

Trabalho Prático 7 - Recuperação da Informação

Matheus Aquino Motta¹

¹Bacharelado em Matemática Computacional
DCC - Universidade Federal de Minas Gerais

matheusaquino199@gmail.com.br 2018046513

Abstract. *During the semester, on the subject of Information Retrieval, we developed a complete search engine, from the implementation of a web crawler, to a working querying system for relevant documents. The purpose of this assignment consisted of evaluating the search engine developed throughout the semester by analyzing the presented results with metrics such as precision and recall. We will give a brief walk through into the last assignments and the engine developed to properly discuss its fundamental results.*

Resumo. *Durante o semestre, na disciplina de Recuperação da Informação, desenvolvemos uma máquina de busca completa, desde a implementação de um coletor de páginas web, Crawler, até um sistema de consultas para a requisição de páginas relevantes para um determinado conjunto de termos, a partir do modelo vetorial e do sinal de Pagerank dos sítios. Agora, nesse relatório iremos brevemente passar pelos componentes do buscador desenvolvidos nos últimos trabalhos e discutir os resultados obtidos no trabalho final. Assim, a proposta dessa tarefa consiste na avaliação da máquina de busca construída a partir de métricas de precisão e revocação dos resultados apresentados.*

1. Introdução

Ao longo do semestre fizemos a construção de uma máquina de busca na *web* que contemplou as partes fundamentais de uma máquina de busca real, desde a construção de coletor de documentos HTML na *web*, e sua respectiva indexação, até um sistema de consultas que a partir do modelo vetorial e sinais de *Pagerank* das páginas, retornaria um conjunto de sítios *web* potencialmente relevantes para uma dada requisição de termos.

No último trabalho prático tivemos de modificar o nosso sistema de consultas, a fim de associar o sinal do *Pagerank* das páginas ao peso de similaridade dos resultados dado pelo modelo vetorial. Com isso, classificamos os resultados de topo gerados pelas máquinas de busca da turma, determinando quais documentos eram ou não relevantes para uma determinada consulta.

Assim, agora com o sistema devidamente construído e resultados classificados, iremos brevemente analisar as partes centrais do sistema, a fim de caracterizar a máquina de busca e descrever os resultados finais e a qualidade da máquina de busca desenvolvida. Para realizar essa análise iremos utilizar curvas de precisão e revocação, geradas através das *urls* de topo adquiridas nas consultas. Entretanto, já é válido ressaltar que os resultados foram extremamente abaixo do esperado, e não respaldam o projeto implementado e a qualidade efetiva da máquina de busca, devido à falhas na avaliação dos resultados.

2. A Máquina de Busca

Agora, iremos passar brevemente por cada componente principal da máquina de busca desenvolvida, que pode ser ilustrada pelo fluxograma abaixo.

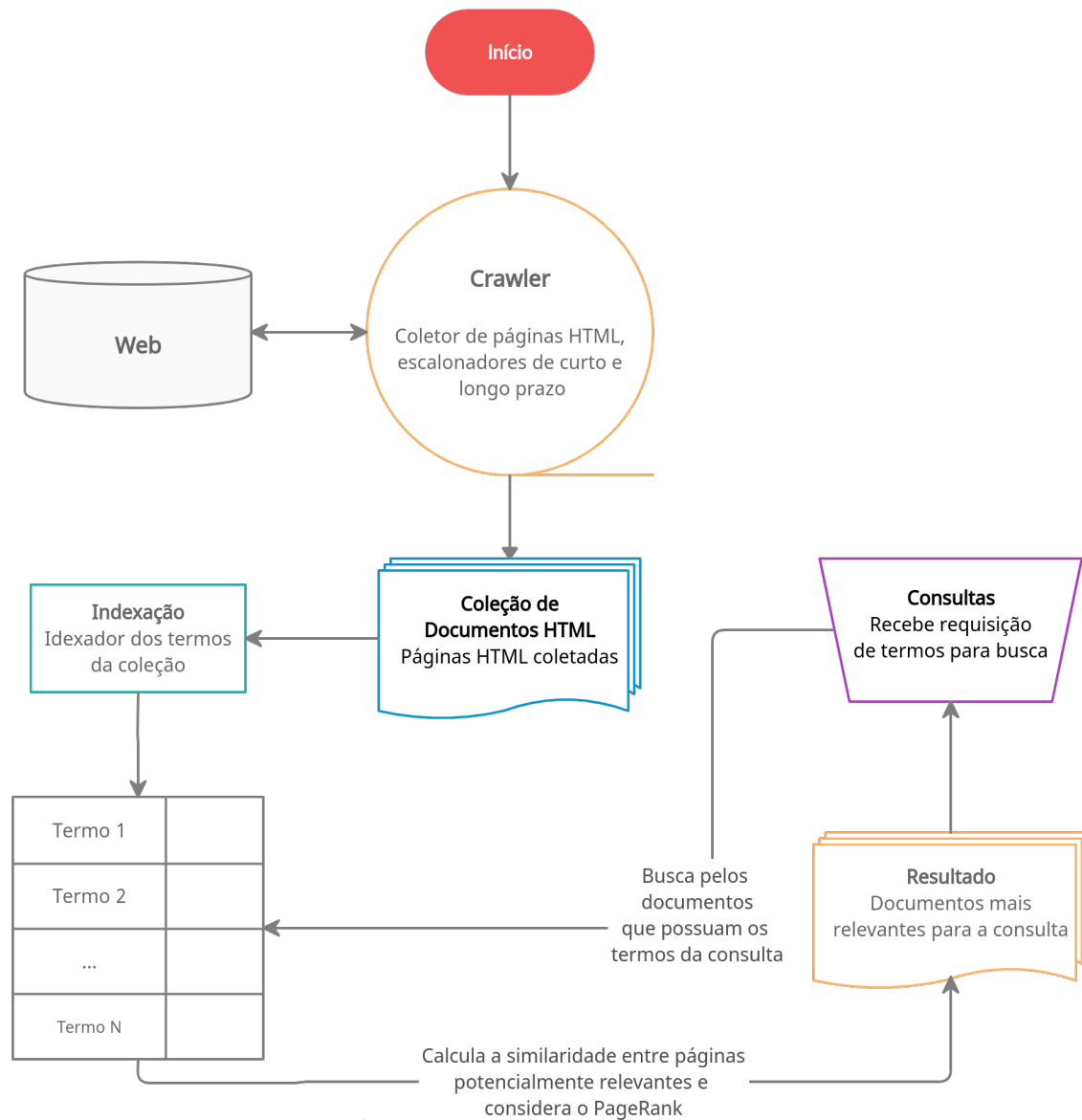


Figure 1. Fluxograma da Máquina de Busca.

