UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

MELQUISEDEQUE CARVALHO SILVA

BUSCA POR PRODUTOS: UM ESTUDO COMPARATIVO DE ABORDAGENS LÉXICAS E SEMÂNTICAS

CAMPINA GRANDE - PB

2024

MELQUISEDEQUE CARVALHO SILVA

BUSCA POR PRODUTOS: UM ESTUDO COMPARATIVO DE ABORDAGENS LÉXICAS E SEMÂNTICAS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Cláudio de Souza Baptista

CAMPINA GRANDE - PB

2024

MELQUISEDEQUE CARVALHO SILVA

BUSCA POR PRODUTOS: UM ESTUDO COMPARATIVO DE ABORDAGENS LÉXICAS E SEMÂNTICAS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Cláudio de Souza Baptista

Orientador – UASC/CEEI/UFCG

Franklin de Souza Ramalho

Examinador – UASC/CEEI/UFCG

Francisco Vilar Brasileiro

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 15 de maio de 2024.

CAMPINA GRANDE - PB

RESUMO

A busca por produtos é uma funcionalidade fundamental que permite aos usuários localizar e adquirir itens específicos, sendo aplicada em diversos contextos, como *e-commerces* e sites de comparação de preços. Este estudo compara abordagens léxicas e semânticas para a realização dessa funcionalidade. Embora a busca léxica possua vantagens em termos de tempo de resposta, ela não captura relações semânticas entre palavras além das similaridades léxicas. Por outro lado, a busca semântica destaca-se ao capturar semanticamente a relação entre os termos, porém, além de possuir maior complexidade, ela também pode ser mais lenta. Neste trabalho, analisamos as abordagens léxico-semânticas para dados de produtos, especificamente em dois conjuntos de dados: Catálogo de Materiais do Governo Federal e descrições de produtos presentes em notas fiscais. Comparamos as estratégias de busca considerando a relevância dos resultados e o tempo de resposta. Os conjuntos de dados possuem características distintas, com o catálogo de materiais sendo mais formal e estruturado, enquanto as notas fiscais contêm textos mais curtos e informais, frequentemente com siglas e abreviações. Este estudo comparativo busca identificar os *trade-offs* entre as abordagens léxicas e semânticas, bem como encontrar as estratégias mais adequadas para cada tipo de dado. Os resultados contribuem para a seleção de mecanismos de busca mais eficazes em catálogos de produtos, considerando diferentes formas de organização dos dados.

BUSCA POR PRODUTOS: UM ESTUDO COMPARATIVO DE ABORDAGENS LÉXICAS E SEMÂNTICAS

ABSTRACT

The search for products is a fundamental feature that allows users to locate and acquire specific items, and it is applied in various contexts such as e-commerce websites and price comparison sites. This study compares lexical and semantic approaches for implementing this functionality. While lexical search has advantages in terms of response time, it does not capture semantic relationships between words beyond lexical similarities. On the other hand, semantic search stands out by semantically capturing the relationship between terms, but it can be more complex and slower. In this work, we analyze lexical-semantic approaches for product data, specifically in two datasets: the Federal Government Materials Catalog and product descriptions found in invoices. We compare search strategies, considering the relevance of results and response time. The datasets have distinct characteristics, with the materials catalog being more formal and structured, while invoices contain shorter and more informal texts, often with acronyms and abbreviations. This comparative study aims to identify trade-offs between lexical and semantic approaches, as well as to find the most suitable strategies for each type of data. The results contribute to selecting more effective search mechanisms in product catalogs, considering different data organization formats.

Busca por produtos: um estudo comparativo de abordagens léxicas e semânticas

Melquisedeque Carvalho Silva
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
melquisedeque.silva@ccc.ufcg.edu.br

Cláudio de Souza Baptista
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
baptista@computacao.ufcg.edu.br

André Luiz F. Alves
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
andre.alves@ifpb.edu.br

ABSTRACT

The search for products is a fundamental feature that allows users to locate and acquire specific items, and it is applied in various contexts such as e-commerce websites and price comparison sites. This study compares lexical and semantic approaches for implementing this functionality. While lexical search has advantages in terms of response time, it does not capture semantic relationships between words beyond lexical similarities. On the other hand, semantic search stands out by semantically capturing the relationship between terms, but it can be more complex and slower. In this work, we analyze lexical-semantic approaches for product data, specifically in two datasets: the Federal Government Materials Catalog and product descriptions found in invoices. We compare search strategies, considering the relevance of results and response time. The datasets have distinct characteristics, with the materials catalog being more formal and structured, while invoices contain shorter and more informal texts, often with acronyms and abbreviations. This comparative study aims to identify trade-offs between lexical and semantic approaches, as well as to find the most suitable strategies for each type of data. The results contribute to selecting more effective search mechanisms in product catalogs, considering different data organization formats.

Keywords

busca semântica, busca léxica, busca de produtos, recuperação de informação

1. INTRODUÇÃO

A Recuperação de Informação (RI) é uma área essencial da Ciência da Computação que facilita o acesso à informação em um mundo cada vez mais digitalizado. Ela tem um papel fundamental ao lidar com a recuperação de documentos, sendo crucial para a construção de sistemas consistentes e indispensável na era da informação.

A RI também desempenha um papel crucial na busca por produtos. Em um mundo no qual há uma quantidade vasta de produtos, a capacidade de recuperar informações precisas se torna cada vez mais significante. A gestão eficaz de informações sobre produtos também é crucial para uma variedade de aplicações comerciais e organizacionais.

No contexto deste trabalho, aplicamos um estudo utilizando a base de dados do Catálogo de Materiais do Governo Federal (CATMAT) [1] e descrições contidas em notas fiscais.

O CATMAT desempenha um papel fundamental ao padronizar e organizar informações sobre produtos. Ele abrange uma ampla gama de itens, desde materiais utilizados em grandes construções civis até pequenos materiais utilizados em confecção de roupas. A base fornece descrições detalhadas, especificações técnicas e codificações padronizadas acerca dos produtos.

As notas fiscais representam uma fonte de dados crítica e abundante sobre transações comerciais. Elas registram detalhes essenciais das transações, incluindo itens adquiridos, preços, datas e identificadores de produtos. A importância das notas fiscais vai além do cumprimento de obrigações fiscais, podendo ser utilizadas na contabilidade, padronização dos relacionamentos entre empresas e aumento da segurança em vendas [2].

No entanto, é importante ressaltar que as características das bases de dados CATMAT e das notas fiscais são distintas. Enquanto o CATMAT possui uma estrutura organizada e formalidade nas descrições de produtos, as notas fiscais frequentemente apresentam textos curtos, informais e repletos de abreviações e siglas.

Para explorar as nuances e desafios associados à busca de produtos em diferentes bases de dados, este trabalho propõe uma análise comparativa entre abordagens léxico-semânticas, apresentando os *trade-offs* entre a relevância e o tempo de resposta de resultados obtidos em duas bases de dados distintas - Catálogo Nacional de Materiais (CATMAT) e notas fiscais. Utilizaremos a ferramenta ElasticSearch [3] para implementar diferentes estratégias de busca, com foco na avaliação da relevância dos resultados e no tempo de resposta. O objetivo é identificar e validar, a partir de métricas adequadas, quais abordagens são mais eficazes em cada contexto, considerando as particularidades das bases de dados utilizadas.

Dessa forma, esta pesquisa visa contribuir para uma melhor compreensão das complexidades envolvidas na busca de produtos em diferentes contextos de dados, fornecendo conhecimentos significativos para o aprimoramento de sistemas de gestão de informações e processos de tomada de decisão relacionados a compras e contratações.

2. TRABALHOS RELACIONADOS

A investigação das abordagens semânticas e léxicas são um tema central na pesquisa em processamento de linguagem natural, dada a coexistência desses diferentes enfoques na análise textual. Entender como essas abordagens se complementam ou se diferenciam é crucial na tomada de decisão do melhor método a

ser utilizado em cada contexto. Nesta seção, serão apresentados alguns trabalhos que exploram tanto abordagens semânticas quanto léxicas em diferentes cenários.

