os órgãos reguladores. Isso permite que os órgãos verifiquem a integridade dos itens e os enviem para a fase de classificação. A classificação é feita com base em classes padronizadas, contidas em um documento chamado pauta fiscal. Cada produto recebe uma descrição feita pelo contribuinte e é atribuído a uma classe de pauta fiscal. A classe informa qual a tributação correta a ser aplicada ao produto descrito.

No entanto, a velocidade de produção e compartilhamento de dados também apresenta uma deficiência. O trabalho manual necessário para analisar cada descrição e atribuir a respectiva classe de produto é lento e propenso a erros, que podem ocorrer devido a atribuições erradas por acidente causadas por desatenção. Essa deficiência resulta em um alto custo para o responsável e para a SEFAZ-PB. Esse custo se traduz em perda de tributação.

Diante desse cenário, uma parceria foi formada entre a SEFAZ-PB e o Núcleo de Tecnologias Estratégicas em Saúde(NUTES/UEPB). Essa parceria em questão é o desenvolvimento de um classificador de produtos que tem como objetivo atender à enorme demanda de classificação das descrições feitas por contribuintes entregues pelas NF-e.

Até o momento presente, uma versão do classificador está integrada ao sistema de faturamento automático da SEFAZ-PB, e apresenta uma acurácia geral de 99% por cento para classificação dos produtos que estão incluídos atualmente no classificador.

O classificador utiliza a descrição encontrada na nota fiscal e na NCM, que refere-se a essa descrição. Isso é feito para primeiramente dividir os produtos em tipos. Os tipos são: bebidas quentes, cerveja, cachaça, água, refrigerante, energético, madeira, cigarro, açúcar e isotônico.

Após isso, as descrições são enviadas a seus respectivos módulos de classificação. Esses módulos tomam como referência as pautas fiscais vigentes. Eles definem as classes que contêm descrições padronizadas que serão comparadas com as descrições feitas pelos contribuintes. Caso um nível de semelhança considerável seja notado, o número de SQ que identifica a classe na pauta será atribuído àquele produto. Isso define, portanto, sua tributação.

Dessa forma, o trabalho atual foca no desenvolvimento de um dos módulos de classificação. Esse módulo é responsável pela classificação de descrições de produtos de água mineral em suas respectivas classes de pauta fiscal.

Um trabalho semelhante foi desenvolvido por Kumar(2019) no artigo “Product Classification using Machine Learning-Part I”, onde ele usa técnicas de manipulação de dados, machine learning e PLN para classificação de produtos em macro e micro categorias, como por exemplo, identificar que um produto é um smartphone e após isso identificar a qual marca de smartphone aquele produto pertence.

No artigo “Resume Screening Classification using Artificial Intelligence and Natural Language Processing” desenvolvido por Arvind Kumar et al(2023), o autor descreve o processo de classificação de curriculum em duas classes “seleção” e “rejeição”. Para tal, são utilizadas técnicas de processamento de linguagem natural e machine learning, na linguagem de programação Python.

1.1 Objetivos

O objetivo deste trabalho é criar um módulo de classificação de pauta para produtos de água mineral em seus respectivos Sq´s de pauta, definidas pelas pautas fiscais geradas pela SEFAZ-PB, com o objetivo de auxiliar no processo de tributação.

Com isso, foi desenvolvido um código em python, responsável pelo pré-processamento das descrições que entram no sistema, e utilização de bibliotecas e frameworks para o cálculo de similaridade entre as descrições pré-processadas dos contribuintes e as descrições que se encontram nas pautas fiscais, para por fim determinar o *sq* de pauta daquela descrição, que corresponde a um código interno da SEFAZ-PB responsável pela identificação de cada pauta fiscal.

1.1.1 Objetivos Específicos

- Realizar o levantamento de regras de negócio para entender as classes de pauta dos produtos referentes às águas minerais.
- Realizar classificação manual de produtos de água mineral, com o objetivo de produzir um gabarito para ser utilizado em um conjunto de casos de teste.
- Desenvolver o módulo de classificação automático em linguagem python, que receba como entrada possíveis descrições de produtos relacionados à água mineral feita por contribuintes, compare o texto contido nessas descrições com as descrições contidas nas classes da pauta fiscal vigente de água mineral e atribuir seus respectivos Sq´s de pauta .

- Fazer com que o módulo produzido se integre facilmente ao classificador, evitando conflitos durante a manipulação de dados entre as diferentes fases de classificação dos produtos.

1.2 Justificativa

A partir das informações disponibilizadas anteriormente, foi possível perceber que a classificação manual das notas fiscais dificulta o trabalho dos órgãos reguladores, graças a diversos motivos já apresentados, o que pode influenciar negativamente o arrecadamento de tributos, e a identificação para um produto que tem quantidades expressivas de vendas em território nacional como água mineral.

Assim, surge a oportunidade de desenvolver um módulo do classificador de produtos fiscais para a categoria de águas minerais, utilizando técnicas de similaridade e processamento de linguagem natural para facilitar o trabalho dos órgãos reguladores ao agilizar o processo de identificação de pautas fiscais, visando assim tornar o arrecadamento de tributos mais rápido, fácil e eficaz para os produtos correspondentes à categoria de água mineral.

1.3 Estrutura do trabalho

Este trabalho está organizado em 5 capítulos. O primeiro capítulo, mostra a introdução, os objetivos, justificativa usadas neste documento, para que todos os leitores entendam melhor a dinâmica do trabalho e seus objetivos.

No segundo capítulo é apresentado o referencial teórico, tópico esse que foca em trazer o detalhamento dos principais conceitos estudados para o desenvolvimento do trabalho, nele é exposto os conceitos e características de cada tópico, unindo a prática do desenvolvimento do sistema com toda a teoria estudada durante o curso.

O capítulo 3 aborda a metodologia utilizada desde a discussão das regras de negócio até o fim do desenvolvimento. O foco dessa seção é mostrar todas as escolhas metodológicas tomadas para a produção do atual trabalho.

No capítulo 4, são discutidos os resultados obtidos pelo classificador, demonstrando com exemplos, como e quais resultados são obtidos.

O capítulo 5 trata das considerações finais a respeito do trabalho, sobre a importância do seu desenvolvimento, e sua contribuição na completude do projeto, além de explicar as limitações que o classificador teria por lidar com o componente humano.

O último capítulo expõe as referências, onde são citadas todas as informações acerca das bibliografias, artigos e páginas da web utilizadas para embasar e adquirir o conhecimento contido neste trabalho.

2 REFERENCIAL TEÓRICO

A seguir será apresentado o referencial teórico que trata os conceitos e características abordadas neste trabalho.

2.1 Pauta fiscal

As pautas fiscais se configuram como tabelas de preços fiscais, que determinam valores presumidos para os itens contidos em cada operação, de forma a aplicar a alíquota e chegar ao quantum do tributo devido (SAMPAIO, 2012).

Além da arbitragem de valores, estes documentos possuem outras informações diversas sobre os produtos passíveis de cobrança, como sua descrição, capacidade, categoria, fabricante, identificador, valor de pauta entre outros.

Dessa forma, ao obter as descrições feitas pelos contribuintes podemos realizar uma comparação com as informações contidas na documentação, e a partir da convergência de dados, é possível obter o valor de pauta para que a tributação seja aplicada de forma adequada.

Deste modo, no trabalho em questão, os documentos de pauta fiscal funcionam de maneira semelhante a um gabarito, onde através do cálculo de similaridade, a identificação dos itens comercializados e descritos de maneira informal, torna-se possível.

