Análise comparativa entre tipos de consulta e modelos de RI com o TREC-COVID

Tainá da Silva Lima

Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil
tainaslima19@gmail.com

Resumo Neste ano tem estado em vigor uma das maiores crises de saúde pública dos últimos tempos: a pandemia do COVID-19. Muitos pesquisadores têm corrido para entender melhor o funcionamento da nova variação do vírus da família coronavírus, para que uma vacina seja produzida rapidamente a fim de evitar maiores perdas. Com isso, um número enorme de artigos científicos estão sendo produzidos, onde, a cada dia que passa, ele só aumenta, dificultando uma busca por informação mais apurada. A coleção de teste do TREC-COVID almeja ajudar pesquisadores biomédicos na recuperação de informação na forma de consultas mais relevantes pré-prontas. Sendo assim, utilizando um subconjunto pequeno do CORD-19 (coleção de artigos sobre o COVID-19), este artigo pretende analisar de modo comparativo a performance dos tipos de consultas oferecidas pela coleção de teste nos diferentes modelos básicos de recuperação: booleano, vetorial e probabilístico. Comparado aos testes realizados pelos próprios criadores do TREC-COVID, não foi obtido bons resultados em termos de ranking, apesar disso foi possível fazer algumas análises comparativas planejadas. Além disso, é discutido pontos interessantes relacionados à importância de um dataset de documentos de qualidade, tokenização e stemming dentro do escopo biomédico.

Palavras-Chave Recuperação da Informação· COVID-19· TREC-COVID.

1 Introdução

Com o advento da pandemia do COVID-19 no início deste ano, muitos setores no mundo todo sofreram baixas, e aqueles que já eram defasados, como a saúde pública brasileira, tiveram seus defeitos ainda mais expostos. O campo da informação não fugiu a essa regra e por conta da necessidade de se entender melhor o novo vírus, teve seu corpo científico literário aumentado expressivamente.

Além disso, na intenção de disseminar informação mais rápido, algumas plataformas de "open science" permitem que alguns artigos científicos sejam divulgados sem terem sido completamente julgados. Gera-se, assim, um embate de querer lançar o conhecimento para o corpo acadêmico o mais rápido possível e ao mesmo tempo tê-lo avaliado e julgado corretamente [1].

A partir desse cenário, foi criado o CORD-19 (COVID-19 Open Research Dataset)[1] [1] à pedido da Casa Branca dos Estados Unidos ao Allen Institute for

[1] https://www.semanticscholar.org/cord19

Artificial Intelligence, que permite que toda a comunidade acadêmica e políticos tenham acesso a informações estruturadas de pesquisas sobre o coronavírus, incluindo sua variação mais recente. Com base na coleção CORD-19 foi criado o TREC-COVID.

O TREC-COVID é uma avaliação comunitária e sua saída principal é uma coleção de teste, cujo objetivo principal é capturar a necessidade de informação de pesquisadores biomédicos em cima da literatura científica gerada na pandemia [1]. Tal avaliação é baseada no framework do Text Retrieval Conference (TREC), que é uma conferência que serve de suporte para pesquisas sobre Recuperação da Informação (RI), promovendo um meio para que possam ser feitas avaliações de larga escala de metodologias da área [2].

Ao deparar-se com a descrição do TREC-COVID em [1], foram elaboradas os seguintes intuitivos questionamentos:

- "Com as consultas da coleção são retornados documentos relevantes de fato?"
- "Como a coleção se comporta em diferentes modelos?"
- "Quais tipos de consulta trazem melhores resultados?"

Seguindo essa linha de raciocínio, este artigo tem como finalidade analisar comparativamente o desempenho da coleção de teste descrita pelo TREC-COVID tanto sob o aspecto das categorias das consultas quanto sob o dos diferentes modelos básicos de RI: booleano, vetorial e probabilístico. A coleção de documentos utilizada nesta implementação foi uma pequena fração do CORD-19 e a própria coleção de teste também foi reduzida. Além das observações mais evidentes, foram calculadas as métricas de avaliação da recuperação de informação Mean Average Precision (MAP) e Precision At 10 (P@10).