A inclusão de semântica para tratar de textos pequenos, como a descrição em notas fiscais abordadas neste trabalho, é um desafio significativo. O artigo de Paalman et al. [4] também trabalha com o contexto em textos muito curtos, como tweets e faturas, enfatizando a dispersão nessas descrições e discutindo sobre os problemas de conexão com a semântica nessas pequenas descrições.

Quando apenas observamos de forma léxica as possíveis descrições de produtos, podemos nos deparar com alguns dos problemas citados no trabalho de Nigam et al. [5], sendo um deles o fator da correspondência léxica poder falhar na recuperação de produtos que correspondam à intenção semântica da consulta. O artigo também aborda o contexto de busca em produtos, enfatizando os problemas dados a partir da falta de compreensão de hiperônimos, sinônimos e antônimos e como isso pode impedir a recuperação precisa de produtos semanticamente relacionados. Para a avaliação de resultados nesse trabalho, foram utilizadas as métricas Precisão, *recall* e NDCG.

Outro trabalho que utiliza as mesmas métricas citadas no trabalho de Nigam et al. é o artigo de Mangold [6]. Utilizando essas métricas, Mangold realizou uma comparação entre diferentes implementações de busca semântica. Inicialmente, no artigo, foram selecionadas algumas consultas e aplicadas substituições de palavras por termos ontologicamente relacionados, como sinônimos, hiperônimos e hipônimos, gerando assim diversos tipos de consulta para utilização na avaliação. Essa abordagem de modificação de consultas também se mostra interessante para estudos que buscam comparar diferentes abordagens de busca, permitindo verificar o comportamento das implementações em diferentes situações.

No artigo de Zhu et al. [7], é utilizada uma abordagem de classificação híbrida para textos curtos. São utilizados exemplos de textos com suas características léxicas, sendo também adicionadas incorporações de caracteres com os significados semânticos que eles possuem. Na pesquisa de Zhu et al., os resultados não foram avaliados utilizando as mesmas métricas dos trabalhos de Nigam et al. e Paalman et al. A abordagem híbrida pode ser uma alternativa para casos em que o conjunto de dados requer a utilização de ambos os enfoques de busca, sendo cada um deles mais impactante em diferentes contextos dentro da base de dados.

Diferentemente dos trabalhos relacionados, este trabalho visa apresentar comparações de desempenho não só no quesito relevância dos resultados, como também no tempo de resposta obtido nas consultas. Essas comparações serão feitas em duas bases diferentes, com características distintas, sendo apresentados resultados a partir das métricas Precisão, *recall* e NDCG, também utilizadas em alguns dos trabalhos relacionados.

3. METODOLOGIA

Nesta seção, detalhamos as abordagens e procedimentos utilizados para realizar a comparação entre as buscas léxicas e semânticas em diferentes conjuntos de dados.

Devido à diferente natureza dos dados de Nota Fiscal e do CATMAT, diferentes estratégias foram utilizadas em cada ponto da metodologia. Portanto, sempre que necessário, as seções e subseções a seguir serão divididas em dois tópicos, sendo um deles para a base de NOTA FISCAL e o outro para a base de catálogos de materiais.

As atividades desenvolvidas na metodologia se dividem nas etapas de Preparação da Base de Dados, Esquema, Execução e Análises realizadas no Experimento. Também são apresentados o ambiente do experimento e os métodos de busca utilizados na pesquisa.

3.1 Preparação da Base de Dados:

3.1.1 Catálogo de Materiais

A base de dados do catálogo nacional de produtos [8] inclui tanto materiais (CATMAT) quanto serviços (CATSER), sendo diferenciados por uma coluna denominada 'tipo'. Como o foco deste estudo é a busca realizada nos materiais, foi realizado um filtro para considerar apenas os materiais do conjunto de dados utilizado. Além disso, o catálogo apresenta diversas colunas contendo descrições diferentes para cada PDM (Padronização Descritiva de Materiais), divididas em classes, subclasses e grupos, sendo cada uma delas uma tabela, como pode ser visto na Figura 1. Cada PDM possui materiais de categorias semelhantes, sendo dado um nome a essa categoria de materiais na coluna “nome_material”. São esses materiais pertencentes a um PDM que serão o objeto de busca neste trabalho. Para facilitar a indexação dos dados no Elasticsearch, foi criada uma coluna que concatena os valores das colunas das diferentes descrições (caso não nulos), denominada "material_text".

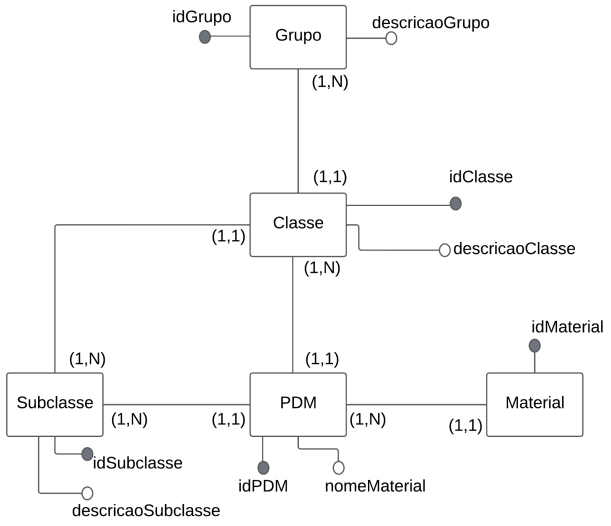

Figura 1. Relacionamento entre as entidades da Base de dados do CATMAT

3.1.2 Nota Fiscal

Na base de dados de notas fiscais, dois campos são cruciais para a realização deste trabalho, sendo eles “codigo_de_barra” e “descricao_produto”. Foram observadas múltiplas entradas com diferentes códigos de barras para a mesma descrição de produto e vice-versa. Para uniformizar os dados, foram removidas ocorrências de uma mesma descrição para diferentes códigos de barras. Como critério de exclusão foi mantido o codigo_de_barra com maior ocorrência para cada descrição. É possível ver esse

tratamento nos dados com um exemplo partindo da Tabela 1, que possui diferentes códigos de barras para a descrição de valor “"BF DISCO FLAP AZUL RETO PLAST”, chegando ao resultado na Tabela 2, após os tratamentos.

Tabela 1: Valores de “codigo_barra” para uma mesma descrição

ID_ITEM	DESCRICAO_PRODUTO	CODIGO_BARRA
63307456	"BF DISCO FLAP AZUL RETO PLAST	7899206184255
63307471	"BF DISCO FLAP AZUL RETO PLAST	7899206184293
63396373	"BF DISCO FLAP AZUL RETO PLAST	7899206184255
63396386	"BF DISCO FLAP AZUL RETO PLAST	7899206184293
70141645	"BF DISCO FLAP AZUL RETO PLAST	7899206184286
70141659	"BF DISCO FLAP AZUL RETO PLAST	7899206184262
70141663	"BF DISCO FLAP AZUL RETO PLAST	7899206184279
70141673	"BF DISCO FLAP AZUL RETO PLAST	7899206184293

Tabela 2: Valor de um “codigo_barra” para uma mesma descrição, após tratamento dos dados.

ID_ITEM	DESCRICAO_PRODUTO	CODIGO_BARRA
63307471	"BF DISCO FLAP AZUL RETO PLAST	7899206184293
63396386	"BF DISCO FLAP AZUL RETO PLAST	7899206184293
70141673	"BF DISCO FLAP AZUL RETO PLAST	7899206184293

3.2 Ambiente do experimento

Para realização dos experimentos foi utilizada uma máquina com processador Intel® Core™ i7-7700K CPU @ 4.20GHz × 8, Memória Ram DDR4 de 32,0 GiB, Sistema Operacional Linux, na distribuição Ubuntu 22.04.3 LTS. Como linguagem para o desenvolvimento, foi utilizado o Python 3.10.0 [9]. Para a indexação dos dados foi utilizado o Elasticsearch na versão 8.12 [10] como ambiente de experimentação. Também foi utilizada a biblioteca Elasticsearch do Python [11], que provê um *client* do Elasticsearch, possibilitando uma maior facilidade e controle ao utilizar das operações de busca e indexação, realizadas no Elasticsearch, junto às funcionalidades da linguagem.