2.1. Crawler, coletor HTML

O primeiro componente implementado foi o coletor de documentos HTML, *Crawler*, desenvolvido nos trabalhos práticos 1 e 2 da disciplina, que tinha como objetivo realizar a coleta de uma coleção de páginas HTML na *web*.

O coletor foi implementado a partir da biblioteca *Chilkat*, onde utilizamos um Escalonador de Curto e Longo prazo, *Short-Term Scheduler and Long-Term Scheduler*, para realizar uma busca em largura por páginas HTML na *web*, priorizando enlaces de domínios

brasileiros variados. Assim, visando satisfazer restrições de polidez da *web*, mantendo certa eficiência a partir do uso de *multi-threads* para a busca.

Os resultados apresentados nesse projeto foram extremamente satisfatórios, haja vista que conseguimos coletar com sucesso uma coleção de 100000 documentos HTML estritamente em domínios brasileiros. Essas, dentre as quais parte significativa consistia de páginas de *url* com pesos baixos, isto é, mais próximas a página principal dos seus respectivos domínios.

A explicação detalhada da implementação, desenvolvimento e resultados do coletor construído pode ser encontrada **aqui**.

2.2. Indexador

Após a construção do coletor, o próximo componente da máquina de busca consistiu na implementação de um indexador para os termos existentes no texto de cada página HTML coletada, relativa aos trabalhos práticos 3 e 4 da disciplina.

Assim, primeiramente no trabalho prático 3 desenvolvemos um sistema simples de indexação para os documentos da nossa mini-coleção coletada previamente. Onde para cada termo no vocabulário da coleção armazenamos em um *hash* o número de documentos no qual o termo ocorre, dado por n_i , e um vetor de 3-uplas nas quais a primeira posição é o documento em que o termo ocorre, e a segunda e terceira são respectivamente o número de vezes em que o termo ocorre no documento, e as posições em que o termo ocorre no texto.

Nesse sentido, no trabalho prático 4 foi fornecida uma coleção significativamente maior de documentos HTML, com cerca de 1000000 de arquivos. Assim, por limitações *hardware* foi necessária uma mudança na abordagem de construção do indexador, na qual agora o index seria construído em múltiplas partes em arquivos separados e posteriormente unidos, por meio de técnicas de ordenação em memória externa similares ao *merge sort*.

Além do arquivo principal de index dos termos do vocabulário, também foi construído um arquivo auxiliar dicionário para facilitar o acesso às informações relativas a cada termo no index principal. Esse, que por sua vez armazena as posições em memória do arquivo do index relativas a cada termo do vocabulário, assim dado um termo, conseguimos acessar diretamente as informações relativas a ele no index.

A explicação detalhada do indexador foge do escopo desse relatório, entretanto uma explicação mais detalhada acerca da implementação e resultados obtidos pode ser encontrada **aqui** para o indexador simples (trabalho prático 3), e **aqui**, para o indexador principal (trabalho prático 4) utilizado para a execução dos demais trabalhos.

2.3. Consultas, Modelo Vetorial e PageRank

Com os arquivos de index já devidamente construídos conseguimos acessar de forma minimamente eficiente informações relativas a um dado termo na coleção, como os documentos em que ocorre, quantas vezes e suas respectivas posições no texto.

Assim, agora dada uma consulta q com k termos iremos extrair quais documentos possuem todos os k termos da requisição dadas. Dessa forma, tomados os documentos

que satisfazem essas condições, iremos ranqueá-los a partir do conceito de similaridade por cossenos, ou mais especificamente, o Modelo Vetorial.

Cada requisição q pode ser representada por meio de um vetor

$$\vec{q} = (w_{1q}, w_{2q}, \dots, w_{kq})$$

para qual os elementos w_{iq} representam o peso do i -ésimo termo na requisição q , que pode ser obtido através da fórmula

$$w_{iq} = (1 + \log_2 f_{i,q}) \times \log_2 \frac{N}{n_i}$$

Onde $f_{i,q}$ representa TF , isto é, a frequência do termo i na requisição q , isto é, o número de vezes que i ocorre na requisição e $\log_2 \frac{N}{n_i}$ representa o IDF do termo i na coleção, onde N é o número total de documentos da coleção e n_i o número de documentos em que i ocorre.

De modo análogo, cada documento d_j será representado por um vetor

$$\vec{d}_j = (w_{1j}, w_{2j}, \dots, w_{kj})$$

para qual os elementos w_{ij} representam o peso i -ésimo termo no documento j , que pode ser obtido por uma fórmula similar à representada acima

$$w_{ij} = (1 + \log_2 f_{i,j}) \times \log_2 \frac{N}{n_i}$$

Só que agora estamos interessados nas informações relativas ao documento j .

Assim, calculando a similaridade entre o vetor da consulta para os vetores de documentos, podemos obter os potencialmente mais relevantes para a nossa requisição a partir de uma pontuação, *score*, de acordo similaridade entre os cossenos dos vetores \vec{q} e \vec{d}_j .

Nesse viés, podemos obter a pontuação de similaridade dada pelo modelo vetorial a partir da equação abaixo

$$\begin{aligned} \text{sim}(\vec{d}_j, \vec{q}) &= \cos(\theta) = \frac{\langle \vec{d}_j, \vec{q} \rangle}{\|\vec{d}_j\|_2 \|\vec{q}\|_2} \\ &= \frac{\sum_{i=0}^k w_{iq} \times w_{ij}}{\sqrt{\sum_{i=0}^k w_{ij}^2} \times \sqrt{\sum_{i=0}^k w_{iq}^2}} \end{aligned}$$

Esse sistema de requisições e ranqueamento implementado foi detalhadamente explicado no relatório do trabalho prático 5 e pode ser encontrado **aqui**.