Organizado em seis seções principais, este artigo apresenta na próxima seção uma descrição do TREC-COVID: a motivação dos autores, o que é uma coleção de teste e sua definição. Nas seguintes, retrata o modo com o qual a implementação da análise foi feita em termos de aplicação computacional, expondo seus resultados e deduções baseadas nos mesmos. Por fim, são feitas as considerações finais sobre o projeto, além de narrar as dificuldades encontradas ao longo de sua construção e planos futuros de evolução do mesmo.

2 O TREC-COVID

2.1 Coleção de teste

Analisando o cenário dos últimos anos de pesquisa em Recuperação da Informação, as coleções de teste têm sido muito utilizadas para avaliar a eficácia das tecnologias da área, além de fornecer um modo de "encapsulamento" de um certo conjunto de necessidades de informação, facilitando o processo de recuperação. Essa ferramenta foi muito popularizada devido à campanhas como o TREC, o Cross-Language Evaluation Forum (CLEF), o NII Testbeds and Community for Information Access Research project (NTCIR) e outros [3]. Uma coleção de teste é um conjunto de tópicos ou de descrições de necessidades de

informação, ou seja, um conjunto de objetos de informação a serem buscados. Junto a isso, há uma listagem de julgamentos de relevância, que indicam quais documentos são relevantes para quais tópicos [3].

Scholer et. al (2016) menciona que, apesar da vantagem de contribuir para a análise dos resultados retornados por uma tecnologia de recuperação, o uso de coleções de teste pode ser problemático. Além das coleções de teste serem custosas e complexas de serem construídas, a característica que por um lado pode ser uma vantagem em termos de praticidade pode ser uma desvantagem, que é o fato de justamente esse tipo de ferramenta ser uma abstração do mundo real, onde decisões do que deve - e pode - ser representado e estudado foi feito por outras pessoas. Isso leva a uma possível limitação no que se é estudado e o que é considerado uma contribuição válida dentro da Recuperação da Informação [3].

2.2 Definição

Uma das principais dificuldades de uma pesquisa sustentada durante uma pandemia é a alta taxa de mudança nas referências bibliográficas científicas na qual ela está pautada. Tal característica está diretamente ligada ao esforço feito para se buscar informação, uma vez que os tópicos de interesse vão se desenvolvendo conforme o avanço do surto da doença e o volume de obras explodem.

Sendo assim, o TREC-COVID é um projeto de avaliação feito de maneira comunitária, desenhado principalmente para gerar uma coleção de teste que reflete as descrições de necessidades de informação de pesquisadores biomédicos em relação ao COVID-19, tendo como base o corpo literário científico construído durantes os últimos meses de pandemia. Segundo Voorhees et. al (2020), de maneira mais objetiva, o TREC-COVID tem como finalidade:

- Criar uma coleção de teste que retrata as necessidades de informação de pesquisadores biomédicos embasado na produção feita por cientistas durante a pandemia
- Avaliar algoritmos e sistemas de buscas para ajudar cientistas e outros a administrarem melhor o corpo literário sobre o COVID-19
- Descobrir métodos que irão auxiliar no gerenciamento de informação científica em futuras crises biomédicas globais

O framework do TREC O TREC-COVID é baseado no modelo do TREC (Text Retrieval Conference) que é uma conferência patrocinada pelo National Institute of Standards and Technology (NIST) e o U.S. Department of Defense, e oferece suporte à pesquisa dentro da comunidade de RI, focado em avaliações de tecnologias de recuperação de grande escala.

Em cada TREC, o NIST dispõe para os participantes uma coleção de teste e outra de documentos para que eles os utilizem como entradas para seus próprios sistemas de busca e que retornem para organização da conferência sua listagem de documentos mais relevantes. Os resultados são avaliados pelo próprio TREC e também há disponível um fórum para que os participantes possam compartilhar suas experiências [2].

As diferenças entre o TREC-COVID e o TREC são: cronograma de execução do projeto comprimido, os tópicos de cada rodada são super conjuntos dos conjuntos das rodadas anteriores e os julgamentos de relevância das rodadas anteriores são dados de treinamento para rodadas futuras [1].