3.3 Métodos de busca do experimento

Para as buscas léxicas, empregou-se o algoritmo de classificação padrão dessa versão do Elasticsearch, o BM25 [12]. Já para a busca semântica, foi utilizado o all-MiniLM-L6-v2, um modelo SBERT(Sentence-BERT) utilizado para a vetorização de termos e otimizado para geração de *embeddings* de sentenças [13]. Esse modelo é treinado com uma tarefa específica de aprendizado de representação para maximizar a semelhança entre pares de sentenças semanticamente equivalentes [14].

3.4 Esquema do Experimento

Nesta seção serão apresentadas as decisões de design tomadas para os experimentos de tempo de execução e relevância.

3.4.1 Tempo de execução

Para os experimentos, consultas foram executadas para as duas abordagens - léxica e semântica - utilizando grupos de testes com quantidade de termos variados, a fim de avaliar o tempo de resposta para diferentes correspondências de diferentes tamanhos, considerando que, consultas com maior quantidade de termos podem demorar mais devido a maior quantidade de palavras a serem verificadas e combinadas. A quantidade de termos utilizada se deu a partir dos n-gramas de cada descrição realizada nas consultas. Ou seja, para cada descrição, variando sobre os termos da mesma, houve uma quantidade diferente de consultas. Como exemplo, podemos observar o produto “PAPEL TOALHA MILI BIANCO 110FLS”, no qual se resultam as seguintes consultas:

- PAPEL
- PAPEL TOALHA
- PAPEL TOALHA MILI
- PAPEL TOALHA MILI BIANCO
- PAPEL TOALHA MILI BIANCO 110FLS

3.4.2 Relevância

Nesta etapa, foram gerados diferentes grupos de testes, sendo eles diferenciados por ruídos de diferentes características. Um grupo inicial foi criado contendo produtos escolhidos de forma aleatória na base de dados. Posteriormente, para cada *query* desse grupo escolhido, foram inseridos diferentes tipos de ruídos, com o auxílio de modelos de Large Language Model. Quatro grupos foram criados para o experimento, sendo o primeiro deles formado por descrições exatas de cada produto, enquanto os outros três foram formados por dados alterados, tendo cada grupo apenas um tipo de ruído em seus dados.

Os testes aplicados envolveram três diferentes formas de alterações nas descrições de produtos, sendo eles:

- Reordenamento dos termos de uma consulta;
- Troca dos termos de uma consulta por sinônimos, representando o mesmo produto com diferentes palavras;
- Erros de digitação em termos da consulta.

Como exemplo, podemos observar a descrição “LAVADORA 10 COLORMAQ” da base de nota fiscal. Após atribuídos cada um dos ruídos, os grupos resultantes foram os seguintes:

- Grupo 1: “LAVADORA 10 COLORMAQ” - contendo a descrição exata do produto;

- Grupo 2: “LAVADERA 10 COLORMAQ” - contendo o erro de digitação em uma consulta;
- Grupo 3 : “MÁQUINA DE LAVAR 10 COLORMAQ” - contendo a descrição utilizando sinônimos;
- Grupo 4: “COLORMAQ 10 LAVADORA” - contendo a troca de termos em uma consulta.

Tendo em vista que buscas nem sempre são realizadas por *queries* com descrições exatas de um produto, o intuito dessa abordagem é validar como os diferentes métodos de busca se comportam em situações reais de consulta, contendo diferentes representações de um determinado produto, como apresentado nos grupos acima.

Com a finalidade de apresentar apenas produtos semelhantes e não os produtos estritamente iguais aos utilizados nas consultas, os produtos de notas fiscais e materiais que foram selecionados aleatoriamente para o Grupo 1 de testes foram retirados dos documentos indexados no Elasticsearch. Na base de dados do CATMAT, o campo "material_text" foi utilizado para a indexação e busca, enquanto para as notas fiscais, o campo "descricao_do_produto" foi utilizado com as mesmas finalidades.

3.5 Execução do Experimento

Neste tópico, serão apresentados os procedimentos realizados na etapa de experimentação deste trabalho. Para estruturação dos passos realizados em cada experimento, foram utilizados como referência os fluxogramas apresentados nas figuras 4 e 6.

3.5.1 Tempo de execução

Para o experimento, a execução das consultas foi realizada de forma intercalada entre cada uma das abordagens, com o intuito de ser o mais justo possível e minimizar ao máximo que fatores externos à realização dos testes, como o cache do sistema, apresentassem grande impacto nos resultados.

3.5.1.1 Nota Fiscal

Como explicado na seção de esquema do experimento, grupos de descrições com quantidade de termos variados foram criados. No caso da base de nota fiscal, foi primeiramente realizada uma divisão dos produtos conforme a frequência dos códigos de barras na base de dados. As seguintes divisões foram realizadas:

- Produtos com códigos de barra com até 5 aparições na base de dados - Grupo de códigos de barra menos frequentes:
- Produtos com códigos de barra com pelo menos 25 aparições na base de dados - Grupo de códigos de barra menos frequentes;
- Grupos com códigos de barras entre 5 e 15 códigos de barras;
- Grupos com códigos de barras entre 15 e 25 códigos de barras.

Após realizada essa etapa, 40 produtos foram selecionados, sendo 10 produtos de cada um desses grupos, cada produto contendo exatamente 5 palavras. Essa escolha se deu visando igualar as consultas em quantidade de termos, além de realizar a etapa n-gramas dos testes, utilizando um limite comum de caracteres. Sendo assim, para cada um dos 4 grupos, 10 palavras foram escolhidas e executadas 5 vezes, resultando em 200 consultas. Cada consulta foi repetida 10 vezes para cada abordagem, léxica e semântica, resultando assim em 4000 consultas totais. Através dessa sequência de testes, foram aplicadas a média e desvio padrão a partir dos dados obtidos.

3.5.1.2 CATMAT

Para a base do CATMAT, onde os dados são bem mais descritivos e extensos que na base de nota fiscal, como é possível de ver nas figuras 2 e 3, e onde também não há um código de barra para determinar um produto específico dentro de todos os outros na base, o experimento foi bem mais simples. Foram selecionados aleatoriamente 40 materiais da base de dados. O mesmo processo de n-gramas foi utilizado para essa base, porém como as descrições dos materiais são mais extensas que as das notas fiscais, cada produto selecionado foi consultado até no máximo a quinta palavra, mesmo que o produto possuísse uma descrição maior que isso. Portanto, foram realizadas 5 consultas para cada uma das 40 descrições. Assim como na base fiscal, os experimentos foram repetidos 10 vezes para cada abordagem, resultando assim em 4000 consultas realizadas.