Além disso, para melhorar o funcionamento do sistema de requisições e qualidade das respostas foi adicionado um sinal de *Pagerank* p_d , à pontuação de similaridade entre consultas e documentos. Isto é, agora a similaridade também pode ser dada por uma combinação linear entre a pontuação dada pelo modelo vetorial $\text{sim}(\vec{d}_j, \vec{q})$ e um valor real p_{dj} , isto é, por uma pontuação $\text{simp}(\vec{d}_j, \vec{q})$, tal que

$$\text{simp}(\vec{d}_j, \vec{q}) = \alpha \text{sim}(\vec{d}_j, \vec{q}) + \beta p_{dj}$$

Onde α e β são constantes de normalização.

Os valores reais de *Pagerank* disponibilizados para cada documento em geral podiam ser da ordem de 10^{-6} , assim, sendo a similaridade do modelo vetorial um valor definido no intervalo $0 \leq \text{sim}(\vec{d}_j, \vec{q}) \leq 1$, foi necessário realizar modificações consistentes nos valores através da aplicação de uma função de ativação sigmoide $s(p_d)$ a todos os valores p_d , definida como

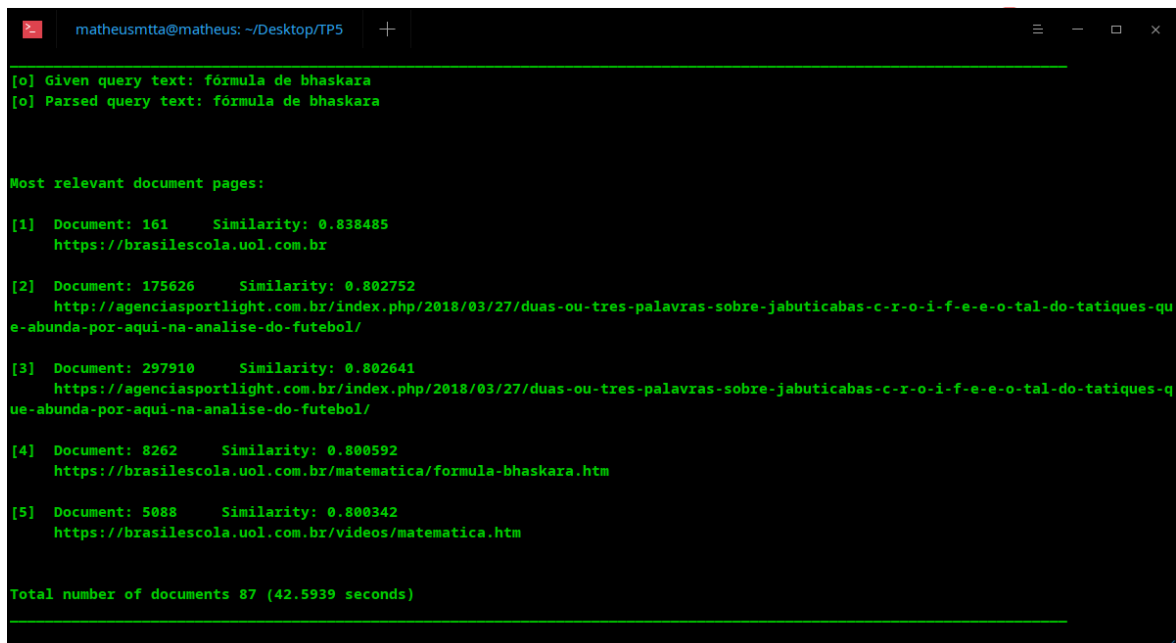
$$s(x) = \tanh\left(\frac{xe^{\pi^2}}{2}\right)$$

de modo que os valores de $s(p_d)$ ficaram minimamente bem comportados no intervalo de $0 \leq s(p_d) \leq 1$ proporcionalmente aos seus valores originais, sendo assim mais impactantes no cálculo de similaridade dado por $\text{simp}(\vec{d}_j, \vec{q})$. Portanto, temos que $\text{simp}(\vec{d}_j, \vec{q})$ será dada por

$$\text{simp}(\vec{d}_j, \vec{q}) = \alpha \text{sim}(\vec{d}_j, \vec{q}) + \beta s(p_{d_j})$$

Dessa forma, tomando um $\alpha = 0.8$ e um $\beta = 0.2$, conseguimos uma combinação linear de resultado que dá maior peso à similaridade vetorial, e um peso de "desempate" ao valor dado pelo *Pagerank*, onde $0 \leq \text{simp}(\vec{d}_j, \vec{q}) \leq 1$.

Os resultados apresentados por essa abordagem demonstraram-se significativamente positivos, outras foram testadas, entretanto essa foi a que mais garantiu páginas tão relevantes para os termos da consulta, quanto para sítios confiáveis.



```
matheusmta@matheus: ~/Desktop/TP5
[o] Given query text: fórmula de bhaskara
[o] Parsed query text: fórmula de bhaskara

Most relevant document pages:

[1] Document: 161      Similarity: 0.838485
    https://brasilescola.uol.com.br

[2] Document: 175626   Similarity: 0.802752
    http://agenciasportlight.com.br/index.php/2018/03/27/duas-ou-tres-palavras-sobre-jabuticabas-c-r-o-i-f-e-o-tal-do-tatiques-q
e-abunda-por-aqui-na-analise-do-futebol/

[3] Document: 297910   Similarity: 0.802641
    https://agenciasportlight.com.br/index.php/2018/03/27/duas-ou-tres-palavras-sobre-jabuticabas-c-r-o-i-f-e-o-tal-do-tatiques-q
ue-abunda-por-aqui-na-analise-do-futebol/

[4] Document: 8262      Similarity: 0.800592
    https://brasilescola.uol.com.br/matematica/formula-bhaskara.htm

[5] Document: 5088      Similarity: 0.800342
    https://brasilescola.uol.com.br/videos/matematica.htm

Total number of documents 87 (42.5939 seconds)
```

Figure 2. Exemplo de consulta, $q = \text{"fórmula de bhaskara"}$.