Construção A organização do TREC-COVID se deu através do sistema de "rounds" (rodadas), onde cada uma é um desafio comunitário de avaliação independente. Uma "round" é formada por várias "runs" (execuções), em que uma "run" representa uma tarefa de busca usando o CORD-19 como coleção de documentos e a coleção de teste construída naquela rodada em específico, retornando os documentos ranqueados por tópico [1].

As "runs" são a fonte para definir um conjunto menor de documentos que serão julgados como "relevante", "parcialmente relevante" ou "não relevante", onde são associados aos números "2", "1" e "0" respectivamente nos arquivos de julgamentos de relevância. Tais julgamentos são feitos por assessores humanos especializados na área biomédica e áreas afins [1].

Estrutura da coleção de teste Como descrito previamente, as coleções de teste são dispostas por meio de tópicos que traduzem uma necessidade de informação do usuário e, no caso do TREC-COVID em específico, de pesquisadores biomédicos. Existem 50 tópicos atualmente, contudo, no tempo de publicação do artigo existiam 30 somente. Os tópicos foram elaborados pelos organizadores do projeto com treinamento biomédico, utilizando questionamentos feitos pelo consumidor ao National Library of Medicine, discussões feitas por médicos influenciadores nas redes sociais e sugestões feitas via Twitter pela hashtag "#COVIDSearch" levantada em março deste ano [1].

Cada tópico possui três tipos de consulta: "query", "question" e "narrative". A "query" é uma consulta pequena formada por palavras-chave sobre o tópico em específico, e a "question" é basicamente uma evolução da "query", sendo de fato uma pergunta utilizando mais afinco a linguagem natural. Já a "narrative" não é uma evolução da "question", ela é uma descrição maior e mais assertiva sobre quais dados são desejados no retorno da busca, utilizando termos mais específicos da área biomédica.

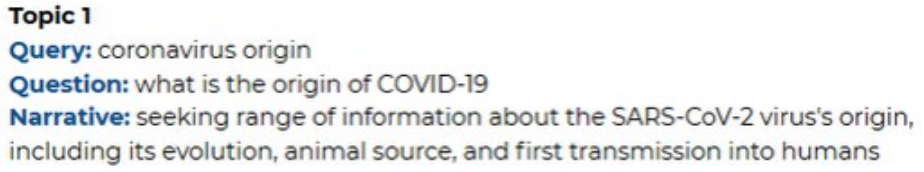

Figura 1. Tópico 1 da coleção de teste do TREC-COVID - Round 5

3 Trabalhos relacionados

As coleções de testes oferecidas pela conferência do TREC são umas das mais famosas no ramo da recuperação da informação, uma vez que são úteis na avaliação dos resultados das estratégias de recuperação. Ao ser utilizado para julgar a qualidade dos documentos fornecidos pela busca, elas permitem que seja feito uma análise da eficácia daquele modelo em específico, principalmente em comparação a outros considerados mais "clássicos"; sendo exatamente este o objetivo de Fernandez et. al (2009) em seu artigo.

Motivado pelo fato de que os métodos de avaliação para comunidade semântica são baseados em estudos centrados no usuário, fazendo com que estas metodologias sejam custosas, não escaláveis e difíceis de reproduzir; Fernandez et. al (2009) apresenta um benchmark de avaliação reutilizável com o propósito de realizar uma comparação cruzada entre os modelos clássicos de RI e aqueles baseados em ontologia. Assim como esta obra utiliza de um framework do TREC, o benchmark é formado por uma coleção de documentos advindas do TREC WT10g, consultas e julgamentos da web track TREC 9 e TREC 2002, ontologias que cobrem alguns tópicos das consultas usadas, bem como bases de conhecimento e anotações armazenadas num banco de dados MySQL [4].