Figura 2. Histograma da quantidade de palavras em descrições - Notas Fiscais

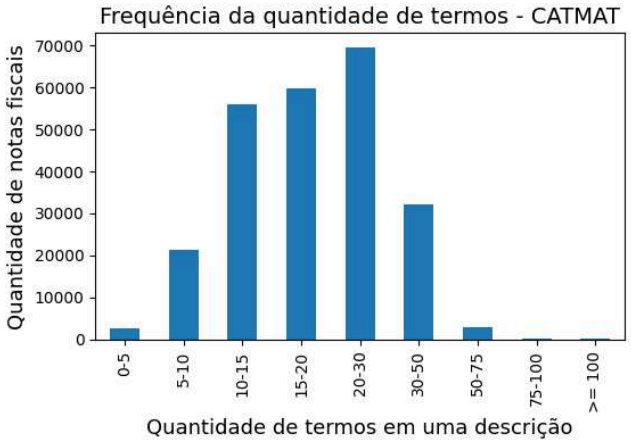

Figura 3. Histograma da quantidade de palavras em descrições - CATMAT

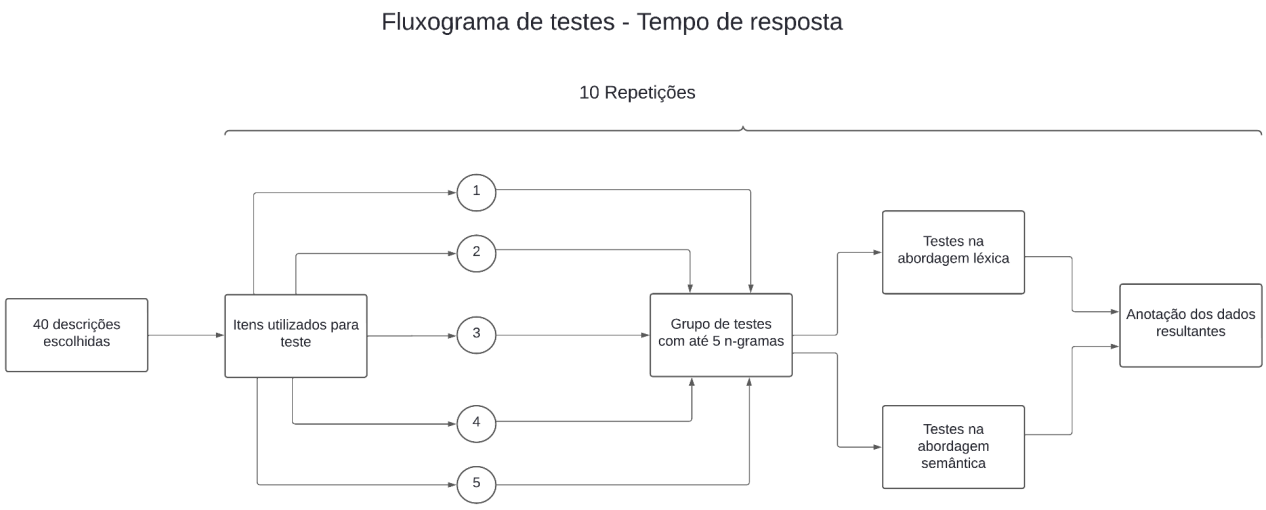

Figura 4. Fluxograma do experimento realizado para avaliação de tempo de resposta

3.5.2 Relevância

Após a seleção dos dados de teste, as consultas foram executadas de forma intercalada entre as duas abordagens, além de serem realizadas com base nos quatro grupos de testes gerados, contendo diferentes ruídos nas descrições. Foram obtidos dados cruciais para a etapa de análise, contendo resultados de relevâncias para diferentes situações. Para atribuir uma pontuação de relevância, também foi utilizada a métrica NDCG [15].

O impacto dessa métrica na pontuação de relevância será melhor abordado nas subseções a seguir.

3.5.2.1 Nota Fiscal

Como critério para relevância, foi considerado o campo "codigo_de_barra" dos produtos. Para um determinado produto pesquisado, todos os produtos que possuem o mesmo valor para o campo "codigo_de_barra" foram tidos como relevantes, tendo em conta que, por compartilhar o mesmo código de barras se tratam do mesmo produto.

Esse ponto pode ser melhor observado a partir do exemplo do "codigo_de_barra" com valor "7891040113491". Para esse caso, podemos observar três descrições diferentes, sendo elas: "Esponja Mágica Scotch-Brite Esponja Mágica Scotch-Brite", "Esponja Mágica Scotchbrite h0002325936" e "Esponja Scotch-Brite Mágica". Todas essas descrições representam o mesmo produto.

Entretanto, alguns produtos, mesmo que com diferentes códigos de barras, possuem semelhantes descrições. Para validar esses casos, foi aplicado uma técnica de similaridade, apresentada no Algoritmo 1. Nesse algoritmo, são medidas as distâncias do cosseno entre duas descrições, sendo uma delas a descrição esperada e outra a descrição retornada. Dois valores de distância do cosseno são medidos, sendo um a distância dos vetores de *embeddings*, obtido através do *sentence-transformers* all-MiniLM-L6-v2, e o outro a distância entre vetores de *Term Frequency – Inverse Document Frequency* (TF-IDF). A maior pontuação entre as distâncias calculadas foi utilizada, considerando que para as abordagens léxicas, a medida do TF-IDF seria a ideal, enquanto para a abordagem semântica, a distância entre os *embeddings* seria uma medida mais adequada.

Algoritmo 1 - Técnica de similaridade utilizada para pontuação de relevância

```
verificacao_relevancia(item1, item2):
    if(item1 = item2) then
       return 2;

    v_tf1 = vetorizacao_com_tfidf(item1)
    v_tf2 = vetorizacao_com_tfidf(item2)

    v_emb1 = vetorizacao_com_embeddings(item1)
    v_emb2 = vetorizacao_com_embeddings(item2)

    distancia_tf_idf = similaridade_cosseno(v_tf1,
v_tf2)
    distancia_embed = similaridade_cosseno(v_emb1,
v_emb2)

    if(max(distancia_embed,distancia_tf_idf) > 0.7 then
       return 1;
    else
       return 0;
```

Para a pontuação de relevância foram seguidos os critérios contidos na Tabela 3.

Tabela 3: Critérios de pontuação para definição de relevância - Notas Fiscais.

Pontuação	Critério
2	Produtos que possuem mesmo valor no campo "codigo_de_barra", porém, com diferentes descrições
1	Produtos que possuem semelhança entre as descrições, porém, não possuem o mesmo valor no "codigo_de_barra", através do algoritmo indicado na Algoritmo 1. Pontuações maiores do que 0.7 foram consideradas relevantes;
0	Produtos que não possuem mesmo valor para o "codigo_de_barra", assim como, também possuem pontuação menor do que 0.7, conforme o Algoritmo 1

Na seleção dos grupos de dados para realização de testes, foram selecionados aleatoriamente 50 códigos de barras que possuíam entre 5 e 50 aparições na base de dados. Um limite inferior foi dado para que produtos pouquíssimos frequentes não fossem escolhidos. Dos códigos de barras selecionados, os mais recorrentes apresentaram 47 aparições na base. A distribuição dessas repetições pode ser vista na Figura 5. Uma paginação de 50 itens para as consultas no Elasticsearch foi escolhida como aproximação desse valor, para conter todos os relevantes de um item em caso de consultas bem sucedidas.

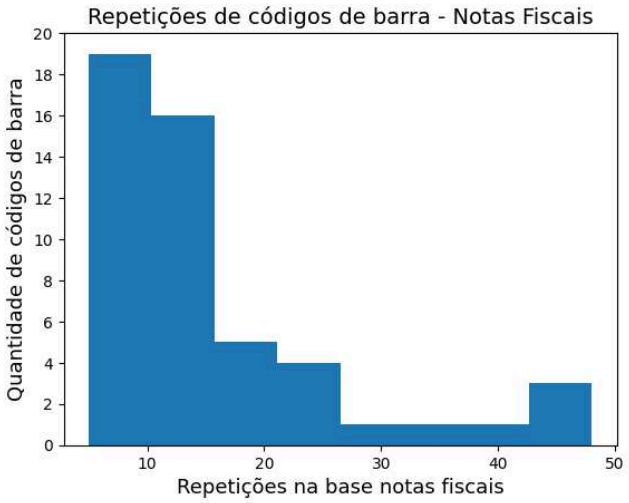

Figura 5. Histograma de repetições dos códigos de barra escolhidos aleatoriamente - Notas Fiscais

Cada um dos 50 produtos passou pela etapa de adição de quatro tipos de ruídos, como descrito anteriormente. As 200 consultas resultantes dessa etapa foram realizadas 10 vezes, sempre variando o tamanho de página para avaliar a relevância dos resultados em diferentes tamanhos de paginação. Para a análise de resultados foram obtidos 4000 resultados de consultas, sendo 2000 para cada uma das abordagens.

