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

PEDRO HENRIQUE DE FARIAS COSTA

USO DE TÉCNICAS DE SIMILARIDADE PARA IDENTIFICAÇÃO DE PAUTA DE PRODUTOS FISCAIS

CAMPINA GRANDE
2021

PEDRO HENRIQUE DE FARIAS COSTA

USO DE TÉCNICAS DE SIMILARIDADE PARA IDENTIFICAÇÃO DE PAUTA DE PRODUTOS FISCAIS

Trabalho de Conclusão de Curso (Artigo) apresentado ao Departamento do Curso de Ciência da Computação da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Computação.

Área de concentração: Inteligência Artificial.

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

CAMPINA GRANDE
2021

C837u Costa, Pedro Henrique de Farias.

Uso de técnicas de similaridade para identificação de pauta de produtos fiscais [manuscrito] / Pedro Henrique de Farias Costa. - 2021.

25 p.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Computação) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2021.

"Orientação : Profa. Dra. Kézia de Vasconcelos Oliveira Dantas , Departamento de Computação - CCT."

1. Ciência de dados. 2. Processamento de linguagem natural. 3. Similaridade entre textos. 4. Pauta fiscal. I. Título

21. ed. CDD 006.35

Elaborada por Lucas H. A. da Silva - CRB - 15/898 BC/UEPB

PEDRO HENRIQUE DE FARIAS COSTA

USO DE TÉCNICAS DE SIMILARIDADE PARA IDENTIFICAÇÃO DE PAUTA DE PRODUTOS FISCAIS

Trabalho de Conclusão de Curso (Artigo) apresentado ao Departamento do Curso de Ciência da Computação da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Computação.

Área de concentração: Inteligência artificial.

Aprovada em: 30 / 07 / 2021 .

BANCA EXAMINADORA

Kézia de Vasconcelos Oliveira Dantas
_____
Profa. Dra. Kézia de Vasconcelos Oliveira Dantas (Orientador)
Universidade Estadual da Paraíba (UEPB)

Paulo Barbosa
_____
Prof. Dr. Paulo Eduardo e Silva Barbosa
Universidade Estadual da Paraíba (UEPB)

Sabrina de F. Souto.
_____
Profa. Dra. Sabrina de Figueirêdo Souto
Universidade Estadual da Paraíba (UEPB)

Dedico este TCC aos meus pais, pois sempre foram e continuam sendo os maiores incentivadores para que eu consiga alcançar meus objetivos.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ML	Machine Learning
PLN	Processamento de Linguagem Natural
SEFAZ	Secretaria de Estado da Fazenda

SUMÁRIO

USO DE TÉCNICAS DE SIMILARIDADE PARA IDENTIFICAÇÃO DE PAUTA DE PRODUTOS FISCAIS

Pedro Henrique de Farias Costa[1]

RESUMO

O presente artigo descreve o processo de desenvolvimento de um classificador criado para padronizar as declarações de produtos e detectar sua pauta fiscal, através do uso da Ciência de Dados e Processamento de Linguagem Natural (PLN), aliados a uma técnica de cálculo de similaridade textual. As descrições dos produtos declarados são realizadas de maneira informal pelos contribuintes do estado da Paraíba, o que dificulta o processo de arrecadação de impostos, viabilizado pelos auditores fiscais da Secretaria de Estado da Fazenda. A partir da análise contextual proposta, foi constatado que a ferramenta desenvolvida conseguiu atingir a acurácia de cerca 95%, conseguindo classificar corretamente aproximadamente este percentual do total de descrições.

Palavras-chave: Ciência de Dados. PLN. Similaridade. Classificador. Pauta Fiscal.

ABSTRACT

This article describes the process of developing a classifier created to standardize product declarations and detect their fiscal agenda, through the use of Data Science and Natural Language Processing (NLP), allied to a textual similarity calculation technique. Descriptions of declared products are informally carried out by taxpayers in the state of Paraíba, which hinders the tax collection process, made possible by the tax auditors of the State Department of Finance. From the proposed contextual analysis, it was found that the developed tool managed to achieve an accuracy of around 95%, managing to correctly classify approximately this percentage of the total descriptions.

Keywords: Data Science. NLP. Similarity. Classifier. Fiscal Agenda.

[1] Graduando no curso de Ciência da Computação da Universidade Estadual da Paraíba (UEPB). E-mail: pedro.farias@aluno.uepb.edu.br.

1 INTRODUÇÃO

Todos os estados do Brasil possuem grande parte da arrecadação dos tributos estaduais controlada pelas Secretarias de Estado da Fazenda (SEFAZ). Esse órgão, que é vinculado ao Ministério da Fazenda do Brasil, atua também fiscalizando as finanças dos estados por meio de várias atividades, com o objetivo de garantir que as obrigações fiscais sejam cumpridas corretamente pelas empresas.

Muitas dessas tarefas são viabilizadas por meio do trabalho realizado pelos auditores fiscais que supervisionam o funcionamento do sistema tributário no país, analisando os pagamentos de impostos e verificando possíveis casos de sonegação ou fraudes por parte dos contribuintes do estado.

No âmbito regional, atualmente, os auditores fiscais da Secretaria de Estado da Fazenda da Paraíba (SEFAZ-PB) têm enfrentado desafios no que diz respeito a realização de cobrança das massivas quantidades de notas fiscais, para cumprir com o seu papel e arrecadar corretamente os tributos que beneficiam o estado.

É comum que os auditores fiscais encontrem nas milhares de descrições de produtos presentes nas declarações das notas fiscais, informações erradas e confusas, sejam elas por motivos de engano ou até mesmo de forma intencional com o objetivo de fraudar a cobrança realizada, visto que não há uma padronização definida para a escrita dessas declarações. Dessa forma, episódios ilícitos são facilitados, já que os contribuintes tiram proveitos do cenário atual para fraudar as declarações e sonegar os impostos, se beneficiando ilegalmente dos tributos que deveriam ser pagos e, gerando no fim do processo, um déficit na arrecadação tributária do estado. Assim, fica clara a grande responsabilidade e as diversas dificuldades enfrentadas pelos auditores fiscais da SEFAZ-PB na atual conjuntura, para cobrar de forma rápida e eficiente os tributos de todos os produtos que são comercializados dentro do Estado.

Esse problema fica ainda mais acentuado quando se trabalha com quantidades exorbitantes de cobranças, visto que pelo fato de não haver um padrão na declaração dos produtos, aliado ao fato de que milhares de produtos são comercializados diariamente e que todos eles devem ser cobrados, torna-se inviável a cobrança rápida da totalidade dessas inúmeras notas, quando essas são realizadas manualmente ou por sistemas que não são bem construídos e evoluídos.

