Agrupamento de itens de notas fiscais referentes a produtos similares

João Pedro D. da Silva[1], Diogo Soares[1], Andre Wüst Zibetti[1], Matheus M. dos Santos[1], Renato Fileto[1], Simone Silmara Werner[1]

[1]Departamento de Informática e Estatística (INE),
Univ. Fed. de Santa Catarina (UFSC), Campus Trindade, Florianópolis-SC-Brasil

{joao, diogo}@grad.ufsc.br, matheus.m.santos@posgrad.ufsc.br
{andre.zibetti, r.fileto, simone.werner}@ufsc.br

Abstract. *Applications such as investigating prices charged in public purchases and possible irregularities require the identification of invoice items that refer to the same product. This is a challenging problem due to the lack of standardization of the textual descriptions of the products in the items. This article proposes and compares 4 methods for grouping electronic invoice items using topic modeling techniques and data like the measurement unit and the NCM (Comum nomenclature of Mercosul) code. The results indicate that the proposal allows grouping products with relatively simple descriptions and have potential to assist the grouping of items with more varied descriptions.*

Resumo. *Aplicações como a investigação de preços praticados em compras públicas e possíveis irregularidades requerem a identificação de itens similares de notas fiscais eletrônicas (NFe). Este é um problema desafiador devido à falta de padronização das descrições textuais dos produtos nos ítens. Este artigo apresenta e compara 4 métodos para agrupamento de itens de NFe usando técnicas de modelagem de tópicos e campos como unidade de medida e código NCM(Nomenclatura Comum do Mercosul). Os resultados indicam que a proposta permite agrupar alguns produtos com descrições relativamente simples e podem auxiliar no agrupamento de itens com maior variabilidade de descrições.*

1. Introdução

Uma das formas pelas quais entes públicos podem realizar a compra de bens e serviços é a licitação, processo que tem por finalidade assegurar a seleção da proposta apta a gerar o resultado de contratação mais vantajoso para a administração pública, evitar contratações com sobrepreço ou superfaturamento na execução dos contratos e incentivar a inovação e o desenvolvimento nacional sustentável [Brasil 2021]. A legislação define sobrepreço como: "preço orçado para licitação ou contratado em valor expressivamente superior aos preços referenciais de mercado". Logo, para que a existência de sobrepreço seja verificada, é necessária a comparação de preços de produtos ou serviços similares. Para verificar se um produto foi adquirido com sobrepreço, pode-se comparar o valor praticado na compra em análise com os valores praticados em outras compras públicas do mesmo produto. Para isso faz-se necessária a identificação de itens similares, desafio que não é trivial pois existem diversas formas de se referir ao um mesmo produto ou similar, na descrição textual do item adquirido.

Considerando esta problemática, [Silva et al. 2023] propuseram uma metodologia para tratamento e agrupamento das descrições de itens de licitações por meio de termos em comum nas suas descrições. Seus resultados experimentais mostram que o método é capaz de agrupar produtos similares. No entanto, essa metodologia forma múltiplos grupos para um mesmo produto, dado que gera um novo grupo sempre que um termo novo é verificado. A metodologia também é suscetível a falhas devido às diferentes formas de descrever um item e a ruído nas descrições. [Kieckbusch 2022] usa aprendizado de máquina para classificar descrições de produtos de notas fiscais em relação ao NCM (Nomenclatura Comum do Mercosul), comparando o uso de modelos supervisionados (*Support Vector Machine - SVM*) e não supervisionados (*Convolutional Neural Network - CNN*). O classificador SVM alcança melhor desempenho, com precisão de 88%. Outros trabalhos utilizam técnicas de aprendizado de máquina como árvores aleatórias para extração de informação de notas fiscais em segmentos específicos [Krieger et al. 2023], aprendizado profundo para extrair informação de imagens de notas fiscais [Yao et al. 2022] e e modelos de linguagem generativos de larga escala para extrair entidades e relações [Brinkmann et al. 2024]. Todavia, nenhum dessas propostas agrupa descrições de amplas variedades de produtos em conjuntos massivos de notas fiscais.

Este trabalho apresenta uma proposta para o uso de técnicas de modelagem de tópicos para identificar produtos similares, combinado com dados disponíveis nas notas fisicais: Nomenclatura Comum do Mercosul (NCM) e unidades de medidas para a formação dos grupos.