A análise foi feita utilizando quatro abordagens diferentes de busca na Web: uma busca por palavra-chave (Lucene), a melhor busca automática do TREC, melhor busca manual do TREC e uma busca semântica que é uma adaptação do modelo vetorial para um modelo de RI baseado em ontologia. O estudo mostrou que a estratégia que se baseia em ontologia teve performances melhores ou semelhantes que as do Lucene e da melhor busca automática do TREC em 60% das consultas em média [4].

Seguindo a mesma linha de pesquisa sobre comparação entre modelos de recuperação da informação, Armenska e Zdravkova (2012) trazem em seu artigo uma discussão e análise da usabilidade de três conhecidos modelos de RI - modelo vetorial com normalização de tamanho de documento pivotado, modelo probabilístico BM25 e o de linguagem - no processo de selecionar o documento que melhor responde a pergunta de consulta.

Diferentemente de Fernandez et. al (2009) e do trabalho apresentado neste artigo, que fazem uso de coleções do próprio TREC ou baseadas nele, nesta análise é utilizado uma coleção de teste extraída das aulas frequentadas pelos autores que cobrem assuntos como a história da computação, conceitos básicos de Tecnologia da Informação e outros. Como não há o uso de uma coleção TREC, ao invés dos julgamentos de relevância aplicados nesta obra e em Fernandez et. al (2009), foi empregado como resultado de referência de ótimo a saída de resposta de um aprendizado de máquina. Os autores verificaram que o caso do modelo vetorial obteve uma corretude da classificação de 92%, onde, nos melhores casos, os probabilísticos e de linguagem atingiram os 90% e 92.5% respectivamente [5].

A ideia de análise levantada neste artigo se assemelha em muitos aspectos a outros publicados em anos anteriores, como os mencionados nesta seção. Apesar de terem em comum o uso de uma coleção de teste advinda da conferência TREC e almejar analisar o comportamento de diferentes modos de se recuperar

informação, Fernandez et. al (2009) se propõe a focar na comparação entre um modelo baseado em ontologia e os clássicos utilizados pela área, enquanto este trabalho pretende-se não só observar o desempenho entre modelos mas também entre categorias de consulta. Armenska e Zdravkova (2012) trazem também uma investigação análoga em termos de algumas abordagens de RI mas diferem na origem da coleção de teste.

4 Implementação

As conclusões apresentadas por meio deste artigo foi baseada na execução de uma aplicação computacional em que, de modo resumido, se propõe a tomar como entrada a coleção teste do TREC-COVID, a coleção de documentos CORD-19 e os modelos clássicos de recuperação da informação (booleano, vetorial e probabilístico); tendo como saída uma análise comparativa entre os tipos de consultas oferecidos pela coleção de teste, bem como entre os diferentes modelos utilizados para recuperar-se informação.

As versões dos arquivos da coleção de teste, julgamentos de relevância e também da coleção do CORD-19 empregados neste trabalho foram as associadas ao "Round 5" do TREC-COVID, diferentemente do que foi publicado em [1], que na época, só havia sido feito o "Round 1". Além disso, toda a implementação foi feita através da linguagem de programação Python em conjunto com bibliotecas disponíveis para a mesma.

Sua organização consiste de três grandes partes: a geração dos subconjuntos da coleção de documentos e consultas, a busca em si e a avaliação dos resultados gerados, como pode ser visto através do esquema da figura abaixo. Cada uma das partes será melhor descrita nas subseções a seguir.

