Análise dos resultados das consultas
Recuperação de Informação

Daniel Ferreira Abadi[1]
2018088062

[1] Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brasil

1. Introdução

Ao longo do semestre foi construído um buscador do zero, desde a parte onde recolhemos as páginas da web brasileira, ao tratamento e limpeza das mesmas e, por fim, a construção do buscador. Esse último foi construído utilizando o modelo vetorial e o sinal de "pagerank". Agora, como uma etapa final, apresentaremos os dados das 20 consultas realizadas, cujos resultados foram avaliados pelos próprios alunos.

2. Pagerank

Antes de apresentarmos a análise dos dados, focaremos em como o sinal de "pagerank" foi utilizado. Como nos foi sugerido, poderíamos utilizar uma função para suavizar e deixar o valor do "pagerank" compatível com a similaridade gerada pelo modelo vetorial. Tal função foi escolhida sendo a sigmoide, onde temos:

$$f(x) = \frac{1}{2} + \frac{1}{2}tanh(\frac{\lambda x}{2})$$

Sendo λ um peso que podemos escolher. Para este trabalho o peso escolhido foi de $10^{3.1}$, e x o sinal nos provido. Abaixo podemos ver gráficos do antes e do depois dos dados passarem por essa função.

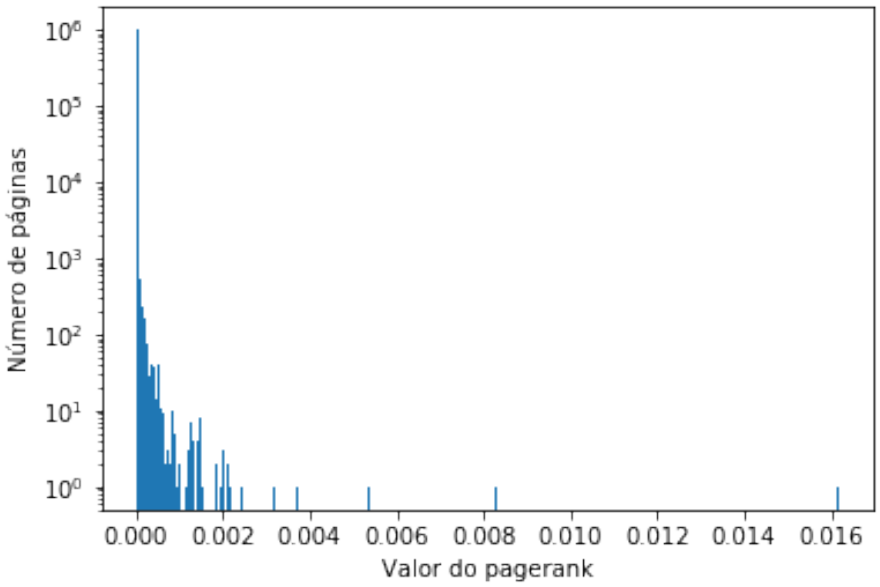

Figura 1. "Pagerank" antes de passar pelo tratamento, em escala logarítmica

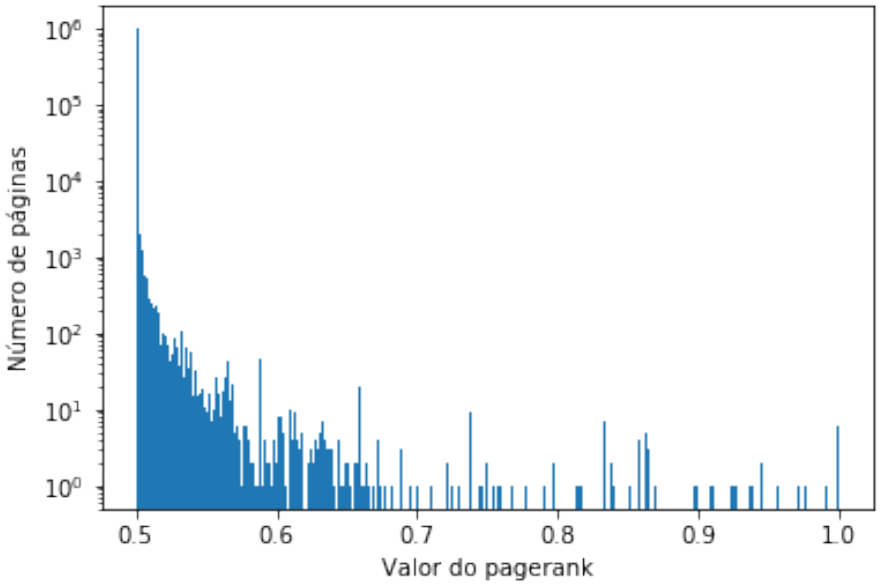

Figura 2. "Pagerank" após passar pelo tratamento, em escala logarítmica