Figura 1 - Exemplo de pauta fiscal de água mineral

ÁGUA MINERAL/TIPO	SQ	UNID.	ICMS ST SUGERIDO PB R$	ICMS ST SUGERIDO OUTRAS UF's R$	FUNCEP R$
Copo Descartável "PET" de 200 a 300ml (MINERAL)	7	Unid.	0,07	0,1	
Copo Descartável "PET" de 200 a 300ml (ADICIONADAS)	8	Unid.	0,06	0,08	
Garrafa "PET" de 300 a 350ml Sem Gás (MINERAL)	9	Unid.	0,08	0,11	
Garrafa "PET" de 300 a 350ml Sem Gás (ADICIONADAS)	10	Unid.	0,06	0,09	
Garrafa "PET" de 300 a 350ml Com Gás (MINERAL)	11	Unid.	0,09	0,13	0,01
Garrafa "PET" de 300 a 350ml Com Gás (ADICIONADAS)	12	Unid.	0,07	0,1	0,008
Garrafa "PET" de 351 a 600ml Sem Gás (MINERAL)	13	Unid.	0,14	0,2	
Garrafa "PET" de 351 a 600ml Sem Gás (ADICIONADAS)	14	Unid.	0,11	0,16	
Garrafa "PET" de 351 a 600ml Com Gás (MINERAL)	15	Unid.	0,18	0,25	0,015
Garrafa "PET" de 351 a 600ml Com Gás (ADICIONADAS)	16	Unid.	0,14	0,2	0,012
Garrafa "PET" de 1.500ml Sem Gás (MINERAL)	17	Unid.	0,23	0,28	
Garrafa "PET" de 1.500ml Sem Gás (ADICIONADAS)	18	Unid.	0,18	0,22	
Garrafa "PET" de 1.500ml Com Gás (MINERAL)	19		0,27	0,32	0,026
Garrafa "PET" de 1.500ml Com Gás (ADICIONADAS)	20	Unid.	0,23	0,28	0,021
Mini Pote "PET" de 05 litros descartável (MINERAL)	21	Unid.	0,45	0,53	
Mini Pote "PET" de 05 litros descartável (ADICIONADAS)	22	Unid.	0,36	0,42	
Mini Pote "PET" de 10 litros descartável (MINERAL)	23	Unid.	0,91	1,06	
Mini Pote "PET" de 10 litros descartável (ADICIONADAS)	24	Unid.	0,73	0,85	
Garrafão "PET" de 20 litros retornável (MINERAL)	25	Unid.	0,45	0,53	
Garrafão "PET" de 20 litros retornável (ADICIONADAS)	26	Unid.	0,36	0,42	
Garrafão "PET" de 20 litros retornável (NATURAL)	27	Unid.	0,45	0,53	
Obs: Os produtos não elencados neste ANEXO ÚNICO, deverão ser aplicado o MVA de 100%, 120% e 140% conforme estabelece o Protocolo nº 11/1991.					

Fonte: portaria N°00034/2022/SEFAZ

2.2 Ciência de dados

A ciência de dados é uma tecnologia que permite ao usuário extrair informações valiosas de um conjunto de dados e a partir destes dados extrair conhecimento.

Dessa forma tal conhecimento, Segundo Amaral (2016, p. 3) “é a informação interpretada, entendida e aplicada para um fim”. Por conseguinte, a tomada de decisões e realização de tarefas que são manualmente inviáveis, são feitas de maneira rápida e eficiente, utilizando-se das informações obtidas.

No presente trabalho, a Ciência de Dados irá possibilitar a manipulação inicial dos dados através de sua leitura e análise para determinar quais estratégias utilizar, sabendo quais dados manter, quais excluir, quais modificar, para formar uma base dados concisa com a estratégia de classificação determinada pelo acordado entre a equipe do NUTES e o time da SEFAZ-PB.

O conjunto de dados a ser analisado consiste de uma lista de descrições de produtos relacionados à água mineral feita pelos contribuintes, e da pauta fiscal de água mineral.

Como as descrições dos contribuintes não são padronizadas, muitas delas possuem a necessidade de serem modificadas para que o classificador possa reconhecê-las, sem desconsiderá-las ou classificá-las de maneira errada.

Através da ciência de dados, torna-se possível a análise do conjunto geral de descrições para determinar quais os erros mais comuns e quais padrões de correções são mais rentáveis, tanto em termos financeiros quanto de processamento.

Figura 2 - Exemplo de conjunto de dados de descrições dos contribuintes

produtos	sqPauta
*AGUA INDAIA MINERAL 1,5ML	40
*AGUA INDAIA MINERAL 500ML	40
*AGUA MINERAL INDAIA 1.5L	40
*AGUA MINERAL INDAIA 500ML	40
0000900015 - AGUA SCHIN MINER S/GAS 1,5LPET 6UN PBR	17
0000900023 - AGUA SCHIN MINER S/GAS 0,50LPET 12UN PBR	13
0000900032 - AGUA SCHIN MINER C/GAS 0,50LPET 12UN PBR	15
20 L AGUA ADICIONADA DE SAIS	42
AG ACQ PAN SG 250ML	5
AG MIN SAN PELLEG GRF C/GAS 505ML	3
AG MINER FRAN PERRIER C.TRIB.41,12%	42
AG MINER FRAN PERRIER GFA 330ML	1
AG MINER ITA SAN PELEGR GFA 250ML	2
AG MINER ITA SAN PELEGR GFA 750ML	4
AG SAN PELL CG 505ML	3
AG SAN PELL CG 750ML	4
AGUA INDAIA 500ML	40
AGUA MINERAL 20L	42
AGUA - 20L PURIFIC	42
AGUA 20 LITROS	42
AGUA 20 LTS	42
AGUA 20L INDAIA	42

Fonte: SEFAZ-PB

2.3 Processamento de Linguagem Natural (PLN)

O processamento de linguagem natural é uma área de pesquisa e aplicação que explora a possibilidade de computadores entenderem e manipularem texto ou fala em linguagem natural, a fim de criar informação a partir de dados obtidos(DÍAZ, et al, 2021).

Esse processo, que pode ser classificado como uma subárea da inteligência artificial, tem foco em manipular textos de forma a torná-los compreensíveis para as máquinas, visto que por padrão as máquinas não compreendem a linguagem natural dos seres humanos com todas as suas irregularidades.

A aplicação de PNL é fulcral, posto que as máquinas não conseguem processar a linguagem natural humana, necessitando assim de representações formais que contribuam para o seu armazenamento ou manipulação (CATAE, 2012).

Essas aplicações têm influência em variadas esferas, como por exemplo, obtenção de informação a partir de sentimentos, humor e opiniões contidas em frases escritas, possibilitando que partes interessadas possam extrair feedbacks e evoluir pontos que antes estavam deficientes.

Com relação aos textos preditivos, o Processamento de Linguagem Natural possibilita a criação de ferramentas que têm funções como corretores ortográficos e preenchimento automático, funcionalidades estas que são utilizadas em grande escala por aplicações de busca e conversação.

Outro exemplo de ferramenta seria o filtro de e-mail, onde torna-se possível a categorização das informações em classes como por exemplo, spam, e-mails sociais, promocionais, ou com informações legítimas, posto que a tecnologia irá analisar não só o conteúdo representado pelas palavras, mas também o contexto e a intenção do usuário, melhorando as buscas e resultados exibidos.