Essa consulta apresentou 3 (Resultados 1, 4, 5) resultados significativamente relevantes do Brasil Escola, altamente relacionados ao conteúdo potencialmente buscado na consulta. Entretanto, os resultados mostram que *Pagerank* possuía um peso maior que o desejado, haja vista que o resultado de número 4 seria o mais relevante em potencial, em detrimento do resultado de número 1, que possui um *Pagerank* maior (O tempo de requisição elevado é dado pela presença da preposição "de" nos termos da consulta.).

3. Resultados

Assim, com a máquina de busca construída e testada, disponibilizamos os resultados de topo para 20 consultas, que seriam posteriormente avaliados para uma futura análise da qualidade dos resultados produzidos pela máquina de busca. Para cada consulta foram tomados 5 resultados de topo, totalizando em aproximadamente 100 *urls* distintas que seriam avaliadas. Nesse sentido, o gráfico abaixo mostra o número de resultados, i.e, páginas *web* avaliadas, dentre os resultados da máquina de busca utilizando o sistema de recomendação final, que faz uso do *Pagerank* e outro que utiliza apenas o modelo vetorial.

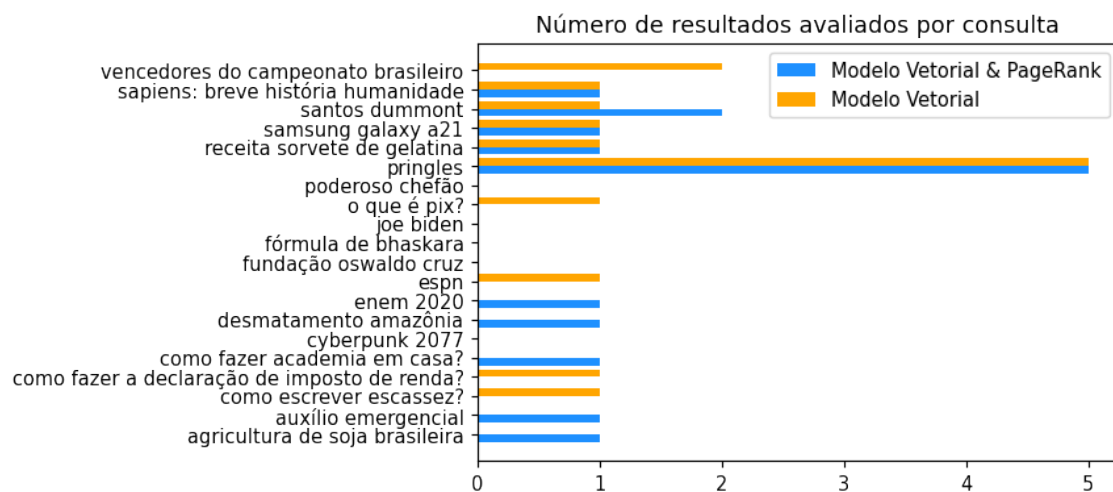


Figure 3. Número de resultados avaliados por consulta.

Vale ressaltar que os resultados para a consulta "pringles" foram todos avaliados, entretanto, a consulta original seria "pringles 118g", para a qual não haviam documentos que satisfizessem a consulta.

Dentre os documentos avaliados, para analisar a qualidade dos resultados estamos interessados naqueles que foram considerados relevantes para suas respectivas consultas.

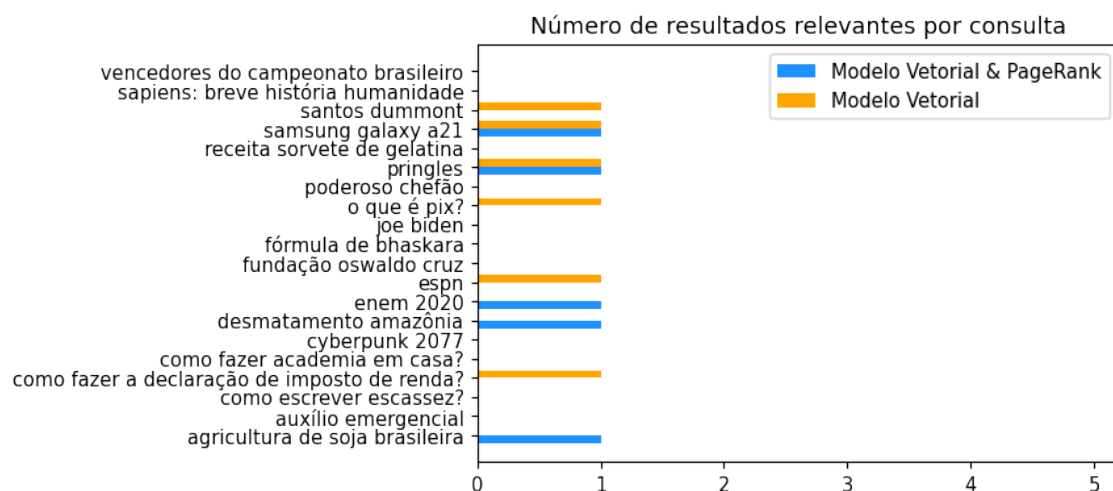


Figure 4. Número de resultados avaliados relevantes por consulta.

Cada documento foi avaliado com uma nota entre 1 e 5, e assim foi considerado relevante apenas se pelo menos metade das avaliações fossem maiores ou iguais a 3.

Dessa forma, para fazer a análise de qualidade dos resultados foram traçadas curvas de precisão e revocação. Onde seja R o conjunto de documentos relevantes, A a resposta gerada pela máquina de busca, tomando a interseção entre os conjuntos $R \cap A$, podemos calcular a precisão do buscador

$$\text{Precisão} = \frac{|R \cap A|}{|A|}$$

Isto é, a proporção de documentos relevantes recuperada pela máquina de busca na consulta em relação ao número de documentos retornados. E a revocação da máquina de busca,

$$\text{Revocação} = \frac{|R \cap A|}{|R|}$$

Isto é, a proporção de documentos relevantes retornados em relação ao número de documentos relevantes para aquela consulta.