Após essa transformação, o valor gerado foi multiplicado por uma constante, digamos C, e somado ao valor da similaridade que foi multiplicado por $1 - C$.

3. Análise geral

Para a realização da análise foi utilizada a avaliação das páginas que foram retornadas pelas consultas. O modelo de avaliação se deu por uma nota de 1 a 5, sendo as notas 1 e 2 resultados irrelevantes, e de 3 a 5 resultados considerados relevantes, mas com graus de relevância diferentes. E, com base nesses resultados, vamos medir a qualidade da busca calculando a precisão e a revocação para cada uma das páginas retornadas nas consultas.

A medida de revocação é dada pela fração dos documentos relevantes que foram recuperados. E a precisão é a fração dos documentos recuperados que são relevantes. Para ficar mais claro, temos as seguintes fórmulas:

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

Onde R é o conjunto de documentos relevantes para uma consulta e A é o conjunto de respostas geradas por uma máquina de busca para uma consulta.

Como dito anteriormente, os alunos avaliaram as páginas obtidas como resposta nos buscadores da turma, foram avaliadas apenas 600. Infelizmente só foram avaliadas as páginas que mais apareceram entre os resultados, tendo várias páginas ficado de fora de tal avaliação.

Abaixo temos dois gráficos barra mostrando a quantidade de páginas retornadas que foram avaliadas e a quantidade que foi considerada relevante, tanto pelo modelo vetorial com o "pagerank", quanto pelo mesmo sem o "pagerank".