Neste cenário pode-se afirmar que é de fundamental importância que todos os produtos que são comercializados no estado venham ser tributados corretamente de acordo com a Pauta Fiscal[2], definida como sendo a fixação da obrigação tributária pelo poder público, por um valor pré-fixado da operação, tomado como teto, independente do efetivo e real valor da operação. Além disso, é necessário que as fraudes e sonegações vindas dos contribuintes sejam detectadas e evitadas, a fim de impedir déficit na arrecadação tributária estadual e assim fazer com que esses recursos voltem para a população em forma de benefícios nas mais diversas áreas. É interessante frisar ainda a importância de que todo o trabalho empenhado no processo da realização das cobranças venha demandar o mínimo de esforço possível.

---

[2] Disponível em: <http://www.e-auditoria.com.br/publicacoes/artigos/o-que-e-e-como-funciona-pauta-fiscal/#:~:text=A%20Pauta%20Fiscal%20%C3%A9%20um,e%20real%20valor%20da%20opera%C3%A7%C3%A3o>. Acesso em: 10 de set. 2021

Diante do exposto, é possível fazer uso da tecnologia e de algumas das suas diversas técnicas, com o objetivo de criar e utilizar algoritmos que possam contribuir com todo esse processo e que tornem viável e praticável todas essas melhorias, facilitando o trabalho dos auditores e mantendo a qualidade na realização das cobranças.

Dessa forma, esta pesquisa objetiva demonstrar a aplicação do Processamento de Linguagem Natural (PLN), da Ciência de Dados e de uma técnica de similaridade textual para aperfeiçoar o processo utilizado na realização das cobranças aplicadas nos produtos que são comercializados dentro do Estado da Paraíba, além de reduzir o trabalho manual que atualmente é desempenhado pelos auditores fiscais do estado. O PLN se mostra essencial visto que é uma tecnologia capaz de analisar, entender e produzir a linguagem humana em diversos idiomas distintos (ALLEN, 2003, tradução nossa), além disso para atuar de forma correta e com ferramentas apropriadas na manipulação e entendimento dos dados é fundamental a aplicação da Ciência de Dados que estuda o dado em todo seu ciclo de vida, da produção ao descarte (AMARAL, 2016), não menos importante para o sucesso do projeto é a utilização de uma técnica que possibilite realizar cálculo de similaridade textual que tem por objetivo medir o grau de equivalência semântica entre duas sentenças (ZHU et al., 2013).

Nas próximas seções, serão abordados os tópicos de fundamentação teórica, com conceitos relacionados às ferramentas utilizadas no desenvolvimento do projeto, baseados em referências científicas já debatidas na academia, além da metodologia usada durante todo o processo de criação da ferramenta de classificação e os resultados gerados com uso da tecnologia desenvolvida.

2 CONCEITOS E TEORIAS

Neste capítulo serão explanados os conceitos que tratam das áreas de conhecimento presentes na pesquisa. Serão estudados conceitos sobre Ciência de dados, Processamento de Linguagem Natural, similaridade entre textos e pauta fiscal, além das bibliotecas Python utilizadas no desenvolvimento da aplicação.

2.1 Ciência de Dados

É uma área que combina vários campos de conhecimento, como estatística, matemática, computação e áreas de atuação, com o objetivo de analisar dados advindos dos mais diversos dispositivos como celulares, sensores, dispositivos vestíveis, sistemas etc. E que podem ser encontrados no formato eletrônico, não eletrônico, analógico ou digital (AMARAL, 2016).

A Ciência de Dados é uma tecnologia que possibilita ao usuário obter informações valiosas de dados trabalhados e extrair deles conhecimento. Tal conhecimento, segundo Amaral (2016, p. 3), “é a informação interpretada, entendida e aplicada para um fim”. Assim, com posse dessas informações, é possível fazer tomadas de decisões e realizar tarefas que são manualmente inviáveis, de modo extremamente rápido e geralmente com maior êxito quando comparado ao trabalho manual.

A Ciência de Dados está atualmente em grande ascensão, pelo fato de conseguir trabalhar de forma eficiente com grandes volumes de dados, e trazer resultados positivos nas mais diversas áreas, como por exemplo no setor financeiro,

onde ela tem sido utilizada em grande escala para predizer futuros comportamentos de clientes, e, dessa forma, permitir a análise de risco de forma mais assertiva, possibilitando assim que instituições disponibilizem créditos para os usuários com mais segurança e consequentemente com menos perdas de arrecadação.

Além disso, a ciência de dados tem sido amplamente utilizada nos sistemas de recomendação, onde é realizado o mapeamento do perfil do usuário e sugeridos para eles conteúdos mais personalizados e que se enquadrem melhor nas suas preferências. É cabível ainda citar a atuação da Ciência de Dados na área da saúde, onde é utilizada para analisar e comparar diversos tipos de vírus, bactérias, sintomas e diagnósticos de forma extremamente rápida, possibilitando a detecção de determinadas doenças através de informações sobre os pacientes, sem a necessidade exclusiva de uma análise médica.

Na presente pesquisa, a Ciência de Dados vai possibilitar a leitura e visualização dos dados trabalhados, além de permitir que sejam feitas manipulações e operações nas bases de dados de modo a tratar, remover e alterar informações nas bases, com o objetivo de organizar os dados e fazer com que permaneçam apenas as informações relevantes para o desenvolvimento do classificador.

2.2 Processamento de Linguagem Natural (PLN)

Para contribuir com o desenvolvimento da solução, é fundamental a utilização de técnicas de Processamento de Linguagem Natural (PLN). Esta é uma área de pesquisa e aplicação que explora como os computadores podem ser usados para entender e manipular texto ou fala em linguagem natural, a fim de fazer coisas úteis (DÍAZ, et al, 2021). Essa subárea da inteligência artificial tem foco em trabalhar com textos de forma a torná-los compreensíveis e úteis para as máquinas, visto que nativamente as máquinas não têm capacidade de compreender a linguagem natural dos seres humanos.

A aplicação do PLN é essencial porque, pelo fato das máquinas não conseguirem trabalhar com a linguagem natural humana, elas necessitam de representações formais que venham contribuir com o seu armazenamento e manipulação (CATAE, 2012). Essas aplicações acontecem nas mais variadas formas e em diversos segmentos, por exemplo, trabalhando com a análise de textos é possível obter informações sobre o sentimento, humor e opinião que está imposto na frase escrita, oportunizando assim que empresas possam extrair feedbacks e evoluir em certos pontos de falha.