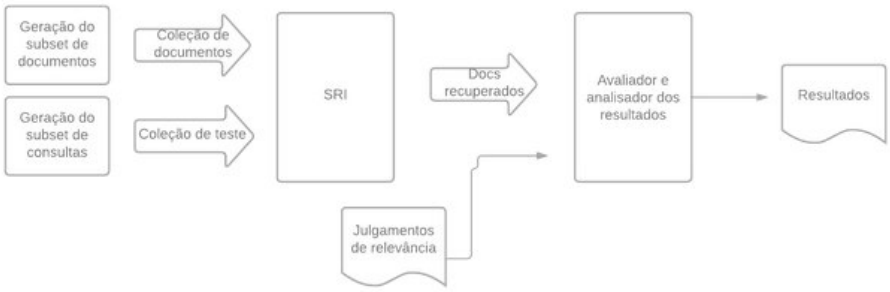

Figura 2. As 3 partes da implementação

4.1 Geração do subconjunto de documentos e consultas

Inicialmente, havia-se como proposta a utilização por completo dos tópicos da coleção de teste do TREC-COVID e fazer o mesmo para a coleção de documentos CORD-19. Contudo, a solução se tornou inexecutável em termos de infraestrutura computacional nessas condições: o arquivo mais recente do CORD-19 que possui os textos dos artigos abrangentes pelo dataset tem mais de 5 GB de tamanho (suas versões antigas não diferem tanto desse valor), sendo inviável de ser trabalhado com a máquina simplista disponível para este trabalho.

Posto isto, tomou-se como uma primeira solução usar o arquivo "metadata.csv", uma planilha oferecida pelo projeto do CORD-19 que possui metadados sobre os artigos contemplados pela coleção, contendo vários dados, incluindo o próprio título e resumo das obras. Portanto, foi utilizado como documentos onde seriam feitas as buscas, os resumos dos artigos. Todavia, mesmo com essa redução brusca na quantidade de palavras a serem processadas, ainda era difícil executar a aplicação com todos os resumos disponíveis, logo foi feita uma redução para 50 (cinquenta) documentos de coleção.

Esses cinquenta documentos não foram escolhidos de modo aleatório, uma vez que, ao fazer isso, poderíamos ter uma coleção em que nenhum de seus itens nem mencionam o termo "coronavírus", o que atrapalha por completo a busca, dado que basicamente quase todos os tópicos da coleção de teste possui ao menos uma das nomenclaturas associadas ao vírus. A coleção foi gerada escolhendo resumos que mencionasse um ou os dois termos: "COVID-19" e "coronavírus".

Em relação às consultas, a princípio seria aplicada à busca todas as construídas pelo TREC-COVID, porém, decidiu-se depois focar somente nos 10 (dez) primeiros tópicos, o que culmina num total de 30 consultas (10 tópicos onde cada um possui 3 tipos de consultas diferentes). Essa decisão foi tomada na intenção de trazer mais documentos que estivessem mapeados com tais tópicos no arquivo de julgamentos de relevância e que se conseguisse melhores resultados no fim.

4.2 Sistema de recuperação da informação

Para este trabalho foi implementado as partes de pré-processamento para documentos e consultas, bem como os próprios modelos de recuperação booleano, vetorial e probabilístico. Para o pré-processamento, a biblioteca NLTK[2] foi amplamente usada, tanto para a tokenização, remoção de stopwords e stemming, onde este último é o algoritmo de Porter.

Já para os modelos de RI, eles foram elaborados baseados nos exercícios feitos pela autora para a disciplina de Recuperação da Informação ministrada pela Professora Giseli Lopes no Departamento de Ciência da Computação da UFRJ, com algumas pequenas adaptações para a nova aplicação. Sendo assim, de modo mais específico, os modelos empregados nesse trabalho foram: o booleano com matriz de incidência, vetorial com similaridade pelo cosseno e ponderação TF-IDF e o probabilístico BM25. Este último apesar de ter sido feito, não obteve

[2] https://www.nltk.org/

resultados melhores do que uma implementação já pronta do BM25Plus da biblioteca "rank_bm25"[3], então, no fim, ficou-se com esta da biblioteca.

4.3 Análise dos resultados

A avaliação dos resultados é ancorada principalmente em duas métricas de avaliação da recuperação da informação: Mean Average Precision (MAP) e Precision At 10 (P@10). Ambas foram calculadas para os modelos vetorial e probabilístico que são aqueles que possuem valor de ranking. Além disso, também foi calculado indicadores de qualidade dos resultados de elaboração própria, como a taxa de documentos relevantes em 1º lugar para o caso probabilístico. Como fonte para a geração do conjunto de relevância, é manipulado o arquivo contendo os julgamentos de relevância concedido pelo próprio TREC-COVID.