Por definição das métricas de precisão e revocação, todas as páginas do conjunto de respostas do topo A deveriam ser avaliadas. Entretanto, como mostrado pelo gráfico acima, pouquíssimos resultados produzidos pela máquina de busca foram avaliados, o que influenciou significativamente a análise do resultado. Haja vista que, embora a máquina de busca tenha retornado alguns bons resultados de topo para parte majoritaria das consultas, como exemplificado pela Figura 2, eles não foram avaliados. Para a consulta "fórmula de bhaskara", notavelmente, nenhum dos resultados foi avaliado, o que resultou em curvas de precisão e revocação irrelevantes que não respaldam os resultados, performance ou qualidade do programa implementado.

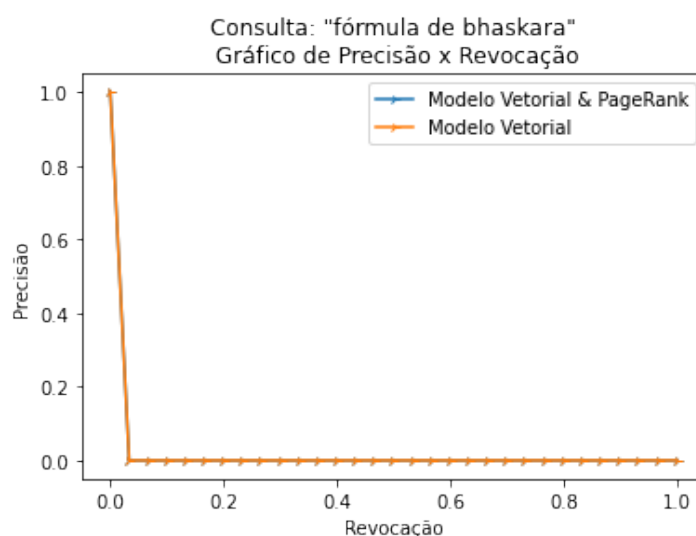


Figure 5. Curvas de precisão e revocação para a consulta "fórmula de bhaskara".

As curvas de precisão e revocação foram traçadas tomando os 5 resultado de topo da máquina de busca para cada consulta e fazendo uma interpolação para a revocação em 30 pontos, a partir da função

$$P(r_i) = \max_{\forall r | r_i \leq r} P(r)$$

Assim, obtemos o seguinte gráfico de precisão e revocação média para as 20 consultas realizadas

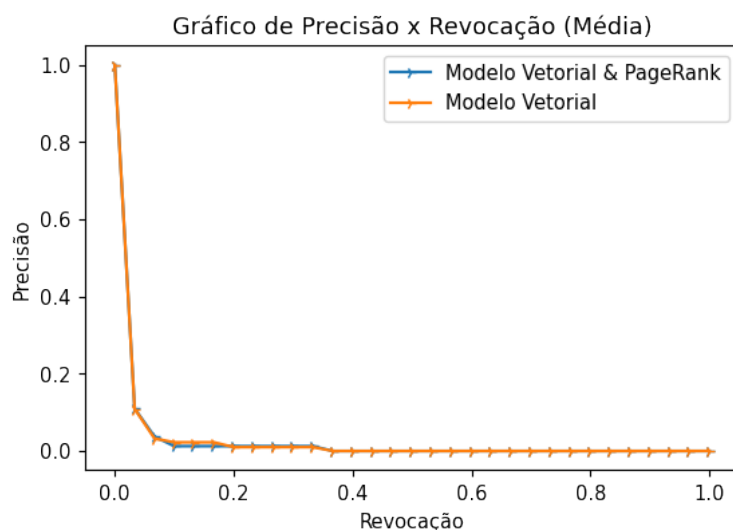


Figure 6. Curvas de precisão e revocação médias para as 20 consultas realizadas.

Embora as curvas de precisão e revocação tenham sido construídas seguindo as especificações dadas em aula, a falta de documentos avaliados tornou a visualização dos resultados significativamente impraticável, impossibilitando uma análise mais profunda acerca da performance da máquina de busca e qualidade dos resultados retornados. Isso notabiliza-se quando visualizamos que para o Modelo Vetorial & PageRank, apenas 15 das aproximadamente 100 páginas únicas foram avaliadas, ademais destacam-se as consultas como *"poderoso chefe"*, *"joe biden"*, *"fórmula de bhaskara"*, *"fundação osvaldo cruz"*, *"cyberpunk 2077"*, que não tiveram nenhum dos resultados disponibilizados avaliados, ou seja, de 20 consultas 6 tem o resultado irrelevante para fins conclusivos. Não obstante, embora algumas consultas tenham tido resultados avaliados, esses eram majoritariamente não relevantes para as consultas realizadas, onde cada consulta teve no máximo um resultado relevante avaliado, totalizando em 5 resultados relevantes avaliados para o "Modelo Vetorial & PageRank" (Obs: Os documentos de topo enviados para análise foram os resultantes das consultas utilizando o "Modelo Vetorial & PageRank").

4. Conclusões

O objetivo central do trabalho consistia na análise da performance e qualidade dos resultados apresentados pelo sistema de recomendação desenvolvido, que por sua vez é uma das partes centrais de uma máquina de busca e fundamental para conclusão do projeto construído ao longo do semestre. Entretanto, a baixa quantidade de documentos avaliados impossibilitou uma caracterização e análise mais detalhada dos resultados, que não respaldou a qualidade do sistema implementado. O que evidencia a não trivialidade da construção de tais sistemas, haja vista que dependem sensivelmente de uma avaliação sistemática dos documentos de interesse na coleção, para a avaliação de sua performance e qualidade.