2. Material e Métodos

2.1. Dados analisados

Para execução deste trabalho utilizou-se um conjunto de dados de notas fiscais de compras realizadas nos municípios do estado de Santa Catarina, no período de abril de 2008 a maio de 2023. Foram considerados 11474 itens presentes nas notas fiscais, divididos em treze produtos distintos das categorias de produtos alimenticíos e combustíveis. Os produtos que constituem o conjunto de dados são apresentados na Tabela 1.

2.2. Padronização textual(PAD)

Considerando o problema de agrupamento de itens pela descrição, [Silva et al. 2023] propôs uma metodologia de pré-processamento das mesmas que consiste em padronizá-las e remover certos termos. Neste trabalho adotaram-se as seguintes etapas para o pré-processamento das descrições feitas em [Silva et al. 2023]:

***Upcasting*:** Todos os caracteres foram convertidos para maiúsculo. Essa etapa serve para diminuir a variabilidade dos dados, de forma que as mesmas palavras não difiram devido a estarem em maiúsculo ou minúsculo.

Remoção de acentuação e pontuação: Foram retiradas a pontuação e acentos do texto, com objetivo de reduzir a variabilidade das descrições.

Remoção de caracteres especiais: Todos os caracteres não alfanuméricos foram removidos. Tais caracteres, na grande maioria dos produtos, não são úteis para identifica-los, portanto sua remoção não gera prejuízo para a análise e pode reduzir a variabilidade.

Retirada de termos numéricos: Números presentes na descrição não identificados como quantidades seguidas por unidades de medida foram removidos.

Separação de quantidade e unidades de medida: Utilizando um dicionário de unidades de medidas, que está em constante atualização, as unidades presentes nas descrições e os números imediatamente anteriores a elas foram retirados das descrições. Difentemente de [Silva et al. 2023], neste trabalho essa característica foi separada, para utilização posterior na fase de agrupamento.

Outra situação observada foi a descrição do número de unidades do item (ex: 12 UN), que aparece em casos em que o produto não está sendo vendido de forma individual. As menções a unidades foram identificadas usando de um dicionário com diferentes formas de se referir a unidade. Essa informação é guardada e ultilizada para calcular o valor unitário dos produtos, possibilitando assim a comparação dos preços.

2.3. Agrupamento das descrições

Este trabalho considerou uma metodologia com quatro métodos para o agrupamento de descrições utilizando a técnica de modelagem de tópicos proposta por [Angelov 2020], combinada com o tratamento de dados e a formação de subgrupos baseada nas características identificadas nas notas fiscais.

Utilizou-se a implementação de modelagem de tópicos disponível no pacote *top2vec* [Angelov 2020]. Esta técnica foi escolhida por ser não supervisionada, não requerer como entrada o número de grupos e funciona com textos curtos, se adequando à nossa problemática. Este algoritmo parte do pressuposto de que muitos documentos com semânticas similares são indicativos da presença de um tópico. Ele toma uma lista de strings como entrada, nesse caso a lista de descrições dos itens de nota fiscal e define um número de tópicos que identifica no conjunto de dados, atribuindo a cada item um tópico. Os procedimentos de cada método avaliado são descritos a seguir.

Método 1: Utilizou-se este método como base para comparação com os demais métodos propostos. Nele não se usou nenhum tratamento inicial, apenas aplicou-se a técnica de modelagem de tópicos(TOP) diretamente nas descrições dos produtos.

Método 2: Neste método realizou-se a padronização das descrições considerando as etapas de upcasting, acentuação e pontuação, e caracteres especiais. Manteve-se os caracteres numéricos e unidades de medida. Após a padronização aplicou-se a modelagem por tópicos(TOP).

Método 3: Realizou-se a etapa de padronização textual das descrições (PAD) considerando todas suas etapas, na sequência aplicou-se a técnica de modelagem de tópicos (TOP) e no último passo utilizou-se as quantidade e unidades de medida (UM) extraídos das descrições para formar subgrupos.

Método 4: Considera o código NCM anteriormente à utilização da modelagem por tópicos para formação de grupos. Foi decidido utilizar somente os quatro primeiros dígitos do código NCM, os quais são denomidados NCM posição (NCMp). Isso porque ao utilizar uma seção menor, diminuem-se os erros por conta de preenchimentos errôneos.