5 Resultados

Mesmo tendo como entrada coleções reduzidas e implementações de autoria própria, que não se comparam a sistemas de recuperação da informação sofisticados e conceituados da área, ainda era esperado um bom resultado quando analisado os documentos retornados para cada tipo de consulta em cada modelo, em relação aos julgamentos de relevância oferecido pelo TREC-COVID. No entanto, devido a três hipóteses que serão discutidas ao longo desta seção, a resposta obtida não foi a esperada e soa incorreta, porém dado tais suposições são saídas que podem fazer sentido.

Para efeitos de comparação, no fim, o modelo booleano não foi considerado, principalmente pelo fato de que não há um cálculo de ranking associado a ele, somente o retorno de quais documentos seriam supostamente relevantes. Mesmo assim, ele foi implementado e está descrito na aplicação.

5.1 Dos diferentes tipos de consulta

Do ponto de vista das três diferentes categorias de consultas - "query", "question" e "narrative" - pode-se verificar que, dentre os 21 primeiros tópicos da coleção de teste, as "narrative" obtiveram a maior porcentagem de documentos relevantes em 1º lugar nos rankings para o caso probabilístico, que foi de 33,3% contra os 23,81% da "query" e 19,05% da "question". Ou seja, para cada um dos 21 tópicos, 33.3% dos primeiros lugares possuem um documento ou relevante ou parcialmente relevante. Com isso, podemos ver que em muitos tópicos, mesmo usando tipos de consulta diferentes, ainda se tem muitos documentos irrelevantes em 1º lugar nos rankings.

Focando especificamente no tópico 1 (ver Figura 1) com a estratégia probabilística BM25Plus, é notável a grande quantidade de documentos ranqueados cuja relevância associada a eles é de "irrelevante". Ainda sim, existe uma certa

[3] https://pypi.org/project/rank-bm25/

narrative		
Number of relevant docs		7
Number of irrelevant docs		13
Percentage of relevant docs		33,33%
Percentage of irrelevant docs		61,90%

Figura 3. Documentos relevantes/irrelevantes em 1º lugar do tipo "narrative"do caso probabilístico

semelhança em quais artigos foram retornados nas três categorias. Nesse caso em específico, ainda que a "narrative" tenha gerado mais "acertos" em termos de relevância, a "query" trouxe mais relevantes nos três primeiros lugares.

Probabilistic Model						
Position	Query		Question		Narrative	
	cord_uid	relevance	cord_uid	relevance	cord_uid	relevance
1	t4ns3syl	2	ienet82k	0	tcrxm7jy	2
2	cns5rnj8	1	dc30gkfe	0	8fmykb4c	0
3	tcrxm7jy	2	t4ns3syl	2	t4ns3syl	2
4	ienet82k	0	cns5rnj8	1	ienet82k	0
5	8fmykb4c	0	tcrxm7jy	2	dc30gkfe	0
6	dc30gkfe	0	8fmykb4c	0	zv0ysi8m	1
7	8kqhvxlz	2	8kqhvxlz	2	cns5rnj8	1
8	3qzlo90e	0	zv0ysi8m	1	lk67yfrp	0
9	lk67yfrp	0	3qzlo90e	0	0t2a5500	2
10	0t2a5500	2	lk67yfrp	0	3qzlo90e	0

Figura 4. Top 10 ranking das consultas do tópico 1 para o modelo probabilístico

O fato da busca com a "question" "what is the origin of COVID-19" inserida no tópico 1 ter nos devolvido na primeira posição um documento de relevância "0" foi um evento intrigante para que fosse feita uma investigação para entender o motivo de tal fenômeno ter ocorrido. Verificando, então, o artigo de identificador "ienet82k", intitulado "Coronavirus: Hotspot on coronavirus disease 2019 in India" [6], podemos averiguar que em seu resumo há a menção de que seria discutido em seu texto a origem do vírus, bem com outras informações. Porém, ao estudar o artigo por completo, é válido deduzir a baixa relevância dele para responder a consulta em questão, pois ele não oferece tantos detalhes sobre a origem do COVID-19 quanto seria a intenção de um pesquisador biomédico ao questionar a pergunta para o sistema de buscas.