No segmento de textos preditivos o PLN possibilita a criação de corretores ortográficos e ferramentas de preenchimento automático que são utilizadas em grande escala hoje pelos mecanismos de busca e de conversação. Outra ferramenta interessante desenvolvida a partir do uso do PLN são os filtros de e-mail, onde é possível detectar se um e-mail recebido é na verdade um spam ou se é uma informação legitima. Também permite categorizar as informações como por exemplo se são e-mails sociais ou referentes a promoções, e é bastante utilizado nos resultados de pesquisas, já que os mecanismos vão fazer uso da tecnologia para analisar não apenas as palavras pesquisadas, mas também o contexto e a intenção do usuário, melhorando assim a busca e consequentemente os resultados a serem exibidos.

Para o desenvolvimento da ferramenta analisada nessa pesquisa, o PLN se mostra como uma tecnologia indispensável, visto que serão trabalhadas sentenças não padronizadas e através do seu uso vai ser possível padronizar, corrigir e organizar as descrições de maneira que fiquem no formato ideal de uso para que a ferramenta seja desenvolvida com êxito e consiga gerar bons resultados.

2.3 Similaridade entre textos

Outra tecnologia que tem capacidade de contribuir no desenvolvimento da solução é a similaridade entre textos, pois busca mostrar computacionalmente, de forma numérica, qual o nível de semelhança entre palavras. Essa técnica é considerada uma subárea do PLN, manipulando operações com textos com o objetivo de extrair informações.

O cálculo da similaridade entre textos pode ser aplicado em diversos contextos diferentes e com os mais variados objetivos, como por exemplo, nas ferramentas de busca que utilizam além de outras técnicas a similaridade para gerar resultados para o usuário. Ou, no FlexSTS desenvolvido por Freire (et al, 2016) que é um framework para similaridade Semântica textual. Cabe citar também o trabalho desenvolvido por Cavalcanti (et al, 2017), que propõe uma nova medida de similaridade entre sentenças em português para detecção de plágio interno em fóruns educacionais e o frame.

Para o andamento do projeto, o cálculo da similaridade entre textos será aplicada com o objetivo de obter o coeficiente de similaridade entre as sentenças não padronizadas contendo descrições dos produtos por parte dos contribuintes e as sentenças presentes na Pauta fiscal, para que dessa forma os itens venham ser cobrados de acordo com as normas estabelecidas pela SEFAZ-PB.

Para criação do modelo, foi necessário fazer uso de diversas tecnologias e bibliotecas que possibilitam e facilitam todo o processo de desenvolvimento. Nas seções a seguir elas serão abordadas para melhor entendimento de seus papéis no contexto deste trabalho.

2.4 Pauta fiscal

As pautas fiscais são basicamente tabelas de preços fiscais que arbitram os valores presumidos dos itens em cada operação, a fim de aplicar a alíquota e chegar ao quantum do tributo devido (SAMPAIO, 2012). Esses arquivos possuem diversas informações sobre os produtos elegíveis para cobrança, como sua categoria, fabricante, descrição, marca, embalagem, capacidade, identificador, valor de pauta etc. A partir da descrição do item, informações e capacidade de armazenamento da embalagem, é possível obter o valor de pauta e aplicar a tributação.

Dentro do escopo deste trabalho, os arquivos de pauta fiscal[3] serão utilizados de modo que as informações mencionadas anteriormente, disponibilizadas pelas pautas, funcionarão como um gabarito para que, através do cálculo da similaridade, seja possível realizar a identificação dos itens que são comercializados e descritos de modo informal.

[3] Disponível em: <https://www.sefaz.pb.gov.br/attachments/article/5753/ANEXO%20DA%20PORTARIA%2000101%202018%20%20PAUTA%20DE%20BEBIDAS.pdf>. Acesso em: 9 de set. 2021

2.5 Bibliotecas Python

A linguagem de programação adotada no desenvolvimento do projeto foi Python[4] uma linguagem de programação interpretada, de alto nível e que dá suporte a vários paradigmas. É ainda uma linguagem de código aberto, utilizada para diversos fins, como em aplicações com ciência de dados, Machine Learning (ML), desenvolvimento web, desenvolvimento de aplicativos, automação de aplicações, etc.

Python é a principal tecnologia utilizada nesta pesquisa, visto que está presente em toda parte de seu desenvolvimento, desde o trabalho com o processamento dos dados, até a criação do modelo. Desse modo, as seções posteriores vão abordar de forma mais específica as bibliotecas do Python que foram utilizadas no desenvolvimento da solução.

2.5.1 FuzzyWuzzy

No Python, existem diversas formas de comparar textos, uma delas é a comparação comum da linguagem, onde é verificado se os elementos são idênticos. A biblioteca FuzzyWuzzy[5], no entanto, vai trabalhar com similaridade de sentenças de forma não binária, ou seja, a partir dela é possível ir além da simples comparação entre palavras onde há apenas a verificação se elas são totalmente iguais ou diferentes. Isso se dá pelo fato de que a biblioteca utiliza como base a distância de Levenshtein, que tem por objetivo calcular o número mínimo de inserções, remoções ou alterações de caracteres em uma string até que ela se torne idêntica a outra.

Para melhor visualização do cálculo a partir da distância de Levenshtein é possível ver a aplicação do método nas strings “amor” e “sabor” de modo que para transformar a palavra “amor” na palavra “sabor” é necessário realizar a adição da letra “s” no início da palavra e a substituição da letra “m” pela letra “b”, gerando assim um custo igual a 2.

Todas as bibliotecas utilizadas têm fundamental importância no desenvolvimento do projeto. Entretanto, a FuzzyWuzzy ganha destaque pelo fato de ser a ideia central da solução, sendo as demais auxiliares para que o resultado obtido a partir do seu uso seja o melhor possível.

2.5.2 Pandas

Para contribuir com o processamento dos dados será utilizada a biblioteca Pandas[6], que é uma biblioteca de código aberto, utilizada para manipulação e análise de dados. Ela foi desenvolvida na linguagem de programação Python e é caracterizada por ser rápida, potente, flexível e fácil de usar. Mais especificamente, é uma biblioteca fundamental para realização de análises exploratórias de dados, já que possibilita a realização de diversas operações em bases de dados, como leitura, manipulação e agregação de dados de forma simplificada.

A biblioteca Pandas é de extrema importância para o desenvolvimento deste trabalho, pois por meio dela são realizadas as leituras, escritas e manipulações das bases de dados trabalhadas.

---

[4] Disponível em: <https://www.python.org>. Acesso em: 9 de set. 2021
[5] Disponível em: <https://pypi.org/project/fuzzywuzzy>. Acesso em: 9 de set. 2021
[6] Disponível em: <https://pandas.pydata.org>. Acesso em: 9 de set. 2021

2.5.3 re

Com o uso da biblioteca re[7] é possível manipular strings e expressões regulares, tendo como suas principais possibilidades de trabalho a busca, quebra e substituição de strings. Sua utilização é importante neste trabalho, tendo em vista que estão sendo trabalhadas declarações em texto livro. Essa biblioteca possibilita uma maior facilidade no trabalho de manipulação dos dados textuais.