A Figura 1 apresenta o fluxo de execução do método 4. Neste método o conjunto de dados constituído pelas descrições dos produtos de notas fiscais passa pelo processo de padronização e na sequência é dividido pelo código NCMp. Após este procedimento para

cada subconjunto de dados é aplicado técnica de modelagem de tópicos resultando em grupos de produtos que são novamente divididos considerando as unidades de medida.

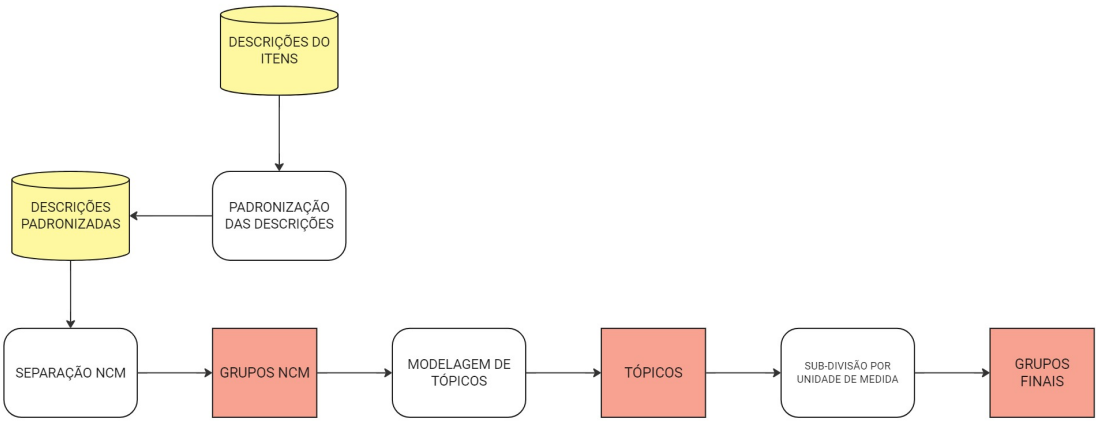

Figura 1. Fluxograma realizado para o Método 4

2.4. Métricas de avaliação

Os resultados foram analisados considerando os grupos originais apresentados na Tabela 1 e os grupos finais encontrados pelos métodos da metodologia proposta. Para cada um dos quatro métodos avaliados obteve-se a sua matriz de confusão (Figura 2), por meio da qual foi possível avaliar em mais detalhes quais os erros que ocorreram na separação dos grupos. Como métricas para os testes e validação da metodologia foram calculadas a precisão, revocação e F1 Score, métricas usadas em [Novaes et al. 2023]. Além delas também calculou-se a acurácia [Paalman et al. 2019].

3. Resultados e discussões

A Tabela 1 apresenta uma visão geral do conjunto de dados, trazendo os produtos, o número de notas com o mesmo produto, número de descrições únicas presente no conjunto e número de descrições únicas após o primeiro tratamento.

Tabela 1. Visão geral do conjunto de dados

Produto	Total de itens	Descrições únicas	Descrições únicas após padronização
(1) Água mineral 20L	1126	1126	746
(2) Água mineral 500ml	1104	1104	857
(3) Creme de leite 200g	538	538	309
(4) Detergente 500ml	1236	1236	910
(5) Diesel comum	931	59	47
(6) Gás natural	1030	18	12
(7) Gasolina aditivada	747	85	69
(8) Gasolina comum	919	435	91
(9) Leite condensado 395g	273	273	168
(10) Leite em pó 400g	868	868	661
(11) Leite UHT 1L	1136	1136	759
(12) Óleo de soja 900ml	712	712	432
(13) Sabão em pó 1kg	854	854	545

3.1. Resultados dos testes

As matrizes de confusão obtidas considerando os grupos originais apresentados na Tabela 1 e os grupos finais encontrados em cada um dos métodos avaliados são apresentadas na Figura 2. Na matriz de confusão observa-se os 13 grupos iniciais considerados corretos e a classe '*'. Para possibilitar a representação, todos os itens alocados a grupos que não sejam um dos 13 considerados corretos são classificados na classe '*'.

A partir dos dados obtidos nos experimentos e apresentados nas matrizes de confusão foram calculadas as métricas de avaliação acurácia, precisão, revocação e F1 Score de cada método. Os valores são apresentados na Tabela 2.