Sendo assim, associado à este caso porém não somente, elucido a hipótese do porquê tais resultados foram poucos satisfatórios:

Hipótese 1 (H1): *Qualidade da coleção de documentos*
Quando se fala em qualidade, não se refere ao fato de que os documentos estão de alguma maneira "corrompidos" ou são inconsistentes. Ao falar de qualidade, diz respeito ao ponto de que os julgamentos de relevância gerados pelo TREC-COVID são baseados em buscas sob os artigos por completo e não somente seus resumos. Isso tornaria de algum modo a listagem de documentos retornados válida, pois, no fim, esta implementação baseia-se em resumos e não textos completos. Todavia, isso pode gerar falsos positivos, como foi o caso do artigo "ienet82k".

5.2 Dos diferentes modelos de RI

Em relação à comparação de comportamentos do ponto de vista de modelos de recuperação da informação, empregando o uso da métrica de avaliação Precision At 10 (P@10), é notável que a abordagem vetorial possui uma leve melhora em relação ao probabilístico em consultas "question", porém possui uma performance pior no caso das "narrative". É claro que, num geral, todos os casos apresentados na tabela abaixo obtiveram valores de P@10 baixos, remetendo novamente à hipótese declarada na subseção anterior.

Topics	Query		Question		Narrative	
	Probabilistic	Vectorial	Probabilistic	Vectorial	Probabilistic	Vectorial
1	0,3	0,3	0,3	0,4	0,2	0,2
2	0	0	0	0	0	0
3	0,2	0,2	0,1	0,2	0,1	0,2
4	0,2	0,2	0,1	0,1	0	0
5	0	0	0	0	0	0
6	0,3	0,3	0,1	0,1	0,1	0
7	0,1	0,1	0,1	0,1	0	0
8	0,1	0,1	0	0	0	0
9	0,1	0,1	0,2	0,1	0	0
10	0,1	0,1	0,1	0,1	0,1	0,1

Figura 5. P@10 para todas as combinações possíveis entre os modelos probabilístico e vetorial com cada categoria de consulta

Por outro lado, se calcularmos o Mean Average Precision (MAP) para todas essas combinações, conclui-se que o vetorial possui melhores números no caso de "questions" e os piores nos outros gêneros de consulta. Esses valores nos fazem refletir novamente sobre a motivação desse episódio generalizado, o que nos permite enunciar as outras duas hipóteses.

MAP			
	Query	**Question**	**Narrative**
Probabilistic Model	0,001335282	0,001335	0,001020251
Vectorial Model	0,001465244	0,000989	0,000820718

Figura 6. MAP para os modelos probabilístico e vetorial para cada tipo de consulta

Hipótese 2 (H2): *Tamanho da coleção de documentos*
O tamanho da coleção é importante para a investigação do acontecimento de tais episódios de valores baixos de P@10 e MAP, dado que muitos dos documentos relevantes de fato ficaram de fora da busca. Além disso, nem todos os documentos relevantes mapeados para um tópico em específico estava na coleção de documentos. Ou seja, a conjuntura de se estar recuperando em meio a alguns documentos irrelevantes, seja sendo genuinamente não relevante ou tido como um por não estar mapeado no arquivo de julgamentos de relevância para o tópico em questão, supostamente poderia afetar as métricas finais.

Hipótese 3 (H3): *Tokenização e stemming não adequados*
Jiang e Zhai (2007) discorrem em seu texto sobre a importância de um tokenizador e um stemmer adequado quando se trata de recuperação da informação em textos biomédicos, uma vez que esses textos utilizam-se muitos termos específicos da área. Tal ideia pode ser claramente vista na experimentação deste artigo. Acredita-se que o uso de um tokenizador direcionado para textos em inglês "comuns" (o oferecido pela biblioteca NLTK), ou seja, não científicos, tenha impactado fortemente nos resultados finais da busca, dado que muitos dos termos biomédicos podem ter sido "separados" erradamente. E, além disso, podem ser sofrido stemming de modo incorreto, o que tornaria a situação ainda mais insatisfatória.
Uma tokenização apropriada pode melhorar a performance da recuperação em até 96%, o que é um número altamente significativo [7]. Jiang e Zhai (2007) vão além e afirmam que diferentes heurísticas de tokenização devem ser escolhidas para diferentes tipos de consulta, por exemplo, o uso da remoção de caracteres especiais e substituição de letras gregas por latinas são as melhores abordagens para aquelas que possuem somente símbolos de genes. Ademais, a aplicação de stemming melhora o desempenho da busca para consultas prolixas [7].