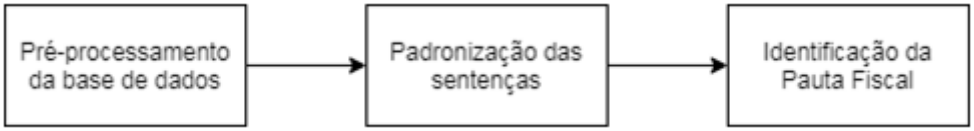
3 METODOLOGIA

A partir da constante interação entre as partes das áreas de conhecimento envolvidas, foram apresentadas e analisadas as principais dificuldades e problemáticas enfrentadas no dia a dia pelos profissionais responsáveis por todo o contexto de cobranças realizadas no estado da Paraíba.

A partir disso, foram levantadas ideias com o objetivo de aperfeiçoar o processo utilizado atualmente na realização das cobranças, além de buscar agregar mais qualidade para o trabalho dos profissionais envolvidos, focando principalmente na criação de uma solução que pudesse automatizar de forma eficaz e célere parte do processo de cobrança utilizado no âmbito da SEFAZ- PB.

Dessa forma, foram definidas três grandes etapas a serem aplicadas, para tornar possível a detecção da Pauta Fiscal, fazendo uso da ciência de dados aliada aos recursos do processamento de linguagem natural e da técnica de similaridade FuzzyWuzzy. A Figura 1 demonstra o processo necessário para a realização da classificação das descrições dos produtos informados nas notas fiscais eletrônicas, de maneira que inicialmente será aplicado o Pré-processamento nas bases de dados para que possam ficar adequadas aos padrões desejáveis de utilização, além disso os textos contendo as descrições do produtos serão padronizados para que finalmente possa ser realizada a detecção da Pauta Fiscal. A posteriori, os passos citados anteriormente serão abordados com mais detalhes, para que fique claro todo o processo praticado até efetiva classificação da descrição do produto.

Figura 1 - Processo para identificação da Pauta Fiscal

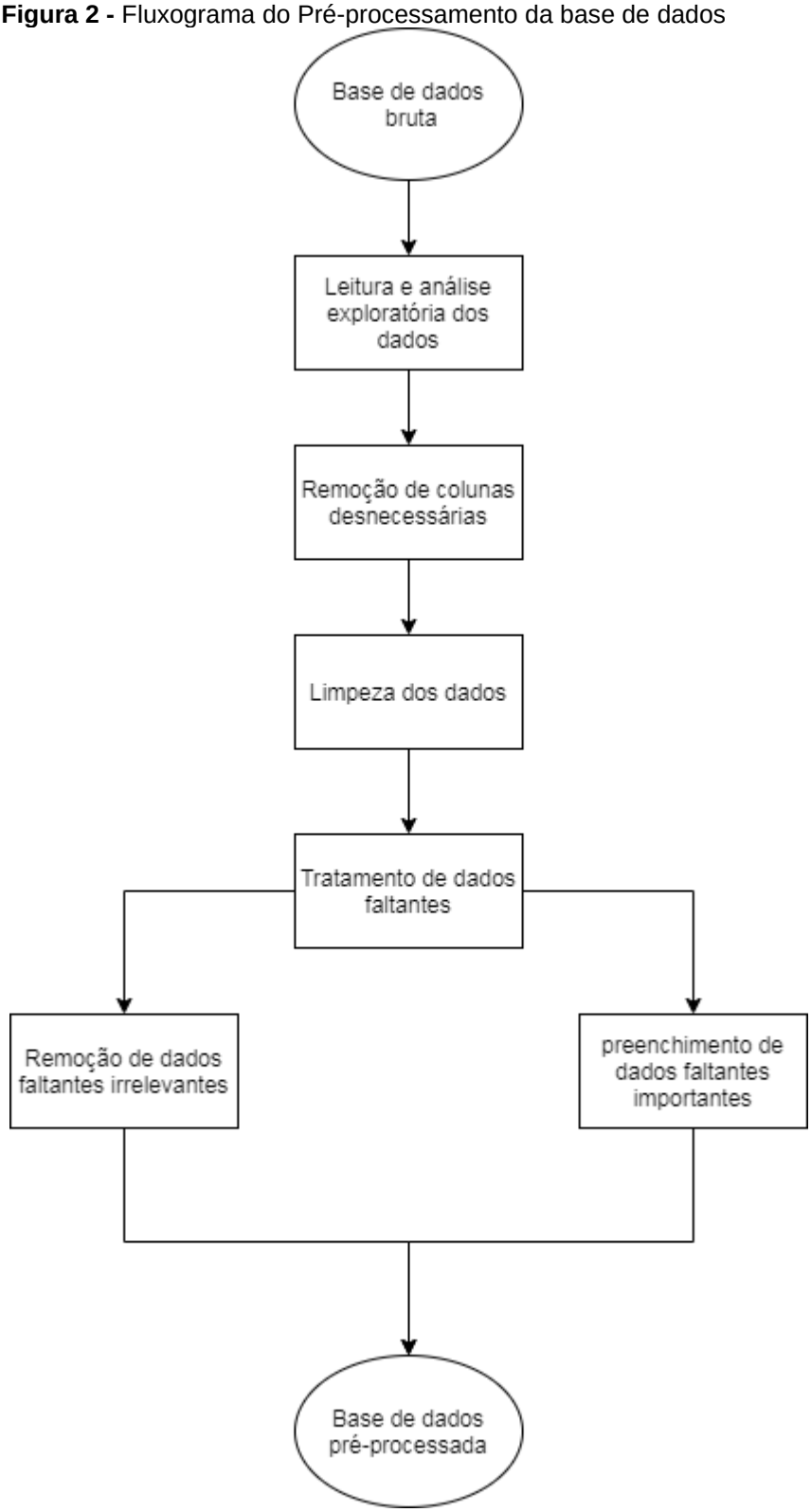


3.1 Pré-processamento da base de dados

O fluxograma abaixo (ver Figura 2) demonstra os passos da aplicação do Pré-processamento nas bases de dados, objetivando organizá-las para que permaneçam apenas informações relevantes a serem utilizadas para criação do classificador. Desse modo, as bases de dados brutas serão submetidas a todos os passos explanados posteriormente, de maneira a gerar bases de dados pré-processadas, com os problemas iniciais resolvidos e os possíveis problemas futuros mitigados, além disso, as bases ficarão organizadas para ser utilizadas nos passos posteriores.