3.5.2.2 CATMAT

Nessa base, os dados possuem classificações para grupos de produtos, sendo divididos por Classe, Subclasse e Grupo, como é possível ver na Figura 1. Porém, diferente da base de notas fiscais, os materiais não possuem um código de barras que difere um item específico em relação aos demais itens, eles apenas estão ligados a uma entidade PDM, sendo possível a partir dela, descobrir a quais classes e subclasses os materiais estão ligados.

Para ser possível selecionar produtos diferentes para os testes, assim como na base de notas fiscais, foi utilizada a divisão mais específica possível da base CATMAT, já que essas divisões diferenciam grupos de produtos. Segundo a Figura 1, a subclasse e a classe são as duas divisões mais específicas ligadas ao produto. Porém, poucos dados na base possuíam um valor cadastrado para a subclasse (Figura 3). Devido a isso, a divisão Classe foi utilizada, sendo selecionados produtos com diferentes valores para esse campo.

Como a intenção deste trabalho é comparar o resultado de abordagens em duas diferentes bases de dados, o mesmo número de paginação selecionado em notas fiscais foi utilizado nesta base. Nesse caso, como opção para escolha de casos de teste, foram escolhidos 50 diferentes produtos de 50 diferentes classes.

O critério para escolha de materiais se deu de forma semelhante a da base de notas fiscais. Sendo o PDM o agrupamento de materiais semelhantes, o campo "nome_material" foi utilizado para validar itens relevantes. Os materiais pertencentes ao mesmo PDM, ou seja, com o mesmo valor para "nome_material", foram considerados relevantes. Para a pontuação de relevância foram seguidos os critérios contidos na Tabela 4.

Tabela 4: Critérios de pontuação para definição de relevância - CATMAT.

Pontuação	Critério
2	Materiais com o mesmo valor no campo "nome_material".
1	Materiais que não possuem o mesmo valor para o campo "nome_material", porém possuem semelhança nos valores pertencentes ao campo "material_text", segundo o Algoritmo 1. Pontuações maiores do que 0.7 foram consideradas relevantes.
0	Materiais que não possuem mesmo valor para o "nome_material", assim como, também possuem pontuação menor do que 0.7, conforme o Algoritmo 1.

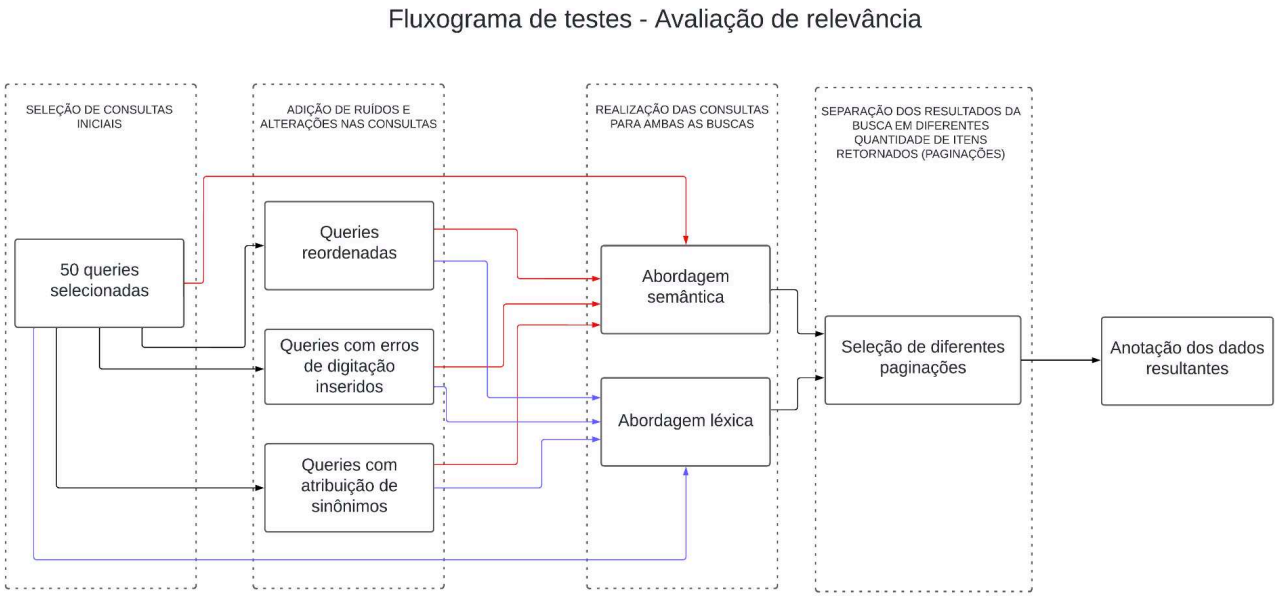

Figura 6. Fluxograma do experimento realizado para avaliação de relevância

3.6 Análise dos Experimentos

Nesta última etapa, foram analisados os valores de tempo de resposta obtidos após toda a série de testes realizados. Foram comparadas as médias e o desvio padrão dos valores obtidos em cada abordagem de busca.

Quanto à relevância, foram guardados os resultados das consultas realizadas utilizando os quatro grupos de teste citados durante o esquema do experimento. A partir desses resultados, foram aplicadas as métricas Precisão, *recall*, *F1-score* e NDCG, para avaliar a relevância dos resultados retornados por cada abordagem em cada contexto presente nos grupos de teste.

4. RESULTADOS E DISCUSSÕES

Nesta seção serão apresentados os resultados obtidos com os experimentos realizados.

4.1 Tempo de resposta

Considerando as repetições dos experimentos, foram calculadas as médias e desvio padrão dos tempos de resposta para as abordagens léxica e semântica. Na Figura 7 é possível ver uma síntese dos resultados dos tempos de resposta, obtidos considerando as bases de dados CATMAT e Notas Fiscais. Observa-se que, no geral, o tempo de resposta obtido na abordagem léxica foi consideravelmente menor que na abordagem semântica. Além disso, também é possível notar que há uma maior estabilidade no método léxico, apresentando uma inconstância bem menor.

Na Figura 8, são apresentadas as médias dos tempos de resposta para cada quantidade de termos, juntamente com seus respectivos valores de desvio padrão. Nessa figura conseguimos observar uma maior oscilação de tempo na abordagem semântica, que apresenta maiores diferenças de desvio padrão para as diferentes quantidades de termos. Já no método léxico, percebemos que há uma maior constância de tempo de resposta, não sendo tão afetada com a mudança de quantidade de termos. Mesmo sofrendo com valores discrepantes, a abordagem semântica do CATMAT sofreu menos com *outliers*, apresentando também uma variação mais uniforme para os tempos de resposta nas diferentes quantidades de termos. Em geral, a média obtida em cada quantidade de termos também variou mais na abordagem semântica.