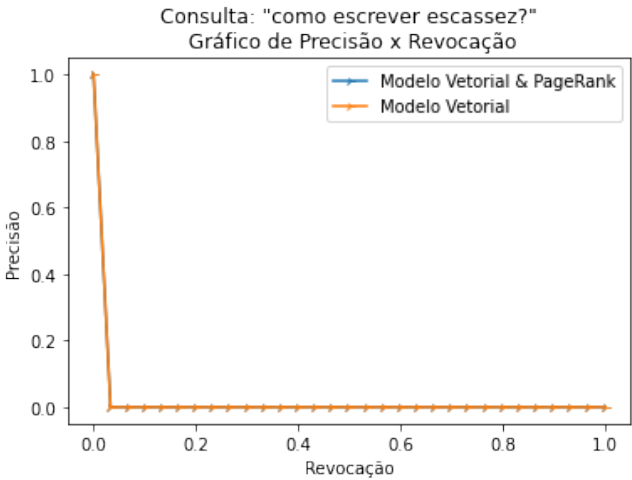
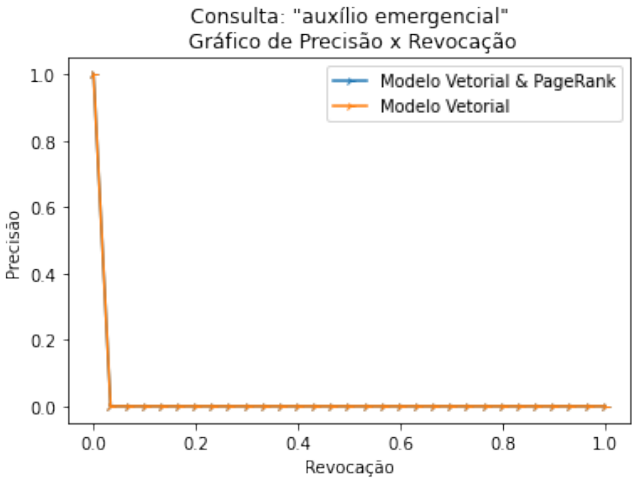
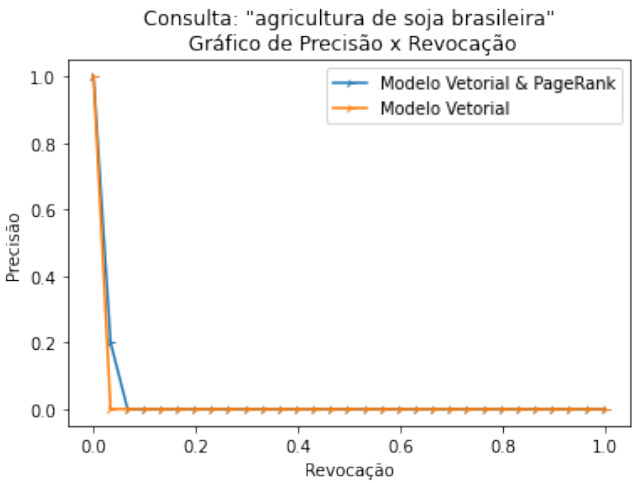
Algumas otimizações a priori podem ser com certeza adicionadas, como um sistema de recomendação que considere a posição relativa entre os termos da consulta no texto e

alterações no método de cálculo de similaridade que ainda dá um peso elevado ao *Pagerank* das páginas, impactando negativamente no resultado de algumas requisições.

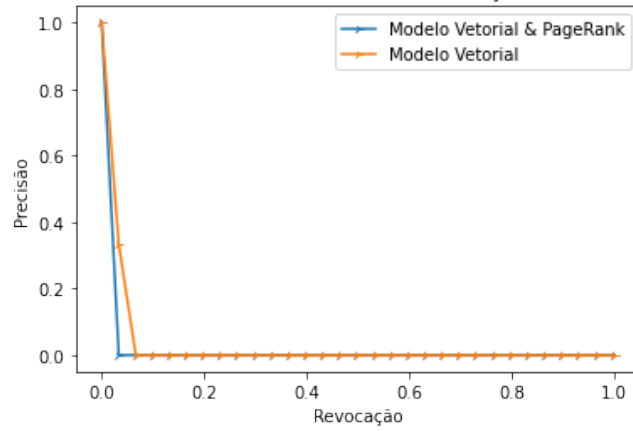
Pessoalmente, além disso, adoraria descobrir quais seriam os impactos do uso da computação quântica nas transições estocásticas de estado dadas pelos passeios aleatórios *Pagerank*, a partir do uso de caminhadas quânticas, e ver como essas poderiam impactar métricas, performance e eficiência desses sistemas em grandes quantidades de dados/documentos.

De um modo geral, a construção de um projeto como esse em um curto período de 4 meses foi extremamente gratificante e desafiadora, devido à complexidade dos trabalhos práticos e densidade do conteúdo apresentado. Tornando extremamente satisfatório o aprendizado e a observação da máquina de busca em funcionamento e os resultados qualitativos apresentados.

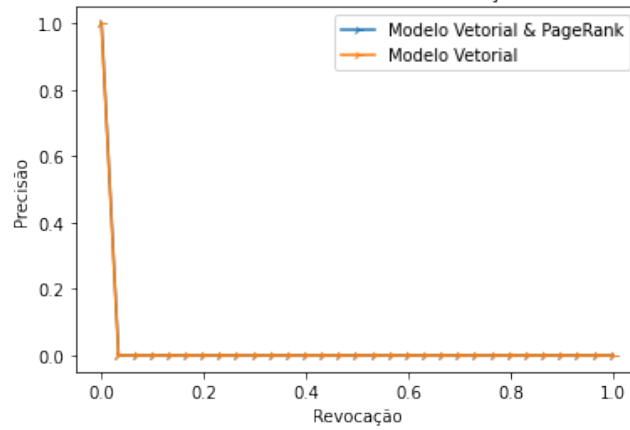
5. Apêndice (Curvas de Precisão x Revocação)