[7] Disponível em: < https://docs.python.org/3/library/re.html>. Acesso em: 9 de set. 2021

Figura 2 - Fluxograma do Pré-processamento da base de dados



3.1.1 Leitura e análise exploratória dos dados

A ciência de dados é uma tecnologia fundamental quando há a necessidade de visualizar, organizar, limpar e extrair informações de dados em grandes quantidades. Dessa forma, em posse das bases de dados e com o auxílio da

biblioteca Pandas, foi realizada a leitura dos dados para uma melhor visualização e manipulação dos mesmos.

Posteriormente, foi feita uma análise exploratória das informações, buscando observar de forma geral possíveis processamentos, alterações e tratamentos a serem aplicados, além de potenciais comportamentos anômalos e problemas que pudessem surgir com os dados durante o processo de criação do classificador.

3.1.2 Remoção de colunas desnecessárias

Os dados obtidos e analisados trazem diversas informações sobre as declarações dos produtos comercializados no estado e as pautas fiscais utilizadas nas cobranças. No entanto, nem todas as informações se mostraram uteis para o desenvolvimento da ferramenta que vai trabalhar com os textos das declarações. Desse modo, foram retiradas as colunas que não demonstraram relevância na criação do classificador, deixando assim apenas as colunas que possuem dados que contribuem com a assertividade da classificação.

3.1.3 Limpeza dos dados

Em seguida, com as colunas a serem trabalhadas já definidas, foi efetuada uma limpeza nos dados, objetivando retirar declarações que não possuíam informações sobre produtos ou que apresentavam dados não condizentes com os campos definidos, buscando assim evitar futuros problemas onde o classificador poderia gerar resultados errados e confusos por fazer uso de informações que fogem totalmente do contexto.

3.1.4 Tratamento de dados faltantes

Durante o desenvolvimento de aplicações que fazem manipulações em bases de dados, é comum a presença de informações faltantes, seja por falta de preenchimento, problemas no armazenamento ou até mesmo por algum motivo intencional. Assim, é preciso tratar esses dados faltantes com cautela para não deixar na base descrições incompletas que atrapalham o andamento da solução, mas também para não perder informações importantes.

3.1.5 Remoção de dados faltantes irrelevantes

Existem diversas formas diferentes de lidar com dados faltantes, seja aplicando média entre os dados, fazendo comparações ou até mesmo removendo-os. Essa solução de remoção foi aplicada nas bases de dados que continham descrições com uma grande quantidade de dados faltantes. Desta forma, estas foram removidas por não contribuir com o desenvolvimento da solução.

Para melhor entendimento do que foi falado é possível visualizar o exemplo da seguinte descrição de item “CERV 250ML” onde nesse caso especifico não é possível aplicar nenhuma técnica que possa complementar a descrição de modo que fique utilizável para classificação, visto que faltam informações essenciais como a marca do produto, a embalagem etc.

3.1.6 Preenchimento de dados faltantes importantes

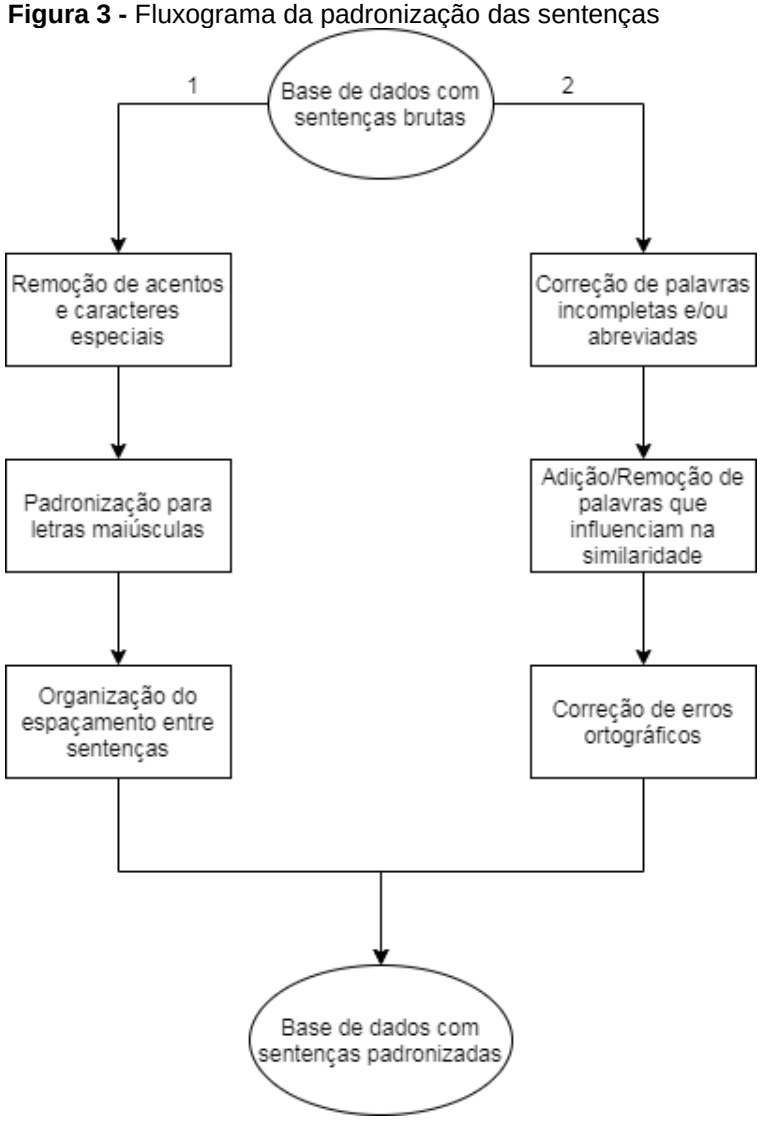
Já quando os dados se mostravam mais relevantes e não possuíam grande quantidade de informações faltantes, foi percebida a necessidade de analisá-los de forma mais minuciosa, onde alguns foram preenchidos e outros corrigidos, tudo isso de forma manual para que não houvesse o descarte equivocado de informações

relevantes e assim as informações que permaneceram pudessem ser utilizadas na criação do classificador. Um exemplo é a seguinte descrição “CERV SKOL 3 LATAS DE 250”, onde é possível complementar qual a unidade de medida do produto, tomando por base o fato de o recipiente ser lata, logo é possível completar que são 3 latas de 250 ML.

3.2 Padronização das sentenças

O fluxograma abaixo (ver Figura 3) demonstra os passos necessários para a padronização das sentenças que contém informações dos produtos. Todo esse processo é necessário para que as sentenças fiquem nos padrões ideais requeridos pelo classificador. Desse modo, as sentenças brutas presentes nas bases de dados serão alteradas com a aplicação das técnicas explicadas nas seções abaixo, gerando assim bases de dados com as sentenças padronizadas e que podem ser utilizadas para detecção da Pauta Fiscal. Para isso, os procedimentos serão aplicados seguindo duas vertentes que atuam de forma paralela, no fluxo 1 são aplicadas técnicas de PLN para tratar da padronização completa das sentenças presentes na base de dados, já no fluxo 2 as técnicas serão aplicadas para tratar de algumas sentenças que possuem problemas específicos.