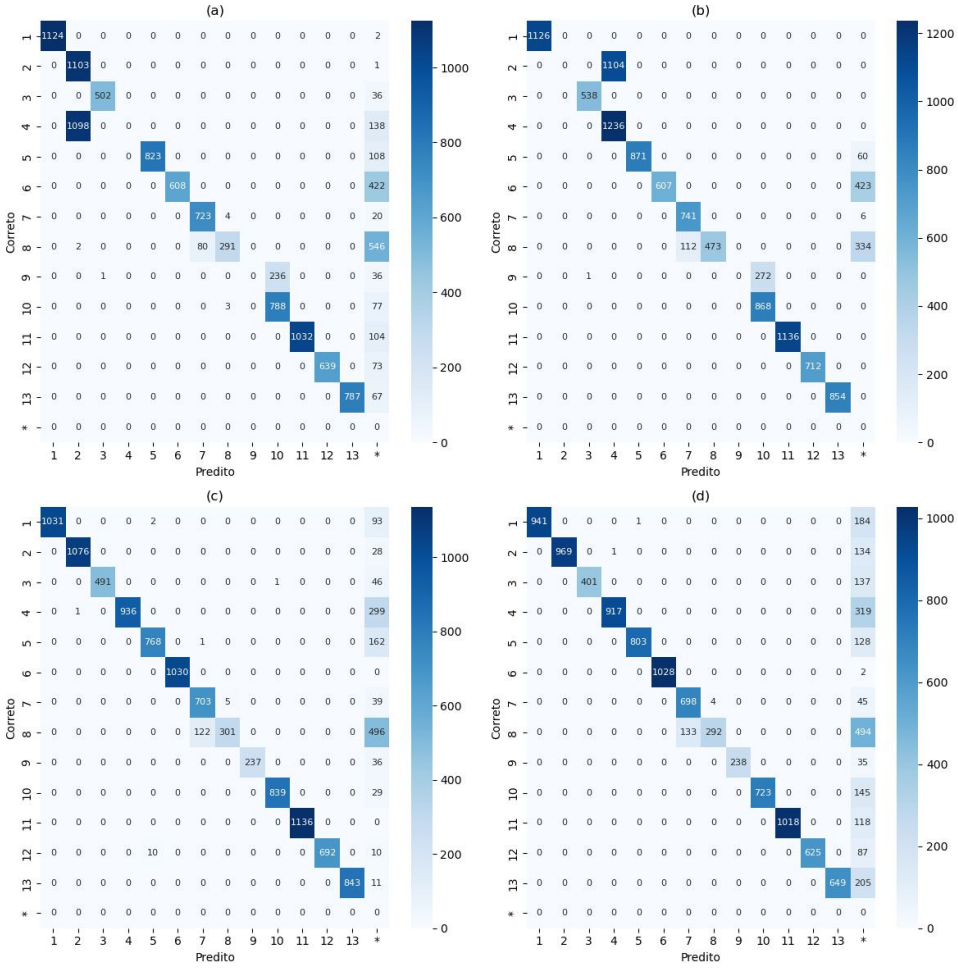

Figura 2. Matrizes de confusão: (a) Método1 (b) Método2 (c) Método3 (d) Método4

Ao analisar a matriz de confusão obtida para o Método 1, Figura 2 (a), percebe-se que os itens de "(10) leite em pó 400g" e "(9) leite condensado 395g" foram mantidos em um mesmo grupo, bem como os itens "(4) detergente 500ml" e "(2) água 500ml". Observa-se ainda que uma grande quantidade de itens dos produtos "(8) gasolina comum" e "(6) gás natural" foram separados em grupos que na matriz aparecem como grupos extras, representados por '*'.

Para o Método 2 (Figura 2 (b)), observa-se que a utilização da etapa de pré-processamento do texto auxiliou na redução de itens alocados a grupos extras, compa-

rativamente ao Método 1 (Figura 2 (a)). Note que a quantidade de itens na classe '*' diminuiu, originando também uma melhora nas métricas de avaliação (Tabela 2).

Em relação ao grupo formado pelos produtos "(2) água mineral 500ml" e "(4) detergente 500ml" a utilização da Método 2 não possibilitou a separação (Figura 2 (b)). Com a remoção da unidade de medida das descrições anteriormente ao agrupamento por tópicos estes itens foram mais satisfatoriamente separados. No terceiro método temos resultados superiores aos anteriores. Na Figura 2 (c) é visível que os dois produtos foram separados e formaram grupos próprios.