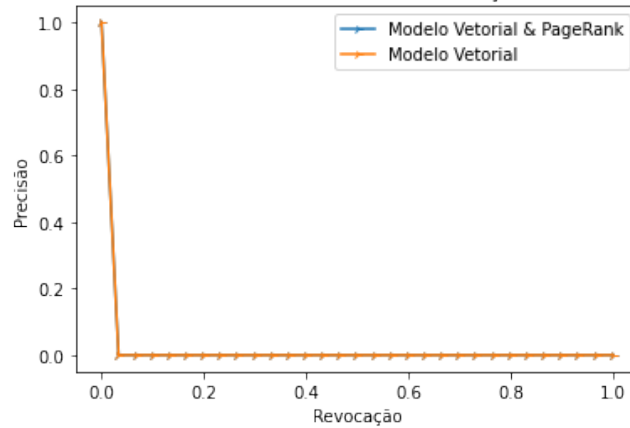
Consulta: "como fazer a declaração de imposto de renda?"
Gráfico de Precisão x Revocação



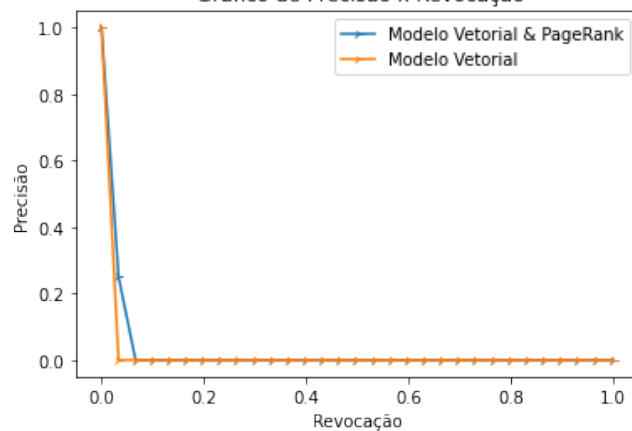
Consulta: "como fazer academia em casa?"
Gráfico de Precisão x Revocação



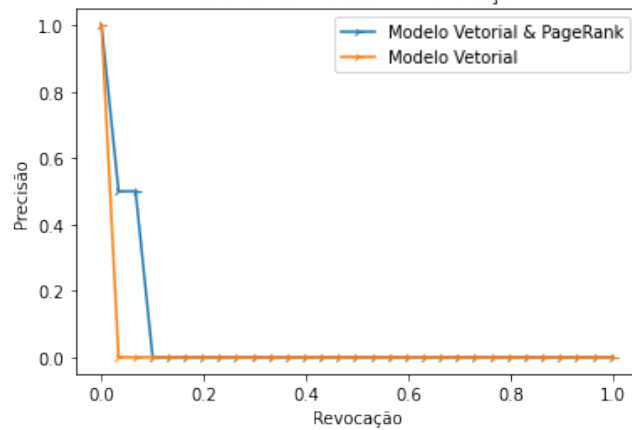
Consulta: "cyberpunk 2077"
Gráfico de Precisão x Revocação



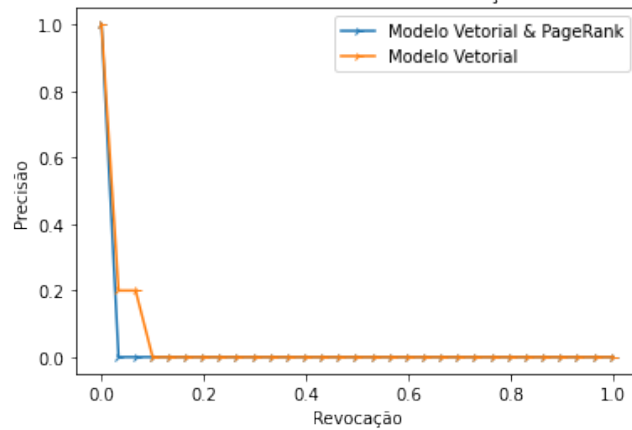
Consulta: "desmatamento amazônia"
Gráfico de Precisão x Revocação

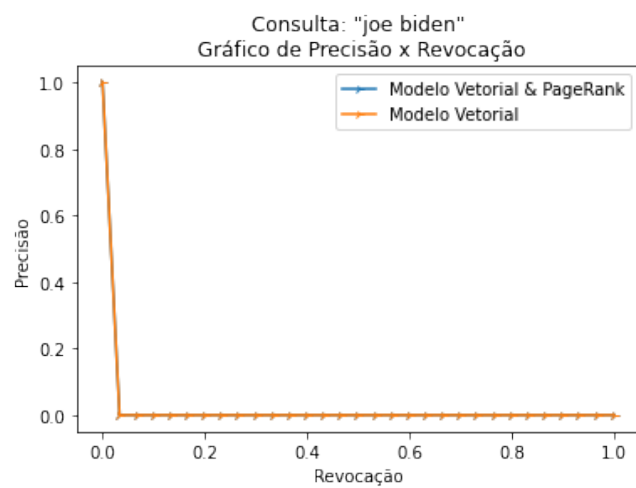
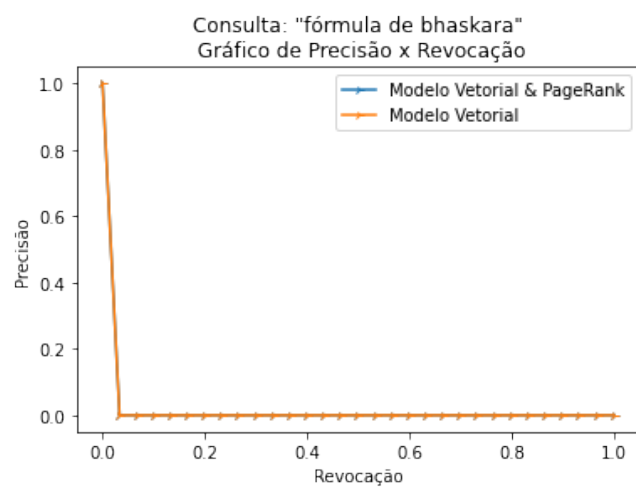
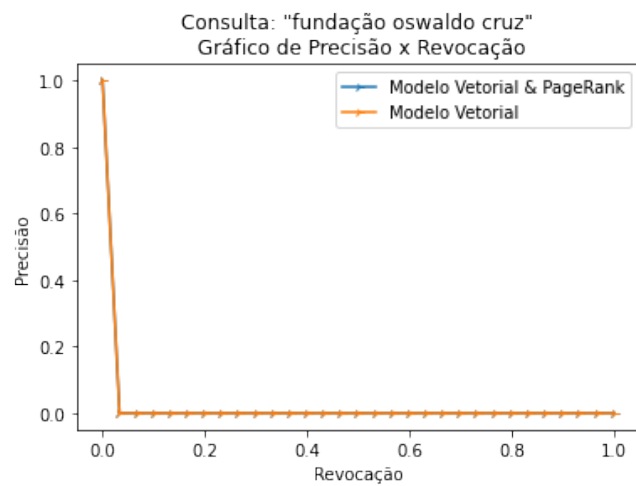