Figura 3 - Fluxograma da padronização das sentenças



Essa etapa do tratamento dos dados, seguiu com aplicação além de tratar de forma mais isolada algumas sentenças que possuem problemas específicos.

3.2.1 Remoção de acentos e caracteres especiais

No que se trata da padronização da base de dados em sua totalidade, inicialmente foram aplicados procedimentos para remover acentos, pontuações e caracteres especiais presentes nas descrições, visto que no treinamento da ferramenta, essas informações não apresentam valor semântico nem sintático no contexto trabalhado, e se não removidos podem acabar interferindo de forma negativa no resultado final.

3.2.2 Padronização para letras maiúsculas

Em seguida, foram padronizadas todas a sentenças para caixa alta (letras maiúsculas), visto que a linguagem de programação Python é Case-sensitive, o que significa que letras maiúsculas são diferentes de letras minúsculas. Dessa forma, como as sentenças não possuem padronização, os textos podem vir com letras maiúsculas e minúsculas, logo se não houver essa uniformização para colocar todos os caracteres em caixa alta, o treinamento será afetado, tendo em vista que sentenças iguais podem ser entendidas como diferentes apenas por possuírem letras maiúsculas e minúsculas em locais diferentes.

3.2.3 Organização do espaçamento entre as sentenças

Nas descrições dos produtos, facilmente são encontradas várias palavras juntas e sem espaçamento, o que acaba gerando confusão para a máquina, pois assim ela vai entender que essas expressões concatenadas são novas palavras. Desse modo, tornou-se estritamente importante adicionar espaçamento entre as sentenças que estavam juntas, deixando-as corretas e evitando assim mais um problema para o treinamento do classificador.

Além disso, foi percebido que era comum encontrar nas declarações vários espaços em sequência, o que também geraria confusão na máquina durante a comparação das sentenças. Esses espaços sequenciais foram identificados em todas as declarações de produtos, e corrigidos de forma que só um espaço permaneceu separando as palavras.

3.2.4 Correção de palavras incompletas e/ou abreviadas

Tratando das sentenças isoladas, é comum encontrar palavras incompletas e abreviadas nos textos das declarações, como por exemplo quando o contribuinte vai declarar um item da categoria de Água Mineral e coloca apenas “Ag. Min”. Desse modo, para a máquina, palavras que possuem o mesmo significado são entendidas como diferentes pelo fato de uma estar completa e outra abreviada. Então, foi necessário detectar as sentenças com esses problemas e completá-las para que ficassem na sua forma completa.

3.2.5 Adição/Remoção de palavras que influenciam na similaridade

Além disso, é comum encontrar descrições onde faltam especificações do produto ou a quantidade de detalhamento é mínima. Então houve a necessidade de adicionar algumas palavras que complementam as informações das descrições para que, assim, elas fiquem completas o suficiente e quando forem submetidas aos cálculos da similaridade seja obtido um resultado satisfatório.

As descrições dos produtos geralmente possuem informações que quando não tratadas, influenciam de forma negativa no resultado do treinamento da ferramenta. São palavras repetidas ou redundantes e stopwords, que são termos irrelevantes para o entendimento da máquina, como por exemplo artigo definido e indefinido. Com isso, foram removidas das descrições as palavras que não têm relevância no cálculo da similaridade, e permaneceram apenas as informações pertinentes para o resultado almejado.

3.2.6 Correção de erros ortográficos

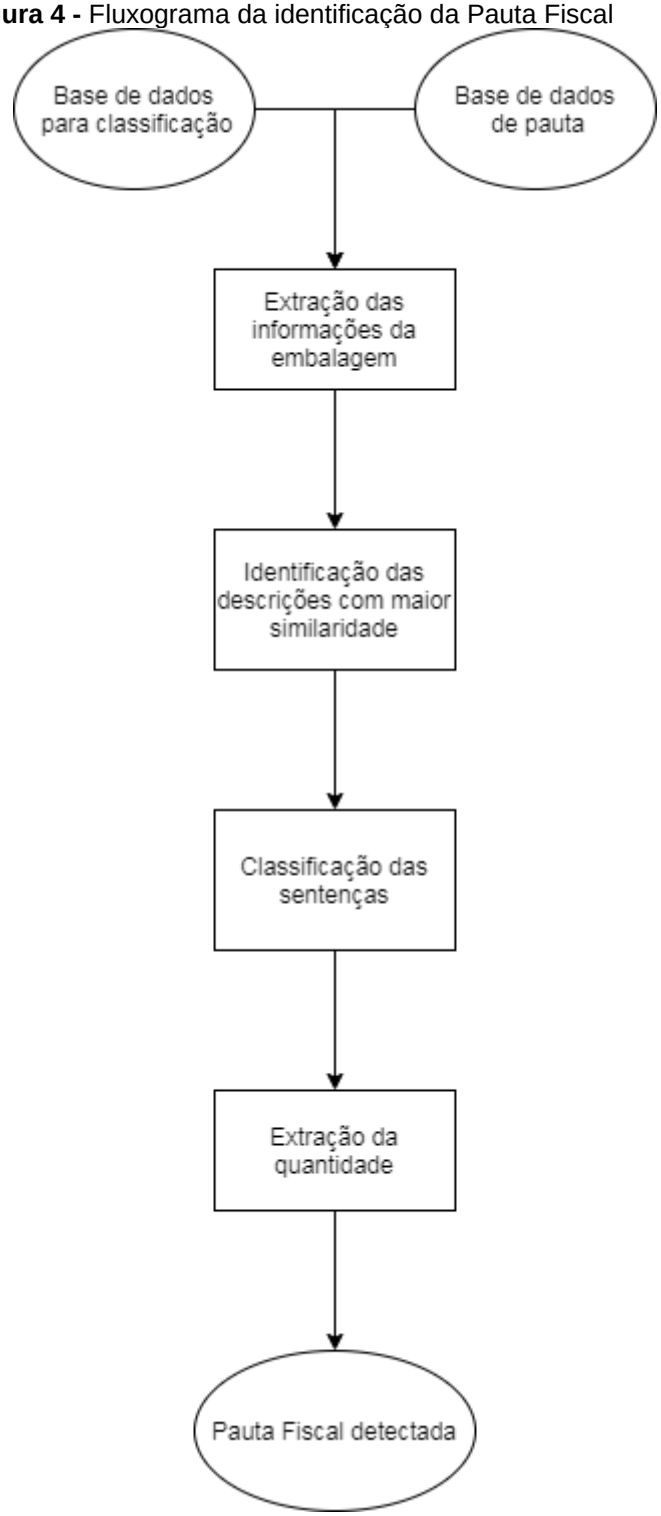
Por fim, foi percebido que com frequência eram encontrados nas descrições dos produtos diversos erros ortográficos, como por exemplo, quando o contribuinte tem o objetivo de descrever em sua declaração um item da classe de Cerveja e acaba escrevendo erroneamente “Ceveja”. Tais erros podem atrapalhar a análise da similaridade, visto que qualquer alteração nas palavras gera divergência na classificação. Assim, foram detectados e corrigidos grande parte dos erros ortográficos presentes nas descrições, finalizando assim os tratamentos realizados nas sentenças.