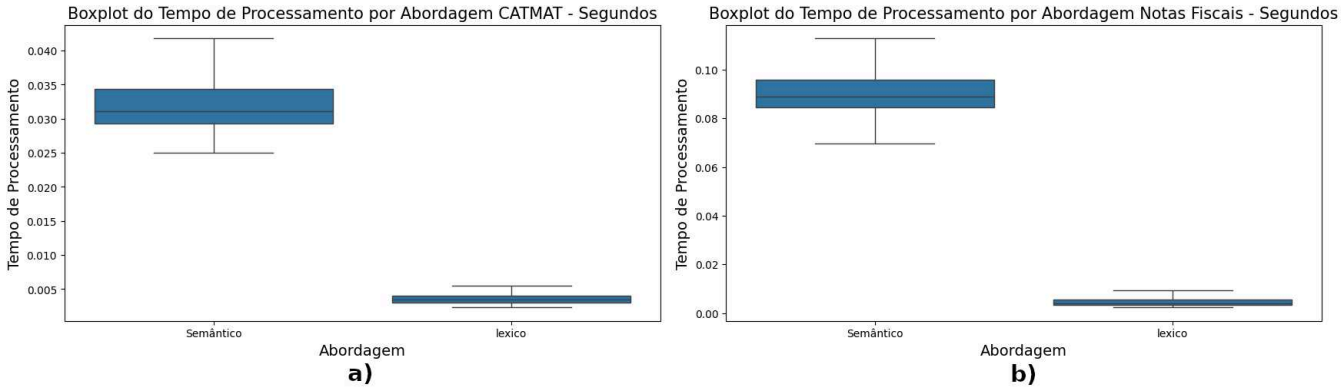

Figura 7. Boxplot do tempo de resposta obtido: a) CATMAT e b) Notas Fiscais

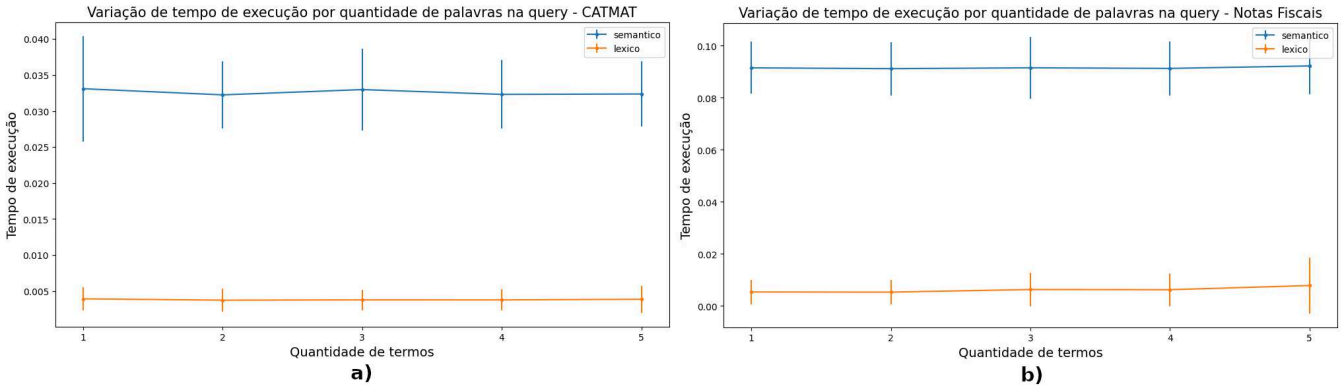

Figura 8. Tempo de resposta para cada quantidade de termos: a) CATMAT e b) Notas Fiscais

Na Figura 9 são apresentados os tempos de resposta para níveis de frequência de produtos na base de Notas Fiscais, variando de códigos de barras com menos de 5 aparições até códigos de barra com mais de 25 aparições. A variação foi pouco sentida pela abordagem léxica em todos os casos, sendo mais uniforme, enquanto na abordagem semântica não houve essa regularidade.

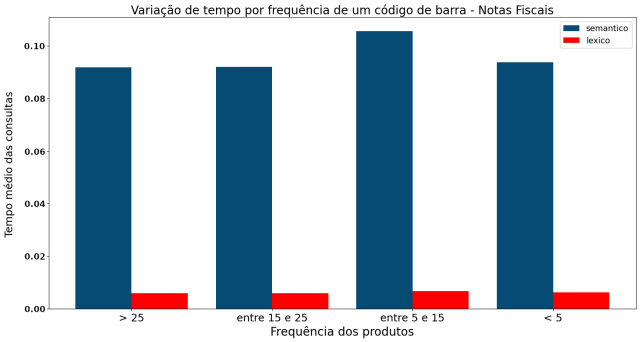

Figura 9. Variação do tempo de resposta por frequências de códigos de barra na base de Notas Fiscais

4.2 Relevância dos resultados

Nesta subseção serão apresentados os resultados e discussões dos experimentos realizados para verificação de relevância obtida em cada método utilizado. As métricas utilizadas para a avaliação dos resultados recuperados pelas abordagens de busca foram o *recall,* a Precisão*,* o *F1-score* e o NDCG.

As figuras 10 e 11 apresentam gráficos contendo os valores obtidos para as quatro métricas em cada base. Cada gráfico apresenta os resultados alcançados em cada métrica, utilizando os quatro tipos de consultas. No eixo x desses gráficos encontramos diferentes tamanhos de páginas, que chamamos de Q, enquanto no eixo y são apresentados os valores de cada métrica.

Na Figura 10.a, observamos valores próximos entre as abordagens nos quatro tipos diferentes de consultas realizadas. As pontuações alcançadas nas *queries* reordenadas foram similares das pontuações obtidas nas *queries* sem qualquer ruído, sendo o tipo de alteração que as abordagens de busca sofreram menor impacto. Enquanto isso, o grupo de consultas com utilização de sinônimos foi o que alcançou menor Precisão. No geral, apesar de obter valores parecidos, o léxico levou vantagem na Precisão em todos os tipos de *queries*. Na maioria dos casos, como nas consultas com erros de digitação, a abordagem léxica manteve quase sempre melhores resultados, com uma baixa margem de diferença para o método semântico.

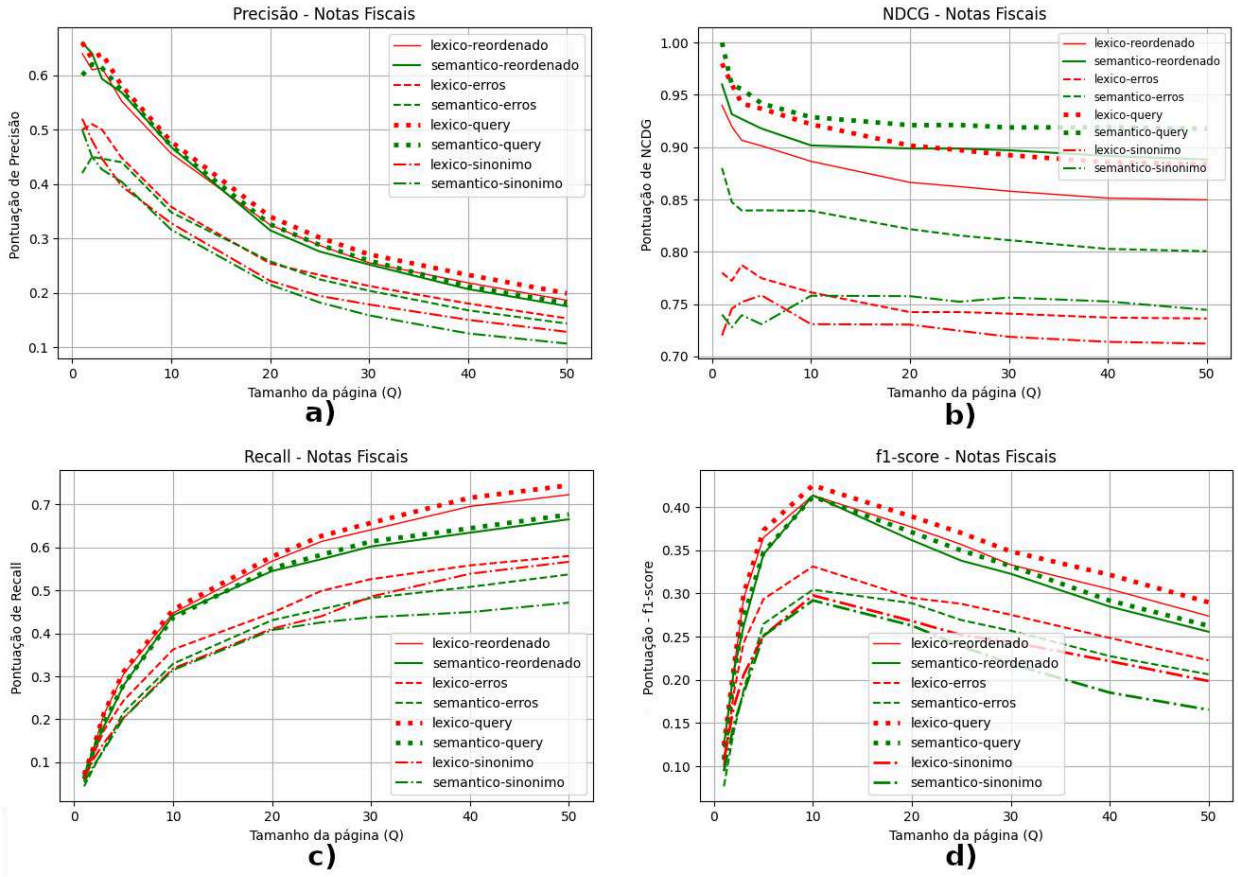

Figura 10. Métricas contendo valores para cada tipo de query - Notas Fiscais: a) Precisão; b) NDCG; c) Recall; d) F1-score