No trabalho em questão, o PNL torna-se indispensável, visto que as descrições dos produtos de águas minerais feitas pelos contribuintes, tratam-se justamente de sentenças não padronizadas, produzidas através da comunicação feita em linguagem natural humana.

Dessa forma, o Processamento de Linguagem Natural visa justamente a padronização, correção e organização destes dados, de forma que os mesmos atinjam um formato que auxilie o máximo, tanto quanto possível a ferramenta de classificação, aumentando consideravelmente a possibilidade de bons resultados.

2.4 Similaridade entre textos

Como pode-se observar nos tópicos anteriores, para que a linguagem natural possa ser processada, faz-se necessária uma "tradução" entre a linguagem natural humana para uma linguagem que as máquinas consigam processar, para isso a similaridade entre textos busca computacionalmente de forma numérica, o nível de semelhança entre palavras. Sendo considerada uma subárea do PLN, essa técnica manipula operações com texto, objetivando extrair informações relevantes.

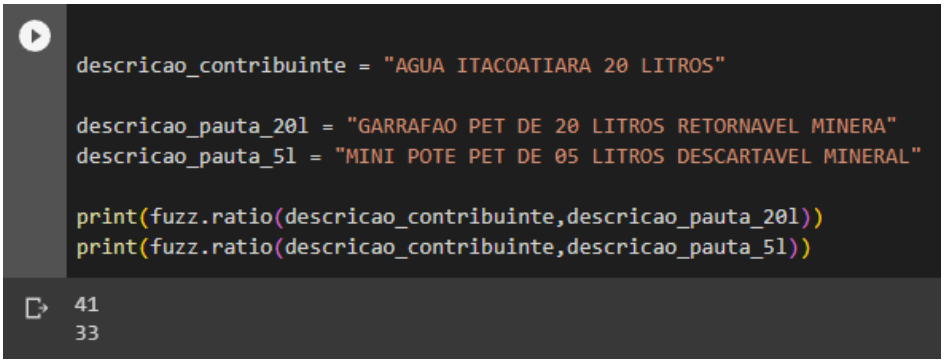
As ferramentas de busca em alguns sites na internet, são exemplos dos mais diversos contextos diferentes objetivos nos quais o cálculo da similaridade se encontra de forma a gerar resultados para o usuário.

Outro exemplo seria o FlexSTS, desenvolvido por Freire (et al, 2016), que trata-se de um framework que calcula similaridade semântica textual.Valendo citar também o trabalho desenvolvido por Cavalcanti (et al, 2017), que apresenta uma nova medida de similaridade entre sentenças especificamente em português com o objetivo de detecção de plágio interno em fóruns educacionais.

Para o trabalho discutido neste documento, a similaridade terá o papel de calcular o coeficiente de similaridade entre as sentenças não padronizadas, produzidas a partir das descrições criadas pelos contribuintes e as sentenças presentes na Pauta fiscal (SECRETARIA DA FAZENDA, Portaria nº 00034, 2022), para que dessa forma a tributação possa ser feita de maneira correta de acordo com as normas estabelecidas pela SEFAZ-PB.

Dessa forma, para que a criação do modelo se torne possível, torna-se necessário a utilização de diversas tecnologias e bibliotecas que possibilitem e facilitem o processo de desenvolvimento do mesmo. Nas seções seguintes estas tecnologias serão abordadas de forma a objetivar um melhor entendimento de suas funções no trabalho em questão.

Figura 3 - Exemplo de cálculo de similaridade



Fonte: Elaborado pelo autor, 2023.

Na figura acima podemos ver um exemplo onde calculamos a similaridade entre as descrições, comparando duas descrições feitas por contribuintes, uma tratando de um garrafão de 20 litros e outra de um mini pote de 5 litros, e respectivamente podemos observar um resultado maior para o cálculo entre ambas as descrições de 20 litros, do que a de 5 litros que é diferente da descrição encontrada na pauta.

2.5 Bibliotecas utilizadas na linguagem Python

A linguagem Python[1] é utilizada para diversos fins, sendo interpretada de alto nível e capaz de dar suporte a diversos paradigmas, ela tem aplicações satisfatórias na área de Machine Learning(ML), desenvolvimento de aplicativos, automação de aplicações, desenvolvimento web e etc. Por isso, a linguagem de programação Python, que é de código aberto, foi escolhida como a tecnologia principal neste trabalho. Ela está presente em todas as etapas, desde a análise e processamento de dados até a criação do modelo.

Na linguagem Python há um conjunto de módulos e funções úteis que reduzem o uso de código no programa, estes módulos são chamados de bibliotecas. Posto isso, as discussões a seguir irão abordar as bibliotecas python utilizadas na produção do trabalho em questão.

2.5.1 Pandas

Para o processamento dos dados, utilizaremos a biblioteca Pandas[2], a qual é de código aberto, sendo usada para análise e manipulação de dados. Desenvolvida em Python, essa biblioteca é conhecida por sua rapidez, poder, flexibilidade e facilidade de uso. É especialmente importante para análises exploratórias de dados, pois permite várias operações em bases de dados, tais como leitura, manipulação e agregação de dados de forma simplificada. Dessa forma, a biblioteca Pandas se torna fundamental para o desenvolvimento deste trabalho, já que é usada para ler, escrever e manipular as bases de dados formadas a partir de descrições de produtos relacionados a águas minerais feitas pelos contribuintes.

2.5.2 re

Neste trabalho, para manipulação de strings e expressões regulares, é utilizada a biblioteca re[3]. Suas principais funcionalidades incluem a busca, quebra e substituição de strings, as quais são essenciais, uma vez que o trabalho envolve declarações em formato de

[1] Disponível em: <https://www.python.org>. Acesso em: 12 de jun. 2023
[2] Disponível em: <https://pandas.pydata.org>. Acesso em: 12 de jun. 2023
[3] Disponível em: < https://docs.python.org/3/library/re.html>. Acesso em: 12 de jun. 2023

texto. O uso desta biblioteca torna possível manipular os dados textuais de forma mais fácil e eficiente.

2.5.3 FuzzyWuzzy

Existem diversas formas de realizar a comparação de textos em Python, incluindo a comparação padrão da linguagem, que verifica a igualdade exata dos elementos. Contudo, a biblioteca FuzzyWuzzy[4] trabalha com a similaridade de sentenças de forma não binária, permitindo uma análise mais avançada além da comparação entre palavras. Com a FuzzyWuzzy, é possível verificar não só a igualdade ou diferença entre as palavras, mas também a similaridade entre elas.

A biblioteca utiliza a distância de *Levenshtein* [5]como base para calcular o mínimo de inserções, remoções ou alterações de caracteres necessárias em uma string para torná-la idêntica a outra.Por exemplo, para comparar as palavras "tráfego" e "tráfico" utilizando a distância de *Levenshtein*, seria necessário substituir a letra "e" pela letra "i" e a letra "g" pela letra "c", resultando em um custo total de dois.

A FuzzyWuzzy é a principal biblioteca utilizada neste trabalho e as demais bibliotecas são usadas para auxiliar na obtenção do melhor resultado possível. Essa ferramenta é fundamental para o desenvolvimento deste trabalho, pois permite a comparação de strings de forma matemática.