3.3 Identificação da pauta fiscal

O fluxograma abaixo (ver Figura 4) mostra os passos necessários para a identificação da Pauta Fiscal dos produtos comercializados e declarados por parte dos contribuintes do estado. Para isso, todos os passos demonstrados posteriormente serão aplicados nas bases de dados para classificação, que contém as declarações de produtos em texto livre, sem regras e sem formatação.

Além disso, os passos serão aplicados também nas bases de dados de Pauta, que são documentos utilizados como base para as cobranças, onde são descritas todas as especificações da categoria de produtos abordada, desde a descrição, embalagem, capacidade, entre outras características. Isso vai possibilitar a obtenção da descrição correta do produto, ao invés das descrições erradas, abreviadas ou com dados faltantes, como são originalmente. Desse modo, é possível obter as descrições corretas e prontas para serem utilizadas na detecção da Pauta Fiscal para realização da cobrança.

Figura 4 - Fluxograma da identificação da Pauta Fiscal



3.3.1 Extração das informações da embalagem

Com posse dessas bases, foi percebido que para um melhor resultado no cálculo da similaridade, era preciso comparar as informações do produto e as informações da embalagem separadamente, visto que para um mesmo produto é possível ter uma variação enorme de embalagem; por exemplo: nas descrições de cerveja, onde é possível encontrar os mais diversos recipientes como lata, latão, long neck, garrafa retornável, garrafa descartável etc.

Dessa maneira, foi aplicado um algoritmo que realiza a segregação nas descrições dos produtos, onde a parte da descrição da embalagem foi dividida do restante da descrição, para que ambas sejam comparadas separadamente.

3.3.2 Identificação das descrições com maior similaridade

Após separar as descrições da embalagem e do próprio produto, iniciou-se a aplicação do cálculo da similaridade. Primeiramente, comparando as informações dos produtos e buscando obter como resultado as sentenças que possuem maior grau de similaridade e assim saber qual é a descrição correta do produto que foi declarado em texto livre.

Por conseguinte, foi aplicado o cálculo de similaridade nas descrições das embalagens, objetivando assim obter as descrições que possuem maior similaridade e consequentemente descobrir qual a embalagem que está descrita na declaração do produto que é realizada em texto livre.

3.3.3 Classificação das sentenças

Para efetivamente classificar e obter a descrição correta para o produto, que é o necessário para identificação da Pauta Fiscal, foram analisados separadamente os índices de similaridade com as informações do produto e as informações de sua embalagem. Sendo assim, foi realizada a junção das descrições com maiores taxas de similaridade na descrição do produto e na descrição da embalagem, simultaneamente. Quando a descrição não possuía informações referentes a embalagem, a classificação se deu observando apenas as informações do produto.

3.3.4 Extração da quantidade

Outro fator de extrema importância para realização das cobranças é a informação de quantos itens foram comercializados, e muitas vezes essa informação vem de forma confusa na descrição do produto, como por exemplo, na descrição “Cerveja C/12”. Neste caso, o classificador deve ser capaz de analisar a descrição e saber que foram comercializadas 12 unidades do produto. Dessa forma, através do uso de Expressões Regulares, foram detectados padrões de escrita dessas informações da quantidade, possibilitando assim a extração da quantidade exata de produtos que foram comercializados.

4 RESULTADOS

A partir da análise das dificuldades enfrentadas pelos auditores fiscais da SEFAZ-PB na realização das cobranças dos tributos do estado, se deu a necessidade de desenvolver um algoritmo que pudesse facilitar e otimizar todo esse processo. Para isso, com o apoio de algumas ferramentas da linguagem de programação Python, foi desenvolvida uma solução automática capaz de detectar a Pauta Fiscal dos produtos através do cálculo da similaridade entre as descrições.

Para validação dos resultados obtidos, uma base contendo 10.782 dados de descrições de produtos da categoria de cerveja e chopp foi submetida ao algoritmo desenvolvido, sendo comparada com a base de Pauta de Cerveja e Chopp que contém 540 linhas com informações essenciais para a realização das cobranças. Dessa forma, as bases foram submetidas a todos os processamentos explicados anteriormente na Seção da Metodologia, e quando realizado o cálculo da similaridade foi obtida uma acurácia de cerca de 95%, tendo desta maneira errado em apenas 5% das descrições presentes na base de dados. Para uma melhor

análise dos dados, a seguir serão explanados alguns dos passos aplicados para detecção da pauta.

Tabela 1 - Recorte de base de Pauta

Descrição	sqpauta	Embalagem	Capacidade	Preco
BOHEMIA PILSEN	7987	LATA	269	2.32
STELLA ARTOIS	7910	GARRAFA VIDRO DESCARTAVEL	330	5.49
SKOL PURO MALTE	7989	LATA	269	2.03

Inicialmente foi recolhido um recorte da base de Pauta (Tabela 1), que como já mencionado, são informações que servem como guia para realização das cobranças. Nessa parte da base, podem ser vistas as principais informações que foram utilizadas para realização do cálculo da similaridade, validação e cobrança. Inicialmente é vista a coluna 'Descrição' contendo informações sobre o nome do produto, esses dados são utilizados como base para o cálculo da similaridade entre informações do produto; também está presente na base a coluna de 'sqpauta' que é o um valor único que funciona como identificador de cada item; posteriormente é possível visualizar as colunas 'Embalagem' e 'Capacidade' contendo a descrição sobre qual a embalagem e a capacidade de cada produto respectivamente. Esses dados são utilizados como base para o cálculo das informações da embalagem; por fim está presenta na base a coluna 'Preço' que contém dados sobre o valor de Pauta a ser cobrado de cada produto específico.

.