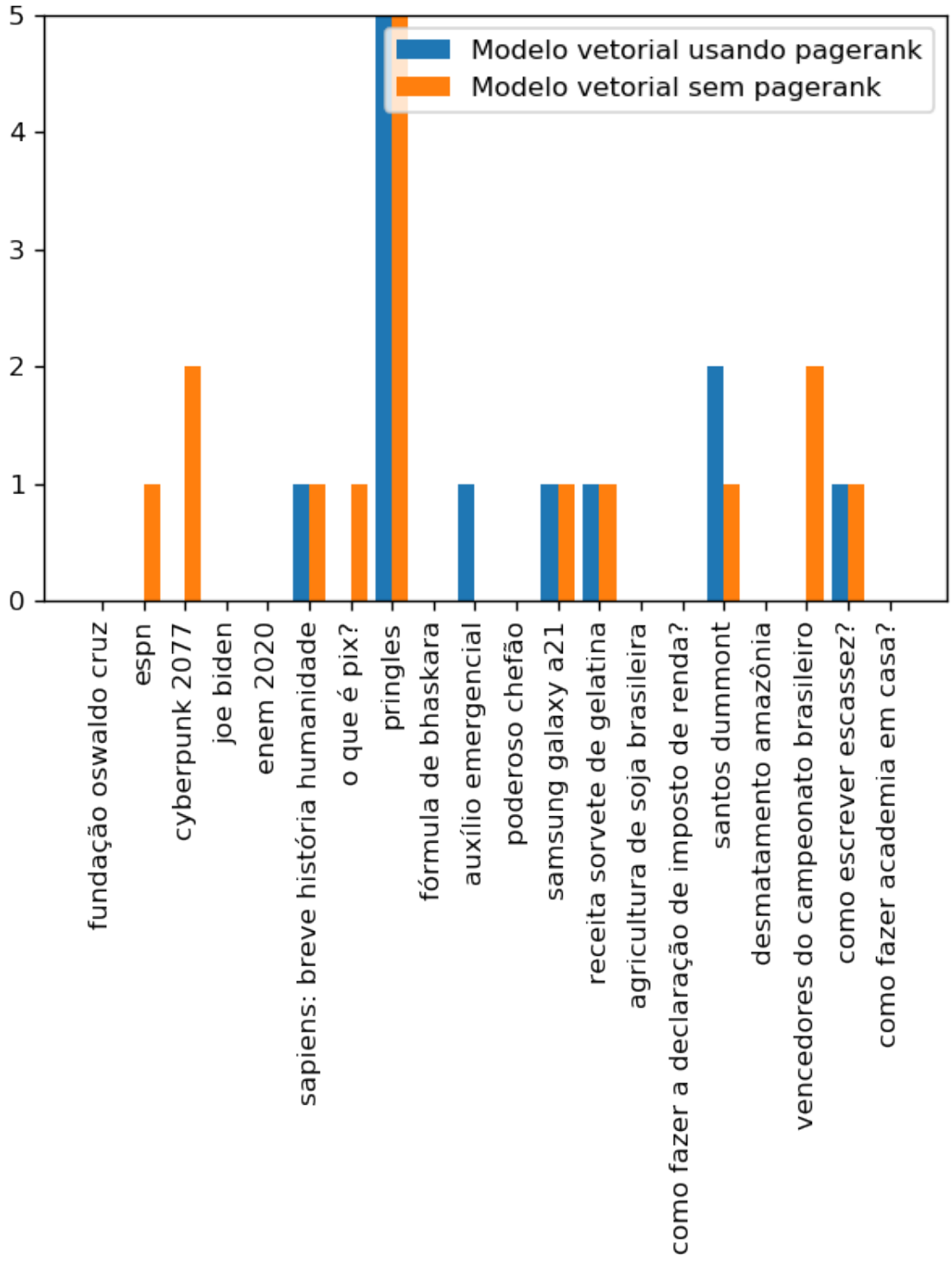

Figura 3. Número de resultados avaliados por consulta utilizando modelo vetorial com "pagerank"

Como podemos ver, apenas 12 resultados do modelo vetorial utilizando o "pagerank" foram avaliados. E, para o modelo vetorial, 16 resultados foram avaliados. Com tais números fica evidente um possível equívoco no método utilizado para avaliação das respostas das máquinas de buscas.