[4] Disponível em: <https://pypi.org/project/fuzzywuzzy>. Acesso em: 12 de jun. 2023
[5] Disponível em: <https://en.wikipedia.org/wiki/Levenshtein_distance>. Acesso em: 12 de jun. 2023

3 METODOLOGIA

Os auditores fiscais responsáveis pela tributação de águas minerais no estado da Paraíba apresentaram e analisaram as principais dificuldades e desafios enfrentados ao classificar produtos de águas minerais. Eles utilizaram a pauta vigente de água e aproveitaram para discutir as regras de negócios estabelecidas através da interpretação das normas de classificação. Isso ocorre especialmente nos casos em que as regras da pauta se tornam um tanto nebulosas.

Dessa forma, ideias foram levantadas com o objetivo de criar um módulo de classificação de pautas de águas minerais, que não só supra as necessidades dos auditores fiscais como se integre ao classificador geral que está em vigência, agregando valor ao trabalho dos profissionais envolvidos e criando uma solução que automatiza de forma eficaz a parcela do processo de cobrança, relacionada aos produtos de água mineral utilizado pela SEFAZ-PB.

Foi definido um processo em três etapas, utilizando ciência de dados, processamento de linguagem natural e a técnica de similaridade FuzzyWuzzy, para possibilitar a classificação automática da Pauta Fiscal. Na Figura 4 é ilustrado o processo de classificação das descrições dos produtos contidos nas notas fiscais eletrônicas.

A primeira etapa consiste no pré-processamento das bases de dados para adequá-las aos padrões desejados, logo após, é o processo de correção das descrições dos produtos para que adotem uma padronização e em seguida, a detecção da Pauta Fiscal é realizada.

Cada etapa descrita anteriormente será explorada com o objetivo de esclarecer em detalhes todo o processo da classificação da descrição das pautas fiscais de águas minerais.

Figura 4 - Processo de classificação da pauta fiscal

Fonte: Elaborado pelo autor, 2023.

3.1 Pré-processamento da base de dados

O fluxograma apresentado na Figura 5 ilustra os passos para a aplicação do pré-processamento nas bases de dados, com o objetivo de organizá-las e manter apenas as

informações relevantes para a criação do módulo do classificador de pautas fiscais para os produtos relacionados a águas minerais.

Essa etapa é necessária, pois a quantidade de dados a serem classificados é grande e formada a partir de descrições não padronizadas de contribuintes feitas em paralelo e sem comunicação nenhuma entre si, o que pode causar muita confusão durante a classificação. Explorar e analisar os dados que serão utilizados nos permite visualizar e aplicar as transformações necessárias nesses dados, para que possam ser processados e então obter-se um resultado satisfatório posteriormente discutido e aprovado pela equipe de auditores fiscais.

Figura 5 - Fluxograma do pré-processamento da base de dados

Fonte: Elaborado pelo autor, 2023.

3.1.1 Leitura e análise exploratória dos dados

O primeiro passo a ser realizado é receber da SEFAZ-PB uma amostra de dados de descrições realizadas pelos contribuintes, a qual é utilizada para uma classificação manual dos produtos em seus respectivos *sq´s* contidos na pauta fiscal, criando assim uma espécie de gabarito que mais tarde será comparada com a classificação realizada pelo algoritmo para quantificar o quão a classificação está correta.

Durante esse processo é possível identificar possíveis processamentos, tratamentos e alterações de dados a serem aplicados, por exemplo, identificar descrições repetidas que atrapalham a visualização de acurácia final, ou em casos como da descrição a seguir "ÁGUA SCHIN MINER S/GAS 0,50LPET 12UN PBR", na pauta fiscal de águas minerais não existe nenhum intervalo ou mesmo produto com a capacidade de 0,50L, mas existe o intervalo de 351ml a 600ml, portanto se este trecho da descrição for transformado em 500ml, a sua classificação por parte do algoritmo é facilitada, pois agora encontram-se na mesma unidade de média.

3.1.2 Remoção dos dados repetidos

A remoção de dados repetidos, é uma das técnicas mais utilizadas para retirar dados desnecessários que influenciam as análises negativamente, possibilitando que o cálculo da acurácia da classificação resulte em um engano.

Evidentemente que esta técnica deve ser aplicada dependendo do contexto, no caso do trabalho em questão, levamos em conta descrições de produtos de água mineral criadas pelos contribuintes, e este tipo de dado quando possui muitas duplicatas, apenas atrapalha no tempo de processamento e na análise de eficiência da classificação, por exemplo, em um cenário hipotético, onde algumas descrições de contribuintes se repetem de forma idêntica, caso essas sejam classificadas corretamente, mas as descrições únicas forem classificadas de maneira errada, a acurácia irá apresentar um valor alto próximo dos 100% de acerto, dando a impressão que o classificador de águas minerais tem resultados ótimos quando na verdade está cometendo erros.

O mesmo vale para o caso onde os valores repetidos são classificados de maneira errada e os únicos de maneira certa, o número obtido pelo cálculo da acurácia acabaria por ser baixo e dar a impressão errada que o classificador não está qualificado para realizar a tarefa, quando na verdade estaria acertando mais vezes do que errando.

3.1.3 Seleção e organização de colunas importantes

Os dados obtidos e analisados forneceram informações importantes sobre as declarações de produtos comercializados no estado e as pautas fiscais utilizadas nas cobranças. Contudo, no processo de desenvolvimento da ferramenta para trabalhar com os textos das declarações, verificou-se que nem todas essas informações eram relevantes.

O critério utilizado para a remoção das colunas foi, selecionar para remoção aquelas que não continham nenhuma informação relevante para comparação de sentenças com as descrições contidas na pauta fiscal vigente de águas minerais, no caso deste trabalho, na base de dados bruta geralmente há uma coluna de comentários que contêm dúvidas geradas a partir da classificação manual, dúvidas essas sanadas posteriormente com os auditores fiscais responsáveis.

Como esta coluna não contém dados de embalagem, capacidade, tipo ou mesmo marca de águas minerais, ela acaba por se tornar irrelevante, então é removida da base de dados.

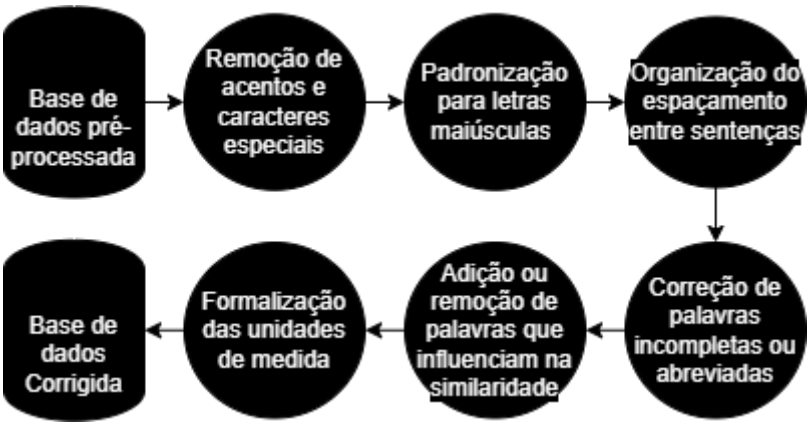
3.2 Correção das sentenças

A SEFAZ-PB tem a responsabilidade de desenvolver uma pauta fiscal, documento este que contém as descrições padrão para os produtos de água mineral, seguindo nomenclaturas encontradas no mercado e unidades de medidas estabelecidas nacionalmente, definindo assim a representação de determinado produto. Todavia, o órgão governamental não tem como obrigar os contribuintes a seguirem a mesma padronização estabelecida pela equipe interna da SEFAZ-PB. Alguns fatores linguísticos presentes na linguagem natural podem tornar as descrições que o algoritmo busca comparar com as contidas na pauta mais diferentes do que deveriam. Esses fatores incluem abreviações, erros de português, erros com relação à escrita correta de determinado termo ou marca e escolha de unidades de medidas diferentes. Por exemplo, a descrição "0000900023 - AGUA SCHIN MINER S/GAS 0,50LPET" deve ser classificada como "Garrafa PET de 351 a 600 ml Sem Gás Mineral".No entanto, ao aplicarmos a distância de Levenshtein, podemos obter um resultado indesejado. Isso ocorre quando na descrição do contribuinte encontramos "0,50L" em vez de "500ml" e as palavras "mineral", "sem" e "gás" estão escritas de maneira diferente.