Já na Figura 11.a, percebemos que as diferenças não são tão próximas quanto na Figura 10.a, sendo a diferença de Precisão no CATMAT mais perceptível. Em todos os tipos de *queries* a abordagem léxica obteve melhores resultados de relevância. Ambas as abordagens apresentaram pouca diferença entre as consultas reordenadas e as consultas sem alterações, sendo possível observar no gráfico como os valores de "lexico-query" e "semantico-query" coincidiram com os valores de "lexico-reordenado" e "semantico-reordenado", respectivamente, para os primeiros tamanhos de página.

Para o NDCG, os valores resultantes contaram com o cálculo de similaridade citado durante a seção de Metodologia. Para a base de notas fiscais, os valores obtidos para a abordagem semântica foram maiores que os valores da abordagem léxica, diferindo da situação das outras métricas, como podemos ver na Figura 10.b. Um dos motivos possíveis para essa melhora em relação à Precisão, são os produtos com descrições semelhantes, mas que não possuem o mesmo código de barra, serem recuperados na busca. Com isso, a pontuação NDCG aumenta por levar em conta essa possível similaridade, enquanto as métricas de Precisão e *recall*, que ignoram essa possibilidade, mantém uma pontuação mais baixa, além de apresentar a abordagem léxica com valores maiores. Como exemplo para esse caso, podemos ver itens da Tabela 5, que possuem descrições com pouquíssimas diferenças, mudando cor ou tamanho, mas suficientes para representar produtos distintos com códigos de barras diferentes.

Tabela 5: Itens com descrições semelhantes e diferentes códigos de barras.

CODIGO_BARRA	DESCRICAO_PRODUTO
7893013025664	Capa Maquina de Lavar Com Ziper Todos Tamanhos e Marcas Electrolux Brastemp Consul - PRETA - TAMANHO G.
7893013025626	Capa Maquina de Lavar Com Ziper Todos Tamanhos e Marcas Electrolux Brastemp Consul - PRETA - TAMANHO M
7893013019113	Capa Maquina de Lavar Com Ziper Todos Tamanhos e Marcas Electrolux Brastemp Consul - BEGE - TAMANHO M

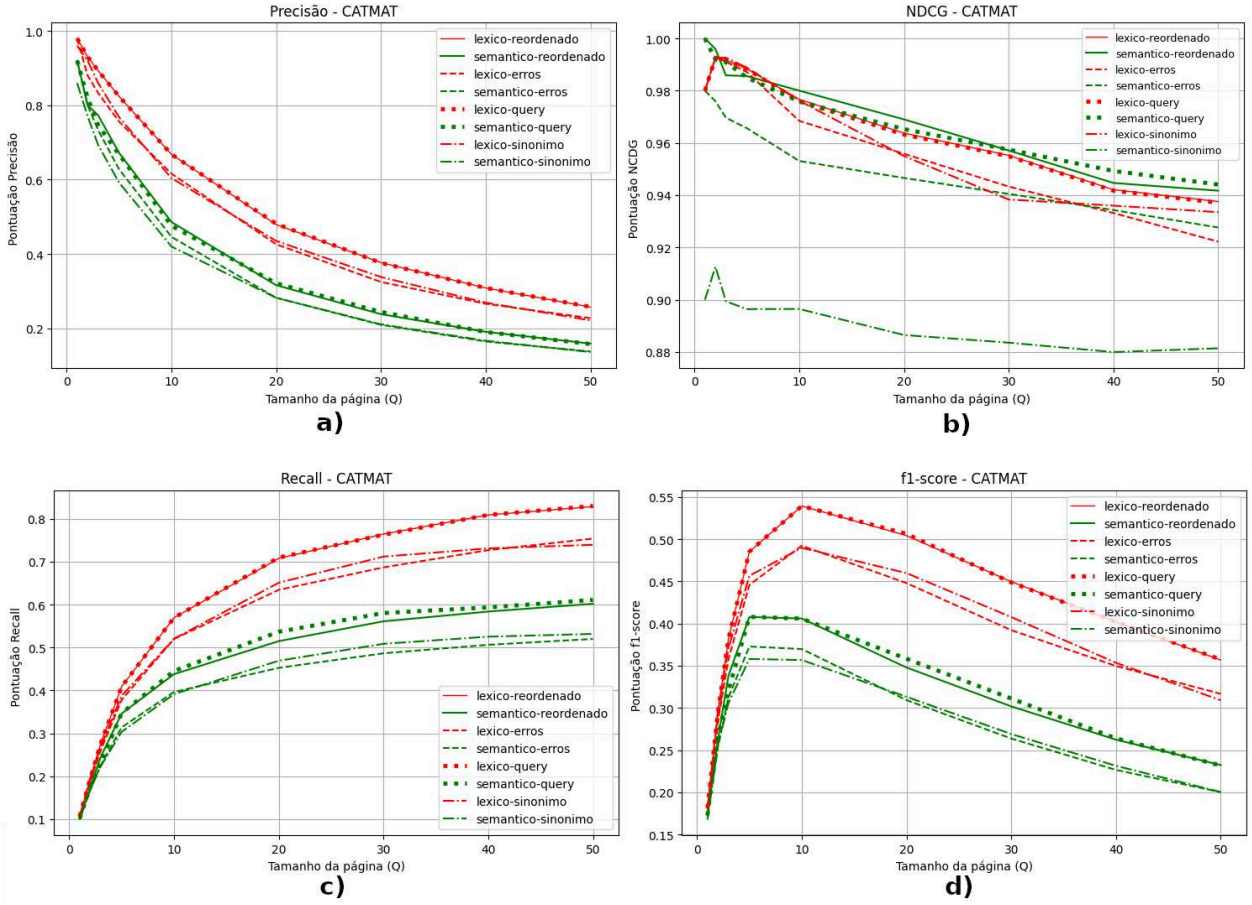

Figura 11. Métricas contendo valores para cada tipo de query - Notas Fiscais: a) Precisão; b) NDCG; c) Recall; d) F1-score

Por sua vez, na Figura 11.b podemos observar a métrica NDCG na base CATMAT, que também atingiu maiores valores que as métricas de Precisão e *recall*. A abordagem semântica obteve pequena vantagem nos resultados para as consultas sem alteração e as reordenadas. Já nas *queries* com erros de digitação a busca léxica obteve melhores resultados para valores de Q=0 até Q=40, enquanto o método semântico teve melhores resultados para Q=40 até Q=50. Apesar dos melhores resultados, no geral, estarem no método semântico para essa métrica, o valor mais baixo demonstrado na Figura 11.b foi a *query* com sinônimos dessa implementação. Um possível motivo para esse valor baixo, é o fato da busca semântica ter encontrado produtos relevantes, porém que não possuíam o exato mesmo nome do material, obtendo assim a pontuação 1, segundo a Tabela 4.