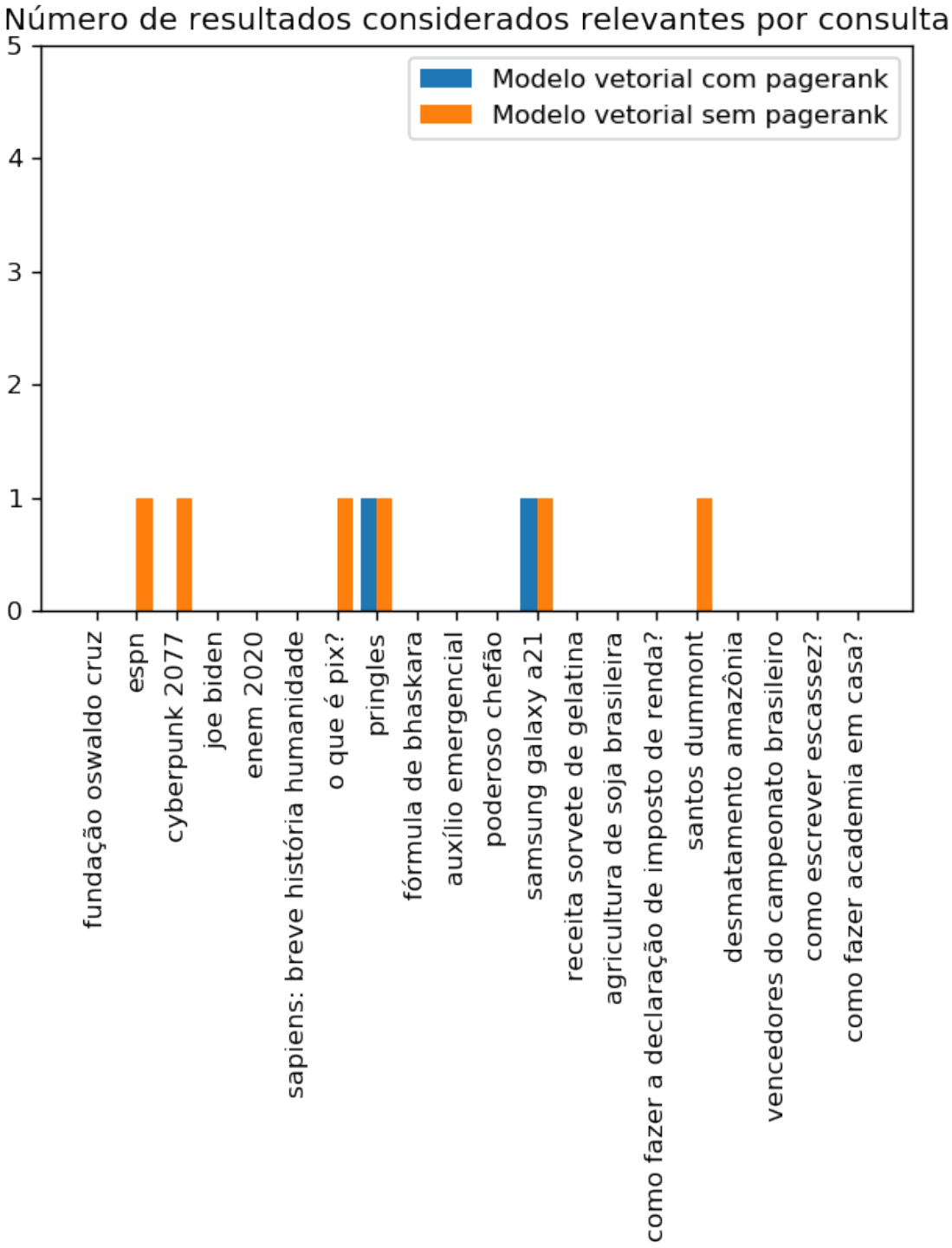

Figura 4. Resultados considerados relevantes

Para embasar a possível falha citada acima, vamos apresentar os resultados obtidos para a consulta “fórmula de bhaskara” utilizando o modelo vetorial com “pagerank”, onde nenhum resultado foi avaliado. A imagem abaixo nos mostra que os resultados para a consulta foram bons, mas infelizmente nenhum dos mesmos chegaram a ser avaliados.

```
Carregando dados...
Tempo de carregamento: 16.396 segundos.
Digite sua consulta: fórmula de bhaskara
Convertendo consulta: fórmula de bhaskara
Resultados encontrados: 87
Primeiros resultados:
ID da página: 162, endereço da página:
https://brasilescola.uol.com.br

ID da página: 5089, endereço da página:
https://brasilescola.uol.com.br/videos/matematica.htm

ID da página: 8263, endereço da página:
https://brasilescola.uol.com.br/matematica/formula-bhaskara.htm

ID da página: 709, endereço da página:
https://exercicios.brasilescola.uol.com.br/exercicios-matematica

ID da página: 421, endereço da página:
https://escolakids.uol.com.br/matematica/

Tempo de consulta: 123.039 segundos.
------------------------ // ------------------------
```

Figura 5. Resultados retornados pela consulta

Os resultados das consultas “fundação oswaldo cruz”, “joe biden”, “enem 2020”, “fórmula de bhaskara”, “poderoso chefão”, “agricultura de soja brasileira”, “como fazer a declaração de imposto de renda?”, “desmatamento amazônia” e “como fazer academia em casa?” não foram avaliados em nenhuma das versões. Abaixo teremos a curva de precisão e revocação média, incluindo todas as consultas nos quais os resultados não foram avaliados.