Por esse motivo a formalização das descrições elaboradas pelos contribuintes é fulcral na fase de pré-processamento.

Primeiramente serão aplicadas técnicas de PLN para tratar da padronização completa das sentenças presentes na base de dados, depois serão tratadas algumas sentenças com problemas específicos. Com isso, será possível gerar bases de dados com sentenças padronizadas, que poderão ser utilizadas para detectar a Pauta Fiscal.

Figura 6 - Fluxograma da padronização de sentenças



Fonte: Elaborado pelo autor, 2023.

3.2.1 Padronização das sentenças

Na busca pela padronização integral da base de dados, foram implementados procedimentos de remoção de acentos, pontuações e caracteres especiais das descrições.

Essas informações não apresentam relevância semântica ou sintática para o contexto abordado pela ferramenta, e, se não removidas, podem prejudicar o resultado final.

3.2.2 Padronização para letras maiúsculas

Na sequência, realizou-se a padronização de todas as sentenças em caixa alta (letras maiúsculas), já que a linguagem de programação Python é Case-sensitive, ou seja, letras maiúsculas e minúsculas são consideradas diferentes.

Por essa razão, como as sentenças não possuem uma padronização prévia, é possível que os textos apresentem letras maiúsculas e minúsculas em locais distintos, o que pode interferir no processo de treinamento.

Desse modo, é importante uniformizar todas as sentenças em caixa alta, para que sentenças iguais não sejam interpretadas como diferentes.

3.2.3 Organização do espaçamento entre sentenças

É comum encontrar nas descrições dos produtos palavras concatenadas sem espaçamento, o que pode confundir a máquina, levando-a a entender essas expressões como novas palavras.

Para evitar problemas durante o treinamento do classificador, tornou-se necessário adicionar espaçamento entre as sentenças que estavam juntas.

Além disso, foi percebido que nas declarações haviam vários espaços em sequência, o que poderia gerar confusão na máquina durante a comparação das sentenças.

Desse modo, esses espaços foram identificados e corrigidos, deixando apenas um espaço entre as palavras.

3.2.4 Correção de palavras incompletas ou abreviadas

No processamento das sentenças isoladas, é frequente encontrar palavras incompletas ou abreviadas nos textos das declarações, o que pode gerar confusão para a máquina.

Isso ocorre porque a máquina pode entender palavras que possuem o mesmo significado como diferentes, devido à diferença entre uma palavra completa e outra abreviada.

Para evitar esse problema, foi necessário identificar as sentenças com esses problemas e completá-las para que ficassem na sua forma completa.

3.2.5 Adição ou remoção de palavras que influenciam no teste de similaridade

Além de especificações faltantes ou com mínimo de detalhamento, é comum encontrar em descrições de produtos palavras que não agregam valor semântico ou sintático no contexto de treinamento da ferramenta.

Essas palavras podem ser redundantes, *stopwords* ou até mesmo repetidas, prejudicando o resultado final. Para evitar esse problema, foram removidas essas palavras das descrições, mantendo apenas as informações relevantes para o cálculo da similaridade.

Muitas descrições de produtos possuem uma quantidade mínima de informações ou a especificação é insuficiente. Por isso, foi necessário complementar as descrições com palavras adicionais que possam fornecer as informações faltantes. Dessa forma, as descrições ficam completas o suficiente para que o cálculo da similaridade possa ser realizado de forma satisfatória.

As descrições de produtos contêm informações que, se não tratadas, podem interferir negativamente no resultado do treinamento da ferramenta. Isso inclui palavras repetidas ou redundantes, além de *stopwords* que não possuem relevância no contexto da análise. Para evitar esses problemas, as palavras sem relevância foram removidas das descrições, mantendo somente as informações relevantes para o cálculo da similaridade.

3.2.6 Correção de erros ortográficos

Por fim, constatou-se que frequentemente eram encontrados na descrição dos produtos diversos erros ortográficos, como por exemplo, quando o contribuinte pretende descrever um item da categoria de Água mineral em sua declaração e acaba escrevendo erroneamente "Ag Min". Tais erros podem prejudicar a análise de similaridade, uma vez que qualquer mudança nas palavras pode gerar divergência na classificação. Portanto, foram identificados e corrigidos a maioria dos erros ortográficos presentes nas descrições, finalizando assim os tratamentos realizados nas sentenças.

3.2.7 Formalização de unidades de medida

A padronização da unidade de medida da capacidade de águas minerais é essencial para a comparação e análise de produtos similares. Quando as descrições dos produtos não

utilizam a mesma unidade de medida, pode ocorrer divergência na classificação e dificultar a comparação entre os mesmos.

Um exemplo disso é a comparação entre as sentenças “AGUA SCHIN MINER C/GAS 0,50LPET 12UN PBR” contida na base de dados e “Garrafa PET de 351 a 600 ml Sem Gás Mineral” contida na pauta fiscal.

Enquanto a primeira utiliza a unidade de medida em litros, a segunda utiliza em mililitros. Essa diferença pode gerar problemas na análise de similaridade, visto que a unidade de medida é um dos fatores que influenciam na classificação dos produtos.

Ao padronizar a unidade de medida, é possível comparar produtos de forma mais precisa e confiável. Para isso, é importante que as empresas utilizem as mesmas unidades de medida em suas descrições de produtos, evitando assim a divergência na classificação e facilitando a análise de similaridade.

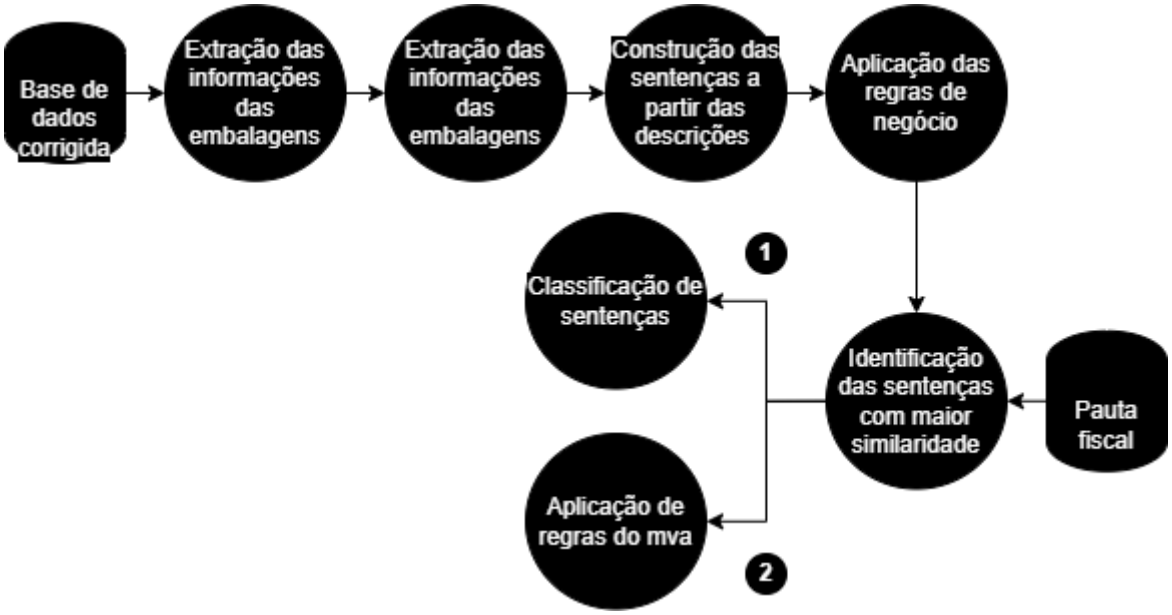
Em resumo, a padronização da unidade de medida é um fator crucial para a comparação e análise de produtos similares, como no caso das águas minerais. Por isso, é importante que as empresas adotem essa prática em suas descrições de produtos, garantindo assim a precisão e confiabilidade na análise de similaridade.