Tabela 2 - Recorte de base para classificação

EAN Com	EAN Trib	Descrição	Produto
7898963098157	7898963098157	CERV DOKTOR BRAU AMERICAN PAL 473ML	CERVEJA E CHOPE
SEM GTIN	SEM GTIN	CERVEJA HEINEKEN	CERVEJA E CHOPE
7896052605385	7896052605330	CERV SCHIN PILSEN 0,30LGFA RT 24UN	CERVEJA E CHOPE

Está exposto na Tabela 2, um recorte da base de dados contendo as informações despadronizadas vindas dos contribuintes. Nessa base é possível visualizar a coluna 'EAN Trib' que não é obrigatório e que auxilia na realização da cobrança, além disso existe na base a coluna 'Descrição' que contém as descrições das declarações que são realizadas em texto livre, esses dados serão submetidos ao cálculos de similaridade com as descrições presentes na base de Pauta para que possa ser detectado qual é realmente o item que está descrito no campo; por fim é possível visualizar a coluna 'Produto' que possui a descrição de qual a categoria do produto em questão.

Tomando por base as descrições do recorte mostrado na Tabela 2, é possível observar exemplos dos principais casos que são identificados pelo classificador, onde existem descrições que não podem ser classificadas porque por não possuem informações suficientes ou porque o produto descrito não está presente na pauta, e,

contrapartida existe o cenário onde a descrição possui as informações mínimas necessárias e pode ser submetida a cobrança.

Na primeira linha da Tabela 2, que contém a descrição ‘CERV DOKTOR BRAU AMERICAN PAL 473ML’ o classificador não consegue encontrar o item correspondente a esse na base de Pauta, visto que ela não possui informação para cobrança desse item especificamente. Na segunda linha da Tabela 2, onde a descrição do item é ‘CERVEJA HEINEKEN’ o classificador novamente falha ao encontrar o item correspondente na base de Pauta, pois as informações são insuficientes para determinar qual foi o item comercializado, visto que é impossível saber a embalagem e a capacidade a partir dessa descrição.

Na terceira linha da Tabela 2, com a descrição ‘CERV SCHIN PILSEN 0,30LGFA RT 24UN’ o classificador consegue atuar de forma eficiente, porque a descrição possui informações suficientes e além disso esse item está presente na base de Pauta. Nesse caso especificamente, com a aplicação do algoritmo os pré-processamento explanado anteriormente são aplicados e desse modo a descrição passa a ser ‘CERVEJA SCHIN PILSEN 300ML GARRAFA RETORNAVEL 24UN’.

Após isso, as informações do produto e da embalagem são extraídas, e é realizado o cálculo da similaridade e detectado qual o item correto para ser cobrado. Além disso, o algoritmo, a partir da informação ‘24UN’, faz a detecção de quantos itens foram comercializados e conclui assim a identificação da Pauta Fiscal do produto.

5 CONSIDERAÇÕES FINAIS

Quando se deu início o trabalho de pesquisa, foi possível observar a dificuldade enfrentada pelos auditores fiscais para realizar todo o processo de cobrança dos produtos comercializados no estado da Paraíba. Desse modo, houve a necessidade de análise e desenvolvimento de uma ferramenta que contribuísse com todo o processo de detecção de pauta fiscal e tributação dos itens negociados.

A partir disso, a pesquisa teve por objetivo desenvolver uma aplicação que fosse capaz de contribuir com todo o processo de cobrança executado sobre as vendas realizadas, de maneira a facilitar o trabalho dos auditores fiscais e continuar realizando uma tributação eficiente. Isto posto, é possível constatar que o objetivo definido foi alcançado, pois o projeto conseguiu demonstrar todo o processo necessário para estruturação e desenvolvimento da ferramenta desejada.

A pesquisa partiu do fato de que na declaração da venda, as descrições dos produtos não possuem formalidade, gerando assim problemas para efetuar a tributação. Para resolver esse problema, foi desenvolvida uma ferramenta de classificação capaz de padronizar essas sentenças, identificar a pauta fiscal e permitir a aplicação da cobrança.

Diante da metodologia proposta, é possível perceber que o classificador desenvolvido enfrente algumas limitações para realização de cobranças em casos que o produto descrito não está presente no arquivo de pauta, ou que há uma ausência demasiada de informações, impossibilitando desse modo que o sistema desenvolvido realize a tributação na totalidade dos itens comercializados.

Por conseguinte, é cabível que posteriormente sejam desenvolvidos e incentivados projetos que busquem atuar em cima das barreiras enfrentadas pelo classificador desenvolvido nesta pesquisa, buscando soluções para ampliar a

quantidade de itens que a ferramenta consegue cobrar de modo que essa quantia se aproxime o máximo da totalidade de itens comercializados no estado.

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.

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

CAVALCANTI, 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: 12 set. 2021

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: 12 set. 2021

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: 12 set. 2021

SAMPAIO, Tereza Carolina Castro Biber. Pauta Fiscal e Perversão. Disponível em: http://www.revistadir.mcampos.br/PRODUCAOCIENTIFICA/artigos/terezacarolinacastrobib ersampaiopautafiscalperversao.pdf. Acesso em: 12 set. 2021

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: 12 set. 2021

AGRADECIMENTOS

Agradeço acima de tudo a Deus, por sempre me abençoar e por ter me direcionado a escolher o curso de Ciência da Computação, área pela qual tenho extrema felicidade de estar inserido. Além disso, sou grato por ter me guiado e me sustentado mesmo em meio aos momentos mais difíceis, por isso a Deus toda honra e toda glória.

Sou extremamente agradecido também aos meus pais, Jânio e Sueneide, pois sempre fizeram o maior esforço para que eu pudesse ter uma boa qualidade de vida e conseguisse chegar até aqui. Não consigo explicar em palavras a gratidão por tamanho incentivo e apoio que recebi por parte deles para seguir firme e nunca desistir dos meus sonhos.

Não posso deixar de expressar minha gratidão a todos os meus familiares, em especial aos meus irmãos Jarmesson, Filipe e Diego, que além de incentivadores estiveram sempre presentes me dando o auxílio necessário para que eu conseguisse alcançar meus objetivos de vida e também de graduação.

Agradeço também a minha namorada Bruna Couto, com quem pude compartilhar de forma mais direta todo o período acadêmico, e que sempre esteve presente e disponível para compartilhar todos os momentos, desde os mais delicados até os mais especiais.

Também quero agradecer aos professores do curso de Ciência da Computação da Universidade Estadual da Paraíba por todo conhecimento compartilhado, em especial a minha orientadora Kézia de Vasconcelos Oliveira Dantas por ter me guiado durante todo o processo de produção desse artigo. Agradeço ainda ao NUTES-UEPB, local onde pude adquirir bastante conhecimento e que foi fundamental para meu crescimento como aluno e como profissional.

Por fim, registro meus agradecimentos a todos os meus amigos e colegas de curso, com quem pude dividir toda essa jornada de graduação.