Consulta: "enem 2020"
Gráfico de Precisão x Revocação

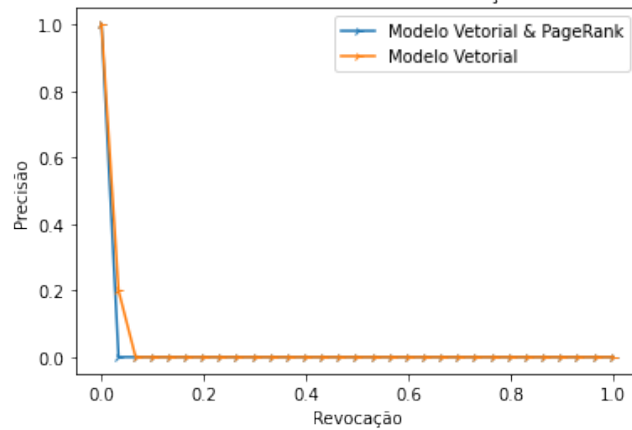


Consulta: "espn"
Gráfico de Precisão x Revocação

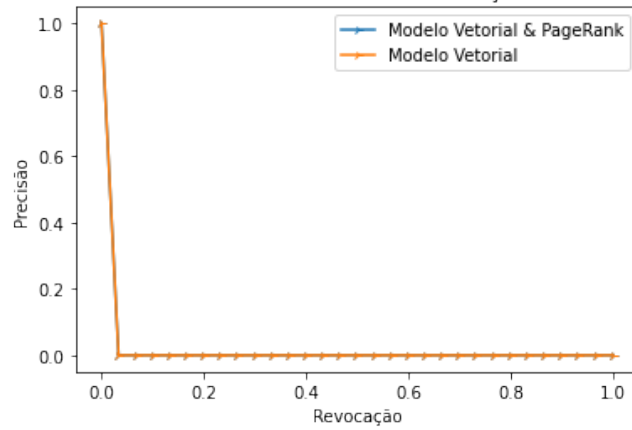




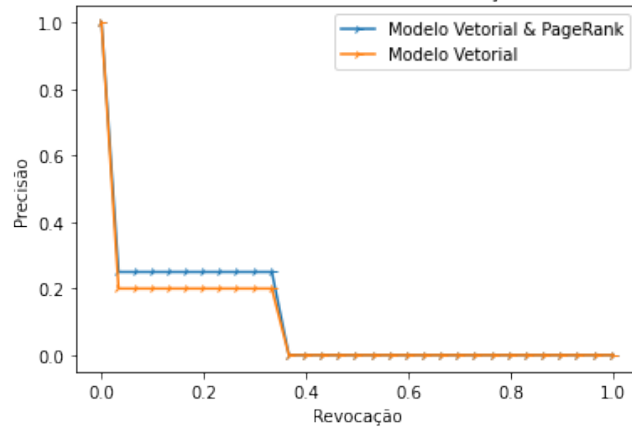
Consulta: "o que é pix?"
Gráfico de Precisão x Revocação



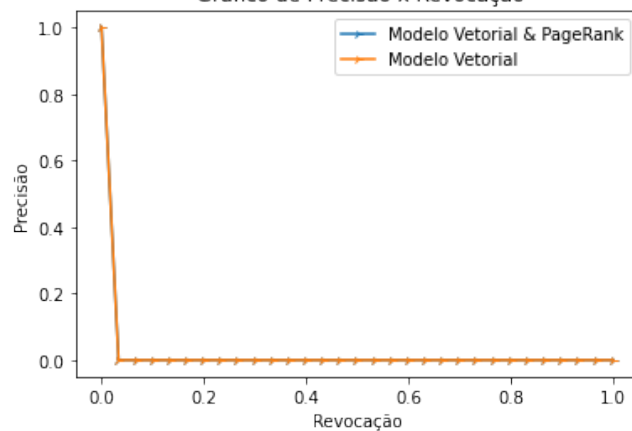
Consulta: "poderoso chefe"
Gráfico de Precisão x Revocação



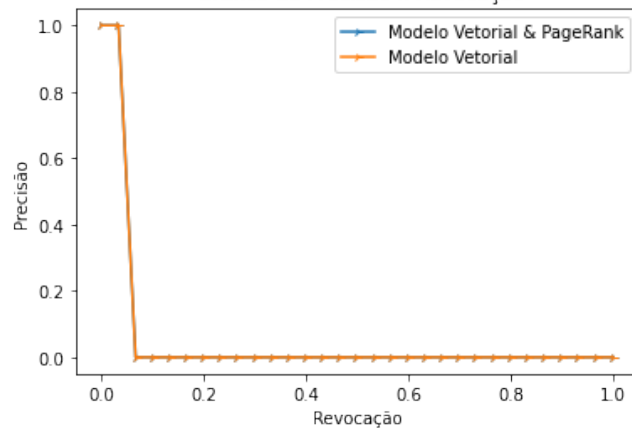
Consulta: "pringles"
Gráfico de Precisão x Revocação



Consulta: "receita sorvete de gelatina"
Gráfico de Precisão x Revocação



Consulta: "samsung galaxy a21"
Gráfico de Precisão x Revocação



Consulta: "santos dummont"
Gráfico de Precisão x Revocação