3.3 Identificação da pauta fiscal

O fluxograma apresentado na Figura 7 mostra as etapas essenciais para identificar a Pauta Fiscal dos produtos declarados pelos contribuintes do estado. Essas etapas serão executadas nas bases de dados para classificação, comparando assim com os dados contidos na Pauta, que são documentos de referência para as cobranças e que descrevem todas as especificações da categoria de produtos abordada, como a descrição, embalagem, capacidade e outras características. Essa abordagem permite obter a descrição precisa do produto, em vez de descrições incorretas, abreviadas ou com dados incompletos, como comumente ocorre nas declarações em texto livre. Ao aplicar essas etapas, é possível obter descrições precisas e prontas para serem utilizadas na identificação da Pauta Fiscal, permitindo a realização da cobrança de forma adequada.

Na etapa de identificação das sentenças com maiores similaridades, o fluxo acaba por se dividir em dois, onde no fluxo um temos a classificação de sentenças que possuem um intervalo de capacidades bem definidos e que possui uma similaridade adequada com as descrições apresentadas na pauta, já no fluxo dois podemos visualizar a classificação das descrições que não conseguiram atingir a similaridade desejada e portanto vão ser classificada segundo as regras de atribuição do MVA, como consta na portaria N°00034/2022/SEFAZ.

Figura 7 - Fluxograma da identificação de pauta fiscal



Fonte: Elaborado pelo autor, 2023.

3.3.1 Extração das informações da embalagem

Ao examinar essas bases, percebeu-se que, para obter um cálculo de similaridade mais preciso, era necessário comparar separadamente as informações do produto e da embalagem. Isso se deve ao fato de que um mesmo produto pode ser embalado de diversas maneiras, como no caso das águas minerais, que podem ser encontradas em garrafas de plástico, vidro, em diferentes tamanhos e com rótulos variados, como garrafas e garrafões de vinte litros, por exemplo. Para solucionar esse problema, utilizou-se um algoritmo que separa as informações da embalagem do restante da descrição do produto, permitindo uma comparação independente entre elas.

3.3.2 Identificação do possível intervalo de capacidade

A pauta de água minerais se difere das outras pautas de algumas maneiras, uma delas é o fato de suas descrições não conterem capacidades específicas para cada possibilidade de classificação de água mas sim um intervalo, como por exemplo “351 a 600 ml sem gás mineral” o que acabaria por prejudicar uma comparação direta com uma descrição criada pelos contribuintes que pode ser por exemplo “AGUA MINERAL 200 ML”.

Dessa forma torna-se de extrema importância identificar o intervalo no qual a capacidade contida na descrição se classifica, para que a sentença que posteriormente será

construída e comparada, não sofra um cálculo de similaridade errôneo devido a disparidades descritivas, apesar do conteúdo das sentenças ser o mesmo.

Para tal, anteriormente é feita uma rápida pesquisa na pauta para recuperar os intervalos disponíveis, já que estas são documentações que podem ser atualizadas e portanto alterar os intervalos utilizados para a classificação de águas minerais.

3.3.3 Construção de sentença a partir do conteúdo das descrições da base de dados para a classificação

Como visto no tópico anterior, as descrições feitas pelos contribuintes podem se diferir e muito das descrições contidas na pauta de água, graças ao padrão de identificação escolhido pela SEFAZ-PB que se baseia em intervalos de capacidades e algumas outras informações de embalagem que já foram obtidas antes dessa etapa.

Portanto a comparação direta da descrição que o algoritmo pretende classificar com a contida da pauta fiscal, resultaria em uma similaridade errônea que apontaria diferença entre as sentenças quando na verdade as duas representam o mesmo produto.

A solução para a problemática em questão é justamente a construção de uma nova sentença a partir dos dados contidos na descrição efetuada pelo contribuinte, de forma que o novo texto se assemelhe o tanto quanto possível das descrições contidas na pauta fiscal, permitindo que o cálculo da similaridade seja feito de maneira adequada, por exemplo "AGUA SCHIN MINER S/GAS 0,50LPET 12UN PBR" teria como nova sentença gerada a partir dela "351 A 600 ML SEM GAS MINERAL", já que contém a informação "S/GAS"(sem gás) e possui 0,50L ou 500ml que está entre 351 e 600 ml.

3.3.4 Aplicação das regras de negócio

Ao observarmos as descrições contidas na base de dados de classificação, encontramos diversas que não se enquadram diretamente em nenhuma das classificações possíveis contidas da pauta fiscal e as vezes nem mesmo nas regras definidas pela atribuição de MVA, caracterizado pelo percentual correspondente à margem de valor agregado a ser utilizada para apuração da base de cálculo relativa à substituição tributária, decorrente de operação interestadual com as mercadorias. Por exemplo, a sentença "AGUA MINERAL 20L", que não contém informações referentes a serem naturais ou não, ter embalagem retornável ou não entre outras. A solução desenvolvida foi a absorção das regras de negócio utilizada pela equipe de classificação da SEFAZ-PB pelo algoritmo, regras estas definidas a partir de entrevistas da equipe e que em sua maioria servem para generalizar algumas regras de

classificação, como por exemplo, a regra de considerar embalagens de água mineral de 20L como retornáveis e naturais quando estas informações não estiverem contida na sentença que descreve o produto em questão.

3.3.5 Identificação das sentenças com as maiores similaridades

Após a construção da sentença de comparação e aplicação de regra de negócio, procedeu-se à aplicação do cálculo de similaridade, realizado entre a sentença produzida a partir das declarações em texto livre com as descrições contidas na pauta, a fim de encontrar quais são mais similares entre si, sempre retornando um valor de similaridade gerado pelo grau de semelhança entre as descrições.

3.3.6 Classificação das sentenças

Após a obtenção do maior valor de similaridade entre a descrição em análise pelo algoritmo esta é comparada a um limiar definido por testes consecutivos buscando o limiar que produz o maior acerto possível, caso o valor calculado pela similaridade seja igual ou superior ao definido pelo limiar, significa que esta descrição pode ser classificada como a regra de pauta a qual ela é similar, atribuindo-se então o sq indicado na documentação para o produto em questão.