6 Conclusão e trabalhos futuros

Este artigo tem por objetivo o estudo do desempenho dos algoritmos clássicos de recuperação da informação, booleano, probabilístico e vetorial, baseando-se na coleção de testes do TREC-COVID e na coleção de documentos sobre o COVID-19 e e sua família coronavírus, CORD-19. Para o estudo, foi desenvolvida uma aplicação em que sua maior parte é constituída de implementações de autoria própria.

Do ponto de vista de resultados, pode-se dizer que eles não foram muito satisfatórios e foram levantadas as seguintes hipóteses para a ocorrência de tal fenômeno: qualidade da coleção de documentos, o tamanho da mesma e a inadequação dos tokenizadores e stemmer utilizados. Apesar disso, é possível concluir que o tipo "narrative" de consulta obteve a melhor porcentagem de documentos relevantes em 1º lugar e que, ao analisar o P@10, as "query" apresentam um comportamento mais "estável" do que as outras. Além disso, a partir da métrica MAP, é perceptível a maior precisão do modelo vetorial em "query" e sua baixa performance nas outras categorias.

Com este trabalho, foi capaz de se perceber a importância de uma boa escolha da coleção de documentos que servirá de "local de busca" para a recuperação da informação, bem como o modo em que os termos tanto dessa coleção quanto a de teste serão processados. Ademais, algumas dificuldades em relação a limitações computacionais influenciaram fortemente nos resultados apresentados.

Sendo assim, pretende-se aperfeiçoar este projeto empregando o uso de um tokenizador e stemmer mais apropriados ou, até mesmo, usar um lematizador ao invés de um tokenizador, como o BioLemmatizer para a linguagem Java, que faz a lematização voltada especificamente para área médica e biomédica. Existe também o processo de lematização implementado em Python incluído na biblioteca NLKT, contudo ele é mais genérico, sendo para qualquer assunto, o que não necessariamente irá garantir melhora nos resultados. Por fim, caso haja a oportunidade de acesso à uma máquina computacional de melhor desempenho e poder de execução, seria interessante a análise utilizando o CORD-19 com seus textos completos.

Referências

1. Voorhees, E., Alam, T., Bedrick, S., Demner-Fushman, D., R. Hersh, W., Lo, K., Roberts, K., Soboroff, I., Lu Wang, L.: TREC-COVID: Constructing a Pandemic Information Retrieval Test Collection. ACM SIGIR Forum. SIGIR (2020).
2. Text REtrieval Conference (TREC) Home Page, https://trec.nist.gov/.
3. Scholer, F., Kelly, D., Carterette, B.: Information retrieval evaluation using test collections. Information Retrieval Journal. 19, 225-229 (2016).
4. Fernandez, M., Lopez, V., Sabou, M., Uren, V., Vallet, D., Motta, E., Castells, P.: Using TREC for cross-comparison between classic IR and ontology-based search models at a Web scale. Semantic Search 2009 Workshop at the 18th International World Wide Web Conference (WWW 2009). 18th International World Wide Web Conference (WWW 2009), Madrid, Spain (2009).
5. Armenska, J., Zdravkova, K.: Comparison of information retrieval models for question answering. Proceedings of the Fifth Balkan Conference in Informatics on - BCI '12. (2012).
6. Cheke, R., Shinde, S., Ambhore, J., Adhao, V., Cheke, D.: Coronavirus: Hotspot on coronavirus disease 2019 in India. Indian Journal of Medical Sciences. 72, 29-34 (2020).
7. Jiang, J., Zhai, C.: An empirical study of tokenization strategies for biomedical information retrieval. Information Retrieval. 10, 341-363 (2007).