Em contrapartida, as demais consultas obtiveram mais produtos com pontuação 2. No caso das consultas com sinônimos na abordagem léxica, a maior influência dos termos não alterados nos documentos pode ter aumentado seus valores, pois muitos dos termos sinônimos inseridos podem não existir na base, tendo pouco impacto nas consultas realizadas.

Especificamente em relação à base CATMAT, uma das razões para o comportamento das métricas de *recall* e Precisão está na análise da base, que agrupa os materiais e os considera relevantes com base na coluna "nome_material", mantida no PDM. Ao contrário da base de notas fiscais, não foi utilizada uma coluna descritiva para diferenciar individualmente os produtos. Isso pode ter causado problemas nos possíveis mapeamentos da busca semântica, devido à generalização de alguns grupos de materiais. Por exemplo, o nome "Bebidas alcoólicas" pode incluir vários tipos diferentes de bebidas, não necessariamente agrupando apenas produtos semelhantes.

Outro problema encontrado está relacionado à presença de nomes de materiais que abrangem conceitos semelhantes, mas são tratados como entidades separadas. Por exemplo, enquanto o nome "Bebidas alcoólicas" é considerado como um grupo distinto, o termo "vinho" tem seu próprio “nome_material”. Como resultado, qualquer item relacionado a vinho que pertença ao grupo de bebidas alcoólicas não recebe a pontuação máxima de relevância. Um exemplo disso na base CATMAT é a busca por "lençol de algodão", que retorna mais de um nome de material, como “‘Lençol Cama’", "Jogo Cama" e "Lençol Cama". Como a busca semântica pode variar ainda mais os termos com base nos sinônimos utilizados, mais produtos que se assemelham, mas não pertencem ao mesmo “nome_material”, podem ser retornados. Além disso, problemas com a base, como um nome de material cadastrado como "Lençol Cama" e outro nome cadastrado como “‘Lençol Cama’”, apenas adicionando as aspas, é um exemplo de como redundâncias da base CATMAT pode influenciar negativamente na busca.

Somado a isso, alguns nomes de materiais agrupados na base utilizam de descrições que possuem termos amplos, como é o caso dos nomes de material "gás refrigerante", "vidro cristal" e "cerâmica odontológica". Sendo esses três exemplos facilmente confundíveis com outros tipos de produto, por possuir palavras como refrigerante e cerâmica, as buscas podem acabar retornando coisas diversas.

Para os valores de *recall*, observados nas figuras 10.c e 11.c, a abordagem léxica apresentou melhores valores em todos os tipos de consultas, assim como nos valores de Precisão, com a diferença de que em nenhum momento o semântico obteve uma pontuação maior.

Decorrente dos valores de *recall* e Precisão, o *F1-score* também apresentou valores maiores de pontuação para a abordagem léxica, como demonstrado nas figuras 10.d e 11.d, sendo ainda mais destoantes as diferenças entre as abordagens na base CATMAT.

5. CONCLUSÃO

Neste estudo foram avaliadas duas abordagens de busca, semântica e léxica, em dois diferentes conjuntos de dados relacionados a produtos, sendo um deles uma base de dados com descrições de notas fiscais e o outro um catálogo de materiais. As abordagens foram avaliadas considerando os critérios de tempo de resposta e relevância dos produtos recuperados por cada abordagem.

Nos resultados obtidos, após as medições realizadas, foi possível observar um menor tempo de resposta e um menor desvio padrão na abordagem léxica. Essa abordagem apresentou resultados mais rápidos para todas as medições realizadas, tanto na base de CATMAT quanto na base de Notas Fiscais. Quanto à relevância, as métricas obtidas apresentaram, no geral, melhores resultados de Precisão e *recall* para a abordagem léxica. Porém, enquanto a diferença de pontuação para essas duas métricas foi destoante na base CATMAT, a base de Notas Fiscais demonstrou proximidade nos valores, apresentando uma abordagem semântica com valores mais próximos dos obtidos no método léxico.

Existe um grande espaço para experimentação e análise, a partir dos dados obtidos nesse estudo. Como, por exemplo, novas comparações podem ser feitas com algumas alterações nas implementações léxicas e semânticas. A abordagem léxica implementada neste trabalho utilizou da lematização padrão do *analyzer* da língua portuguesa no Elasticsearch. Como novas avaliações, outras estratégias de implementação léxica podem ser agregadas, como o *stemming*. Já para o método semântico, outros modelos para geração de embeddings podem ser utilizados e comparados.

Além disso, um trabalho futuro pode envolver uma proposta aprimorada dos critérios de relevância utilizados, bem como a melhoria do algoritmo de verificação de similaridade, que pode ter beneficiado a abordagem léxica devido à sua forma de utilização das distâncias. Também podem ser consideradas mudanças na fase de tratamento das bases de dados, que podem incluir desde o aprimoramento do mapeamento dos produtos relevantes até a incorporação de métodos para lidar com abreviações presentes nos dados. Por fim, outra abordagem promissora é a implementação de uma abordagem híbrida, estudando a mesclagem de técnicas, com espaço para comparações de desempenho com outras estratégias.

6. REFERÊNCIAS

[1] gov.br. Catálogo de Materiais do Governo Federal - CATMAT. Disponível em: https://www.gov.br/saude/pt-br/composicao/sectics/desid/catmat.

[2] SEBRAE. 2023. A importância da emissão da nota fiscal. Disponível em: https://sebrae.com.br/sites/PortalSebrae/artigos/a-importancia-da-emissao-da-nota-fiscal,857ed6387eab5810VgnVCM1000001b00320aRCRD

[3] Elasticsearch. O coração do Elastic Stack. Disponível em: https://www.elastic.co/pt/elasticsearch

[4] Paalman, Jasper & Mullick, Shantanu & Zervanou, Kalliopi & Zhang, Yingqian. (2019). Term Based Semantic Clusters for Very Short Text Classification.

[5] Nigam, P., Song, Y., Mohan, V., Lakshman, V., Ding, W., Shingavi, A., ... & Yin, B. (2019, July). Semantic product search. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2876-2885).

[6] Mangold, Christoph. (2007). A survey and classification of semantic search approaches. In: International Journal of Metadata, Semantics and Ontologies. Vol. 2(1), 2007, pp. 23-34. 2. 10.1504/IJMSO.2007.015073.

[7] Zhu, Y., Li, Y., Yue, Y., Qiang, J., & Yuan, Y. (2020). A hybrid classification method via character embedding in chinese short text with few words. *IEEE Access*, *8*, 92120-92128.

[8] gov.br. Catálogo de Materiais e Serviços. Disponível em: https://www.gov.br/compras/pt-br/sistemas/conheca-o-compras/catalogo

[9] Python. Release Python 3.10.0. Disponível em: https://www.python.org/downloads/release/python-3100/

[10] Elasticsearch. Elasticsearch version 8.12.0. Disponível em: https://www.elastic.co/guide/en/elasticsearch/reference/current/release-notes-8.12.0.html

[11] Python Elasticsearch Client. Disponível em: https://elasticsearch-py.readthedocs.io/en/v8.13.1/

[12] Shane Connelly. 2018. BM25 na prática — parte 2: o algoritmo BM25 e suas variáveis. Disponível em: https://www.elastic.co/pt/blog/practical-bm25-part-2-the-bm25-algorithm-and-its-variables

[13] Huggingface. Sentence-transformers/all-MiniLM-L6-v2. Disponível em: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[14] Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

[15] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Wei Chen, Tie-Yan Liu. 2013. A Theoretical Analysis of Normalized Discounted Cumulative Gain (NDCG) Ranking Measures. In Proceedings of the 26th Annual Conference on Learning Theory (COLT 2013).

Sobre o autor:

Melquisedeque Carvalho Silva é aluno do curso de Ciência da Computação na Universidade Federal de Campina Grande.