3.3.7 Aplicação das regras do MVA

Caso o valor de similaridade seja inferior ao limiar estabelecido, o produto em questão não pode ser classificado como uma das águas minerais descritas na pauta fiscal, um fato que no caso de tratar-se de outras pautas, haveria apenas a consideração do produto tratado não ser contemplado na pauta fiscal, todavia para o caso de águas minerais existe a aplicação do MVA, que não trabalha com descrições restritas como as da pauta,mas sim com classes mais gerais que vão comportar grande parte das descrições, que caso essa regra não existisse seriam classificados como não sendo águas minerais ou mesmo não possuindo uma classificação de pauta fiscal. Portanto adota-se o estabelecido na portaria N°00034/2022/SEFAZ. Onde quando a atribuição de um sq não é possível, o valor de MVA de acordo com a capacidade encontrada na descrição em análise é considerado, valor esse que consiste de uma porcentagem variando de 100-140%, estabelecida de acordo com as regras de classificação documentadas na portaria. Que seriam da seguinte forma: 100% para os produtos listados entre 5 a 20 litros com embalagem descartável ou retornável, 120% para os produtos listados

entre 1000ml a 2000 ml com embalagem “PET” e 140% para os produtos listados entre 200ml a 600 ml com embalagem “PET”.

4 RESULTADOS

A necessidade de criar um algoritmo que pudesse simplificar e otimizar todo o processo de cobrança de tributos do estado, surgiu após uma análise minuciosa das dificuldades enfrentadas pelos auditores fiscais da SEFAZ-PB. Com a ajuda de algumas ferramentas da linguagem de programação Python, foi possível desenvolver uma solução automática capaz de detectar a Pauta Fiscal dos produtos, utilizando o cálculo de semelhança entre as suas descrições para alguns produtos determinados pelos representantes da SEFAZ-PB. No trabalho em questão, é analisada a produção do algoritmo de classificação referente aos produtos de águas minerais.

A fim de validar os resultados obtidos, o algoritmo desenvolvido foi submetido a uma base de dados contendo 1.501 descrições de produtos da categoria de águas minerais. Essa base foi comparada com a base de Pauta respectiva, que contém informações essenciais para a realização das cobranças. E para determinar qual o limiar que possibilita o maior valor de acurácia possível, uma série de testes com limiares diferentes é realizada, na Figura 8, é mostrada a série de testes realizados com similaridades entre 8500 e 9900, e observa-se que no intervalo de 8900 a 9100 obtemos uma acurácia de 90% de acerto, o que significa que o algoritmo classificou corretamente 90% das descrições dos contribuintes.

Figura 8 - Teste de limiar e acurácia

limiar	Acurácia
8500	10%
8600	49%
8700	79%
8800	84%
8900	90%
9000	90%
9100	90%
9200	84%
9300	84%
9400	84%
9500	84%
9600	84%
9700	59%
9800	18%
9900	62%

Fonte: Elaborado pelo autor, 2023.

Na Figura 9 há um recorte da classificação realizada pelo algoritmo utilizando similaridade de 9100, onde as colunas “prodDesc”, “sqPauta” e “classificacao” representam as descrições dos contribuintes, o sq de pauta atribuído manualmente, e a classificação realizada pelo algoritmo respectivamente. E observa-se que dos 20 itens nenhum foi classificado de maneira errada.

Figura 9 - Recorte de classificação do algoritmo

prodDesc	sqPauta	classificacao
AGUA INDAIA MINERAL 1500ML	10274	10274
AGUA INDAIA MINERAL 500ML	10270	10270
AGUA MINERAL INDAIA 1500ML	10274	10274
AGUA MINERAL INDAIA 500ML	10270	10270
AGUA SCHIN MINER S GAS 1500ML PET 6UN PBR	10274	10274
AGUA SCHIN MINER S GAS 500ML PET 12UN PBR	10270	10270
AGUA SCHIN MINER COM GAS 500ML PET 12UN PBR	10272	10272
20 L AGUA ADICIONADA DE SAIS	10283	10283
AG ACQ PAN SG 250ML	140%	140%
AG MIN SAN PELLEG GRF COM GAS 505ML	10272	10272
AG MINER FRAN PERRIER C TRIB 41 12	140%	140%
AG MINER FRAN PERRIER GFA 330ML	10266	10266
AG MINER ITA SAN PELEGR GFA 250ML	140%	140%
AG MINER ITA SAN PELEGR GFA 750ML	140%	140%
AG SAN PELL COM GAS 505ML	10272	10272
AG SAN PELL COM GAS 750ML	140%	140%
AGUA INDAIA 500ML	10270	10270
AGUA MINERAL 20L	10282	10282
AGUA 20L PURIFIC	100%	100%
AGUA 20 LITROS	10282	10282

Fonte: Elaborado pelo autor, 2023.

Para ter uma melhor visualização da acurácia do algoritmo, é possível criar uma matriz de confusão que nada mais é do que uma tabela que demonstra os falsos positivos , falsos negativos , verdadeiros positivos e verdadeiros negativos. Simplificando podemos observar a quantidade de itens que foram classificados corretamente, mas também os itens que foram classificados de maneira errada, tendo uma atribuição conflitante com a contida na base de dados classificada manualmente. É possível visualizar na Figura 10 a matriz de confusão criada a partir da classificação do algoritmo, com todas as possibilidades de classificação representadas nas linhas e colunas, representando assim itens que podem ser identificados pela linha sendo classificados como itens identificados pela coluna, tendo a diagonal principal então como as classificações corretas e os outros valores como classificações indevidas daqueles itens. Na figura em questão observamos que justamente a diagonal principal possui os maiores valores evidenciando os 90% de acerto do algoritmo.

Figura 10 - Matriz de confusão

	sq	100%	120%	140%	0
sq	1164	3	0	28	2
100%	0	14	0	3	0
120%	0	0	8	1	0
140%	1	0	0	221	1
0	1	0	0	13	25

Fonte: Elaborado pelo autor, 2023.

Nos próximos parágrafos, serão apresentados alguns dos passos aplicados para a detecção da pauta.

Quadro 1 - Recorte de base de pauta

prodDesc	sqPauta
Copo Descartável 200 a 300 ml Mineral	10264
Copo Descartável 200 a 300 ml Adicionadas	10265

Fonte: Elaborado pelo autor, 2023.

Uma base de Pauta, como pode ser observado no Quadro 1, foi coletada inicialmente, como mencionado anteriormente, para servir como guia na realização das cobranças. Nessa seção da base, as principais informações utilizadas para calcular a similaridade, validar e realizar a cobrança podem ser encontradas. A coluna "prodDesc" aparece primeiro, contendo informações sobre o nome do produto, a embalagem e a capacidade de cada produto, respectivamente, usadas para calcular a similaridade entre as informações do produto. A coluna "sqPauta" está presente na base como um valor único que funciona como identificador de cada item. Esses dados são usados para definir as informações da embalagem.

Quadro 2 - Recorte de base para classificação

produtos	sqPauta/MVA
GARRAFAO ES AGUA MINERAL 20 LTS VAZIO	0
ÁGUA MINERAL 15L	100%
AGUA MINERAL S/GAS GRF PET 12X1L ACQUA PANNA	120%
ÁGUA S GÁS	140%
ÁGUA SCHIN MINER S/GAS 1,5L PET 6UN PBR	10274

Fonte: Elaborado pelo autor, 2023.

No Quadro 2 é possível visualizar um recorte da base de dados despadronizada fornecida pelos contribuintes, presente no Quadro 2.