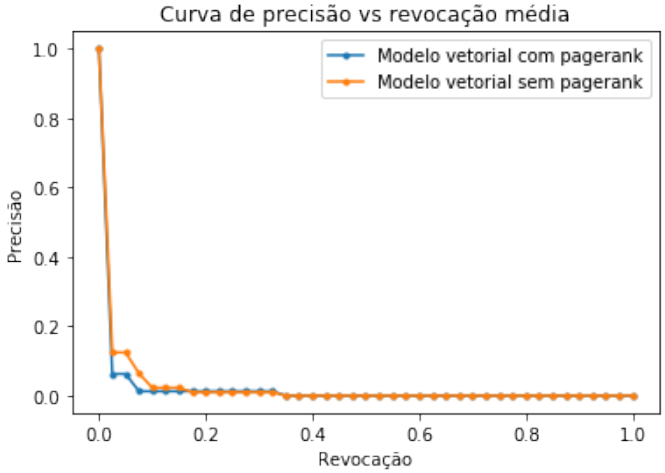

Figura 6. Curva de precisão x revocação

Com essa curva podemos ver que, no geral, o modelo vetorial sem o sinal de “pagerank” se saiu melhor, tendo uma precisão um pouco maior. Com tão poucos resultados conferidos fica inviável propor algo que explique tal situação. Agora mostraremos as curvas de precisão e revocação para as consultas em que pelo menos um dos resultados foi avaliado.

3.1. Consulta “espn”

Para a consulta “espn” um resultado foi avaliado para a versão sem “pagerank” e nenhuma para a outra versão. E o resultado foi considerado relevante, portanto a curva de precisão e revocação ficou dessa forma:

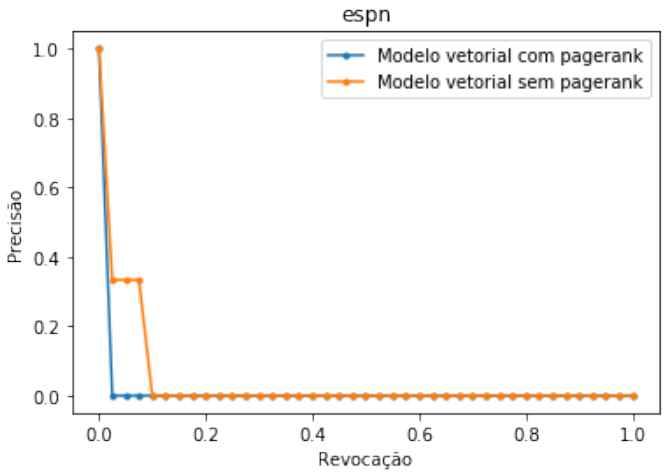

Figura 7. Curva de precisão x revocação

3.2. Consulta “cyberpunk 2077”

Para a consulta “cyberpunk 2077” dois resultados foram avaliados na versão sem utilizar “pagerank” e nenhuma para a outra versão. Um dos resultados foi considerado relevante, portanto a curva de precisão e revocação ficou dessa forma:

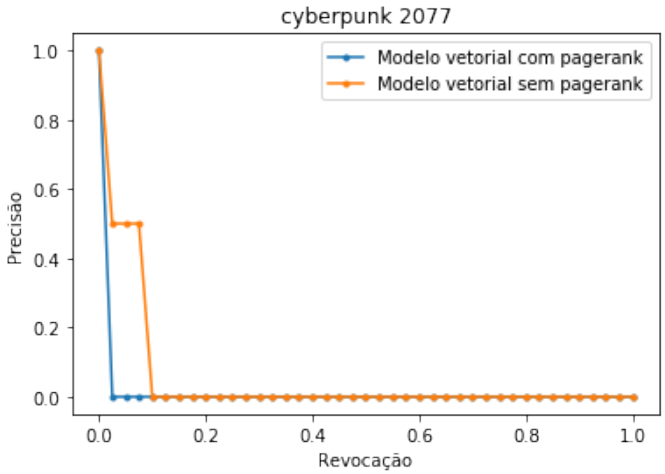

Figura 8. Curva de precisão x revocação

3.3. Consulta “sapiens: breve história humanidade”

Para a consulta “sapiens: breve história humanidade” um resultado foi avaliado para cada versão, mas nenhum foi considerado relevante, portanto a curva de precisão e revocação ficou dessa forma:

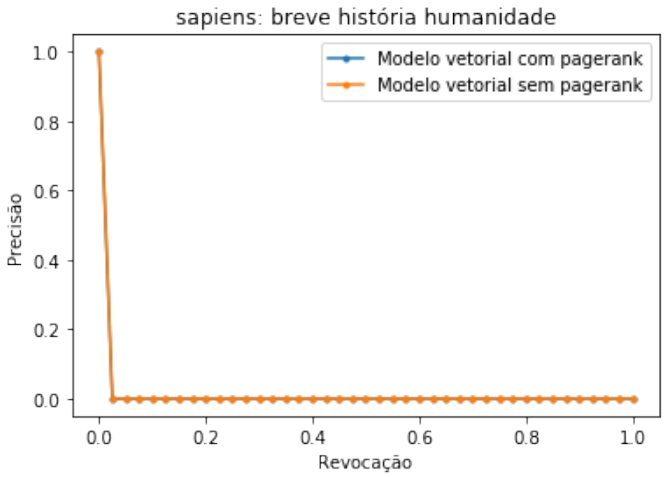

Figura 9. Curva de precisão x revocação

3.4. Consulta “o que é pix?”

Para a consulta “o que é pix?” um resultado foi avaliado para a versão sem “pagerank” e nenhuma para a outra versão. E o resultado foi considerado relevante, portanto a curva de precisão e revocação ficou dessa forma:

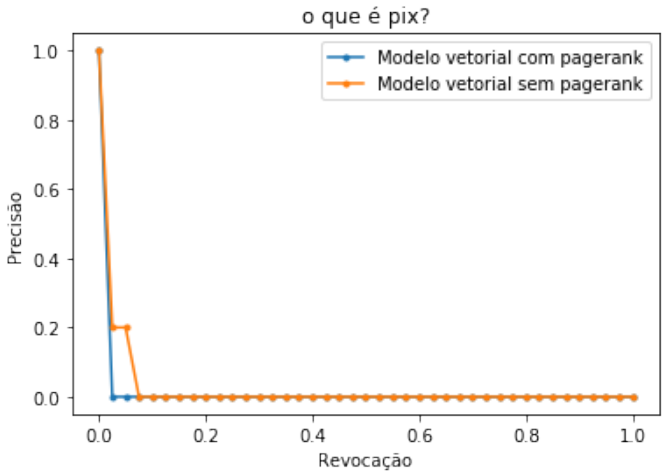

Figura 10. Curva de precisão x revocação

3.5. Consulta “pringles 118g”

Para a consulta “pringles 118g” não foi obtido nenhum resultado da máquina de buscas, portanto a consulta foi mudada apenas para “pringles”. Onde todos os resultados foram avaliados de ambas as versões. Mas apenas um resultado de cada versão foi considerado relevante, portanto a curva de precisão e revocação ficou dessa forma:

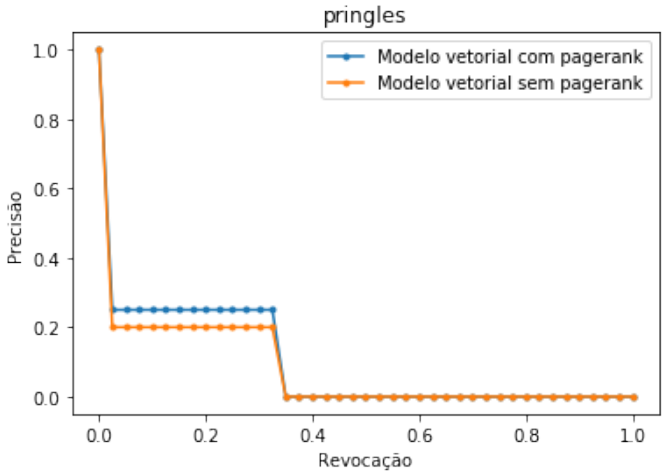

Figura 11. Curva de precisão x revocação

3.6. Consulta “auxílio emergencial”

Para a consulta “auxílio emergencial” um resultado foi avaliado para a versão com “pagerank” e nenhuma para a outra versão. E o resultado foi considerado irrelevante, portanto a curva de precisão e revocação ficou dessa forma:

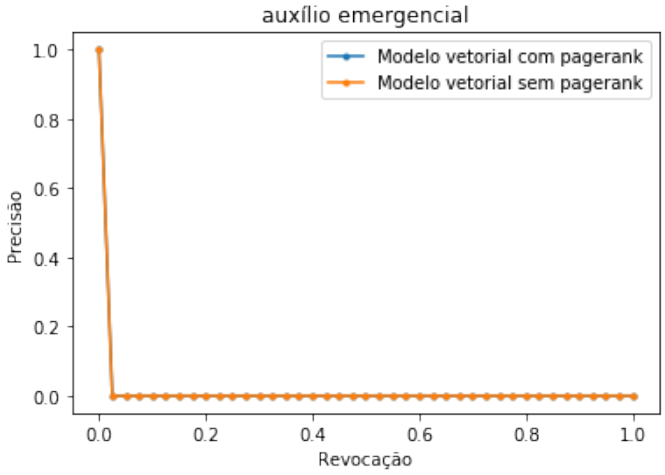

Figura 12. Curva de precisão x revocação

3.7. Consulta “samsung galaxy a21”

Para a consulta “samsung galaxy a21” um resultado foi avaliado para cada versão, e ambos foram considerados relevantes, portanto a curva de precisão e revocação ficou dessa forma:

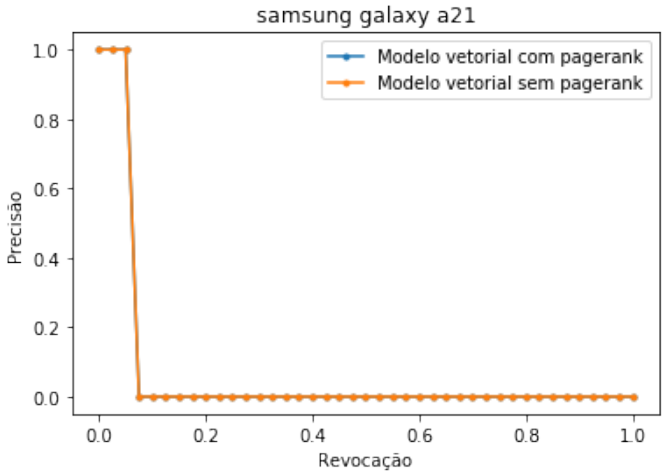

Figura 13. Curva de precisão x revocação

3.8. Consulta “receita sorvete de gelatina”

Para a consulta “receita sorvete de gelatina” um resultado foi avaliado para cada versão, mas nenhum resultado foi considerado relevante, portanto a curva de precisão e revocação ficou dessa forma:

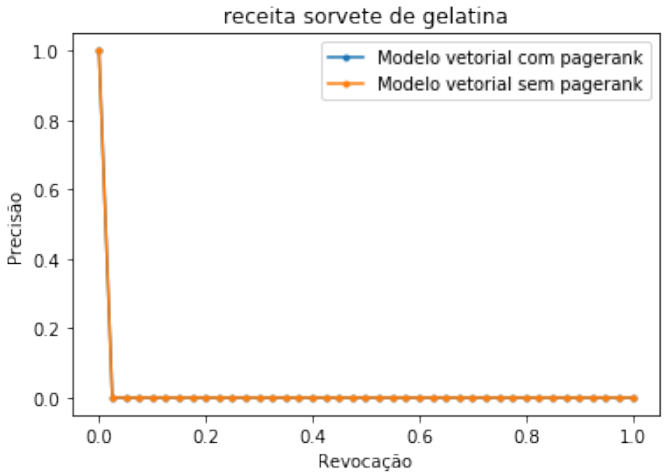

Figura 14. Curva de precisão x revocação

3.9. Consulta “santos dummont”

Para a consulta “santos dummont” um resultado foi avaliado para a versão sem “pagerank” e dois para a outra versão, e apenas o resultado da versão sem “pagerank” foi considerado relevante, portanto a curva de precisão e revocação ficou dessa forma:

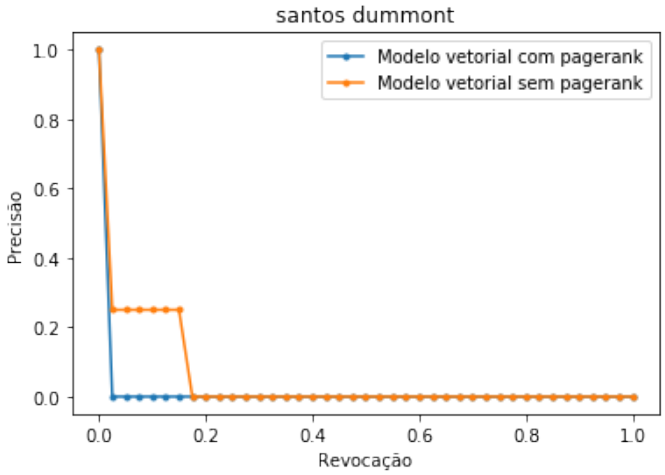

Figura 15. Curva de precisão x revocação

3.10. Consulta "vencedores do campeonato brasileiro"

Para a consulta "vencedores do campeonato brasileiro" dois resultados foram avaliados na versão sem "pagerank" e nenhuma para a outra versão. Mas nenhum dos resultados foi considerado relevante, portanto a curva de precisão e revocação ficou dessa forma:

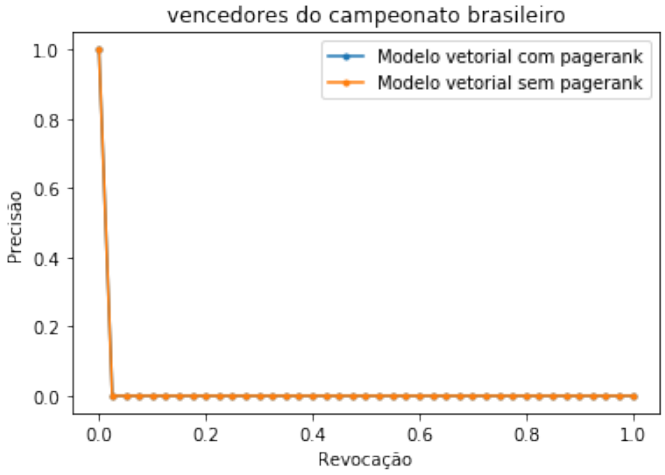

Figura 16. Curva de precisão x revocação

3.11. Consulta "como escrever escassez?"

Para a consulta "como escrever escassez?" um resultado foi avaliado para cada versão, mas nenhum foi considerado relevante, portanto a curva de precisão e revocação ficou dessa forma:

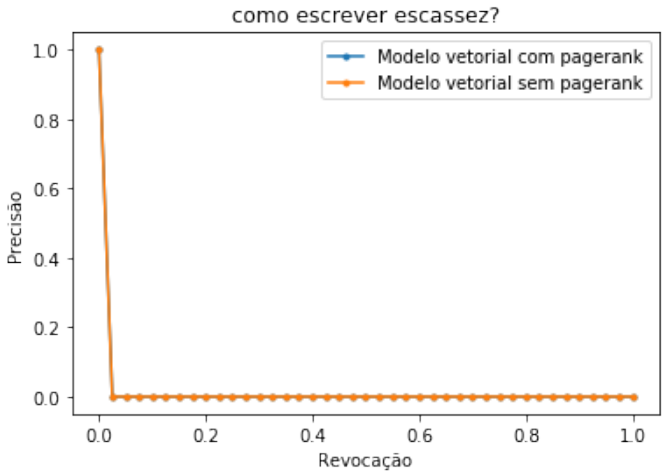

Figura 17. Curva de precisão x revocação

4. Conclusão

Como foi dito anteriormente, poucos resultados retornados foram avaliados pelos alunos, gerando resultados não muito conclusivos sobre a eficiência da máquina de buscas. Entretanto, foi possível perceber uma diferença nos números considerados relevantes para as consultas feitas na versão sem "pagerank", tendo proporcionalmente mais páginas avaliadas como relevantes, e também um maior número de páginas avaliadas.

Com tão poucos dados fica difícil apontar possíveis motivos para as diferenças citadas acima. Inicialmente podemos pensar que houve um grande peso para o sinal de "pagerank", mas é impossível afirmar qualquer coisa.