No Método 4, com a adição do filtro inicial pelo código NCM posição (NCMp), houve uma queda no desempenho. Isso ocorre pois há um preenchimento errado do código NCM nas notas fiscais, resultando no aumento do número de itens na classe '*' comparado com o Método 3. Cabe ressaltar no entanto que o conjunto de dados utilizado neste trabalho considerou apenas 13 produtos distintos, número bastante inferior ao número de produtos distintos praticados nas compras públicas, fato que pode ter tornado a separação por NCM pouco efetiva.

Tabela 2. Acurácia, precisão, revocação e F1 Score dos 4 métodos propostos

Proposta	Acurácia	Precisão	Revocação	F1 Score
Método 1	0.7295	0.7926	0.7295	0.7332
Método 2	0.7980	0.8034	0.7980	0.7803
Método 3	**0.8790**	0.9872	**0.8790**	**0.9165**
Método 4	0.8033	**0.9883**	0.8033	0.8740

4. Conclusão

Este trabalho propôs uma metodologia para o agrupamento de notas fiscais utilizando a técnica de modelagem de tópicos nas descrições dos produtos. Foram realizados experimentos nos quais aplicou-se o algoritmo de modelagem de tópicos em dados de descrições de produtos de notas fiscais usando 4 métodos distintos. O 1º método consiste em aplicar técnica de modelagem de tópicos na descrição original. No 2º método é feito pré-processamento das descrições antes da etapa de modelagem de tópicos. No 3º método é adicionado uma etapa de formação de sub-grupos por termos de unidade de medida extraídos das descrições. Por fim, no 4º método o conjunto de dados é dividido pelo código NCM antes de aplicar as três etapas definidas anteriormente.

Os resultados mostram uma melhora significativa no desempenho da modelagem de tópicos com a adição das etapas de pré-processamento de texto e formação de subgrupos por unidades de medida extraídas da descrição. A etapa de separação inicial por NCM pode ser útil para conjuntos de dados maiores, pois torna o problema menos complexo para o algoritmo de modelagem de tópicos.

Este é um trabalho em andamento e para etapas futuras pretende-se desenvolver um novo conjunto de dados mais abrangente para testes e explorar o uso de Modelos de Linguagem de Grande Escala (LLM's) para extração de características das descrições.

Agradecimentos

Este trabalho conta com recursos financeiros do projeto *Céos: Inteligência de Dados para a Sociedade*, uma parceria de pesquisa entre a UFSC e o Ministério Público do Estado de Santa Catarina (MPSC) com suporte financeiro do MPSC.

Referências

Angelov, D. (2020). Top2vec: Distributed representations of topics. `https://github.com/ddangelov/Top2Vec`.

Brasil (2021). Lei nº 14.133, de 1º de abril de 2021. `https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14133.htm`. Lei de Licitações e Contratos Administrativos.

Brinkmann, A., Baumann, N., and Bizer, C. (2024). Using llms for the extraction and normalization of product attribute values.

Kieckbusch, D. S. (2022). Scan-nf: a machine learning system for invoice product transaction classification through short-text processing. Master's thesis, Univerty of Brasília (UnB).

Krieger, F., Drews, P., and Funk, B. (2023). Automated invoice processing: Machine learning-based information extraction for long tail suppliers. *Intelligent Systems with Applications*, 20:200285.

Novaes, L. P., Vianna, D., and da Silva, A. (2023). Modelagem de tópicos para a tarefa de recuperação de casos legais. In *Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados*, pages 128–140, Porto Alegre, RS, Brasil. SBC.

Paalman, J., Mullick, S., Zervanou, K., and Zhang, Y. (2019). Term based semantic clusters for very short text classification. In Mitkov, R. and Angelova, G., editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 878–887, Varna, Bulgaria. INCOMA Ltd.

Silva, M. O., Costa, L. L., de Barros Bezerra, G. F., Gomide, L. D., Hott, H. R., Oliveira, G. P., Brandão, M. A., Lacerda, A., and Pappa, G. L. (2023). Análise de sobrepreço em itens de licitações públicas. *Anais do XI Workshop de Computação Aplicada em Governo Eletrônico (WCGE 2023)*.

Yao, X., Sun, H., Li, S., and Lu, W. (2022). Invoice detection and recognition system based on deep learning. *Security and Communication Networks*, 2022(1):8032726.