Na coluna “produtos” encontram-se as descrições das declarações em texto livre, que serão submetidas ao cálculo de similaridade com as descrições presentes na base de pauta, com o objetivo de identificar o item descrito no campo. Por fim, na coluna “sqPauta/MVA” é possível ver o respectivo sq atribuído pela classificação manual.

Na primeira linha do Quadro 2, que contém a descrição “GARRAFAO ES AGUA MINERAL 20 LTS VAZIO”, o algoritmo não consegue encontrar um item corresponde na base de pauta, visto que, por se tratar de um garrafão vazio de 20 litros e não de um produto referente a água mineral realmente, a pauta não possui informação para a cobrança deste item. Na segunda linha do quadro 2 podemos observar a descrição “AGUA MINERAL 15L”, que na verdade não se encaixa em nenhum dos intervalos especificado pelas regras contidas na pauta, mas a capacidade informada na descrição acaba por permitir que esta possa ser classificada com o MVA de 100% que é definido pela descrição “ 100% para os produtos listados entre 5 a 20 litros com embalagem descartável ou retornável”.

Semelhante ao caso anterior a descrição”AGUA MINERAL S/GAS GRF PET 12X1L ACQUA PANNA”, que encontra-se na terceira linha do Quadro 2, não pode ser classificada diretamente com um sq de pauta, pois nenhum dos itens previstos na mesma contém um intervalo que dê margem para capacidades de um litro, todavia o MVA de 120% prevê uma possibilidade de classificação à parte para essa descrição, ”120% para os produtos listados entre 1000ml a 2000 ml com embalagem “PET”.

E ainda de maneira análoga as duas anteriores na quarta linha do Quadro 2, contém a descrição “AGUA S GAS” que nem mesmo capacidade possui, ou seja pela aplicação arbitrária na regra deveria ocorrer a não classificação do produto, mas graças à regra de negócio discutida com a equipe da SEFAZ-PB descrições que podem ser identificadas como água, mas que não que pela falta de capacidade não podem ser aplicadas nas regras, deve-se considerar o maior valor de MVA ou seja 140%, apesar da regra de descrição para esse MVA ser “140% para os produtos listados entre 200ml a 600 ml com embalagem “PET”.

O classificador é capaz de atuar eficientemente na quinta linha da Tabela 2, devido às informações suficientes presentes na descrição do item e à sua inclusão na base de Pauta. Após a aplicação do algoritmo e dos processos de pré-processamento mencionados anteriormente, a descrição é convertida em " GARRAFA PET DE 1.500ML SEM GÁS MINERAL". Em seguida, o cálculo de similaridade e identificação do item é feito com todas

as descrições da pauta, objetivando encontrar a mais similar para fins de cobrança tornando a identificação da Pauta Fiscal do produto mais fácil.

5 CONSIDERAÇÕES FINAIS

A SEFAZ-PB firmou parceria com o NUTES (UEPB) para desenvolver um classificador de produtos, para reduzir as dificuldades para cobrar produtos comercializados na Paraíba. Atualmente, uma versão do classificador está integrada ao sistema de faturamento automático da SEFAZ-PB, incluindo as categorias de bebidas quentes, cerveja,cachaça, água, refrigerante e energético, madeira, cigarro, açúcar e isotônico.

Este trabalho tem grande valia tanto para o NUTES/UEPB quanto para a SEFAZ-PB, por isso foi renovado e continua trabalhando para agregar cada vez mais categorias. Para a categoria de águas minerais, como solução, um algoritmo em python foi desenvolvido para auxiliar na classificação de pauta fiscal e tributação dos itens vendidos, com o objetivo de facilitar o trabalho dos auditores fiscais e manter a tributação eficiente. O trabalho foi bem-sucedido ao demonstrar todo o processo necessário para a estruturação e desenvolvimento da ferramenta, utilizando de tecnologias que tem aplicação no mercado,e de conhecimentos para aplicação destas tecnologias, adquiridos durante o curso de computação.

Entretanto, o classificador teve algumas limitações com relação a identificar produtos que pela descrição se assemelham muito às águas minerais, quando na verdade não o são. Portanto, futuros projetos devem buscar soluções para melhor interpretação das sentenças, possibilitando a desconsideração de descrições que à primeira vista aparentam se encaixar nas regras de classificação, mas entretanto sua real semântica não está aliada às normas de atribuição.

REFERÊNCIAS

ALLEN, James. Natural Language Processing. MASSACHUSETTS INST. TECHNOL., R.L.E. PROGR. REP., U.S.A, 2013.

AMARAL, Fernando. Introdução à Ciência de Dados: mineração de dados e big data. Rio de Janeiro: Alta Books, 2016.

CACALCANTI, Anderson; FERREIRA, Rafael; FERREIRA, Máverick; NETO, Sebastião; PASSERO, Guilherme; MIRANDA, Péricles. Uma nova abordagem para detecção de plágio em ambientes educacionais. In Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE).Disponível em: https://www.br-ie.org/pub/index.php/sbie/article/view/7646/5442. Acesso em: 14 mar. 2023.

CATAE, Fabrício. Classificação automática de texto por meio de similaridade de palavras: um algoritmo mais eficiente. São Paulo, 2012.

Kumar. Product Classification using Machine Learning-Part I. Medium, [S.l.], 2019. Disponível em: URL. Acesso em:https://towardsdatascience.com/product-classification-using-machine-learning-part-i-5a1cd0c2caf2 .16 de jun. 2023.

Freire, J., Pinheiro, V., & Feitosa, D. (2016). FlexSTS: Um Framework para Similaridade Semântica Textual. Linguamática, 8(2), 23-31. Obtido de https://www.linguamatica.com/index.php/linguamatica/article/view/v8n2-3. Acesso em: 14 mar. 2023.

Puerta-Díaz, M., de Mira, B. S., Martínez-Ávila, D., Ovalle-Perandones, M.-A., & Grácio, M. C. C. (2021). O Processamento de Linguagem Natural nos Estudos Métricos da Informação: uma análise dos artigos indexados pela Web of Science (2000- 2019). Encontros Bibli: Revista eletrônica De Biblioteconomia E Ciência Da informação, 26, 01-24. Disponível em: https://doi.org/10.5007/1518- 2924.2021.e76886. Acesso em: 14 mar. 2023.

SAMPAIO, Tereza Carolina Castro Biber. Pauta Fiscal e Perversão. Disponível em: <http://www.revistadir.mcampos.br/PRODUCAOCIENTIFICA/artigos/terezacarolinacastrobib ersampaiopautafiscalperversao.pdf>. Acesso em: 14 mar. 2023.

ZHU, Tian. & LAN, Man. 2013. ECNUCS: Measuring short text semantic equivalence using multiple similarity measurements. Atlanta, Georgia, USA, p. 124. Disponível em: https://aclanthology.org/S13-1017/ Acesso em: 14 mar. 2023.

SEFAZ. Portaria n° 00037/2022, de 12 de Março de 2022. Título da Portaria (se houver). Diário Oficial Eletrônico .Fixa valores constantes no Anexo Único, para efeito de recolhimento do ICMS devido por Substituição Tributária, nas operações internas e interestaduais com os produtos água natural, água mineral e água adicionada de sais, revogar as Portarias n°s 000173/2021/SEFAZ. e 00027/2022/SEFAZ. João Pessoa, 22 de mar. 2022.

SINHA, Arvind Kumar et al. Resume Screening Classification using Artificial Intelligence and Natural Language Processing. International Journal of Artificial Intelligence. Pune, India, 1-8, Jan de 2021.