



Modelo BERT para Recuperação de informação em Referências bibliográficas

Workshop de Inteligência Artificial ICT-UNIFESP

Integrantes

MATHEUS RAMOS ESTEVES - 156.732

VINICIUS MATUMOTO DIOGO - 156.734



Sumário

- Introdução;
- Motivação;
- Metodologia experimental
 - Banco de Dados;
 - Tratamento de Dados;
 - Modelo BERT;
 - Recuperação de Informação;
 - Testes Realizados;
 - Métricas de Avaliação;
- Resultados Obtidos;
- Conclusão;
- Trabalhos Futuros;
- Referências.

Motivação

- Pedido do Dr. Thiago Macedo
- Conversas com o Dr. Fabio Faria

Search term(s)

Title
Add another term+
Search

Subject

All classifications will be included by default.

Computer Science (cs) Physi [1] [J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," 2018, *arXiv:1801.06146*.](#)

Economics (econ) Quan [2] [R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. 25th Int. Conf. Mach. Learn.*, Jul. 2008, pp. 160–167.](#)

Electrical Engineering and Systems Science (eess) Quan [3] [M. F. Kabir, K. Abdullah-Al-Mamun, and M. N. Huda, "Deep learning-based parts of speech tagger for Bengali," in *Proc. 5th Int. Conf. Inform., Electron. Vis. \(ICIEV\)*, May 2016, pp. 26–29.](#)

Mathematics (math) Statis [4] [J. Li, R. Li, and E. Hovy, "Recursive deep models for discourse parsing," in *Proc. Conf. Empirical Methods Natural Lang. Process. \(EMNLP\)*, Oct. 2014, pp. 2061–2069.](#)

Include cross-listed papers Exclude cross-listed p

[1] [J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," 2018, *arXiv:1801.06146*.](#)

[2] [R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. 25th Int. Conf. Mach. Learn.*, Jul. 2008, pp. 160–167.](#)

[3] [M. F. Kabir, K. Abdullah-Al-Mamun, and M. N. Huda, "Deep learning-based parts of speech tagger for Bengali," in *Proc. 5th Int. Conf. Inform., Electron. Vis. \(ICIEV\)*, May 2016, pp. 26–29.](#)

[4] [J. Li, R. Li, and E. Hovy, "Recursive deep models for discourse parsing," in *Proc. Conf. Empirical Methods Natural Lang. Process. \(EMNLP\)*, Oct. 2014, pp. 2061–2069.](#)

[5] [L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *WIREs Data Mining Knowl. Discovery*, vol. 8, no. 4, Jul. 2018, Art. no. e1253.](#)

[6] [Z. A. Merrouni, B. Frikh, and B. Ouhbi, "Automatic keyphrase extraction: A survey and trends," *J. Intell. Inf. Syst.*, vol. 54, pp. 391–424, May 2019.](#)

[7] [K. M. Hammouda, D. N. Matute, and M. S. Kamel, "Corephrase: Keyphrase extraction for document clustering," in *Proc. Int. Workshop Mach. Learn. Data Mining Pattern Recognit.* Berlin, Germany: Springer, Jul. 2005, pp. 265–274.](#)

[8] [P. Sharma and Y. Li, "Self-supervised contextual keyword and keyphrase retrieval with self-labelling," *Preprints*, 2019, doi: 10.20944/preprints201908.0073.v1.](#)

Metodologia experimental

A. Banco de Dados

- Web Scraping usando Selenium e HTTP
- BibTex
- Mais de 52 mil referencias

	A	B	C	D
1	referencia	autor	titulo	ano
2	Sheldon Katz. Rational curves on calabi-yau threefolds. 1992. arXiv: arXiv:alg-geom/9202010	Katz, Sheldon	Rational Curves on Calabi-Yau Threefolds	1992
3	Sheldon Katz and David R. Morrison. Gorenstein threefold singularities. 1992. arXiv: arXiv:alg-geom/9202011	Katz, Sheldon and Morrison, David R.	Gorenstein Threefold Singularities with Small Resolution	1992
4	Yun-Gang Ye. Complex contact threefolds and their contact curves. 1992. arXiv: arXiv:alg-geom/9202012	Ye, Yun-Gang	Complex Contact Threefolds and Their Contact Curves	1992
5	Sheldon Katz. Rational curves on calabi-yau threefolds. 1992. arXiv: arXiv:alg-geom/9202013	Katz, Sheldon	Rational Curves on Calabi-Yau Threefolds	1992
6	Sheldon Katz and David R. Morrison. Gorenstein threefold singularities. 1992. arXiv: arXiv:alg-geom/9202014	Katz, Sheldon and Morrison, David R.	Gorenstein Threefold Singularities with Small Resolution	1992
7	Yun-Gang Ye. Complex contact threefolds and their contact curves. 1992. arXiv: arXiv:alg-geom/9202015	Ye, Yun-Gang	Complex Contact Threefolds and Their Contact Curves	1992
8	David R. Morrison. Mirror symmetry and rational curves on quintic hypersurfaces. 1992. arXiv: arXiv:alg-geom/9202016	Morrison, David R.	Mirror symmetry and rational curves on quintic threefolds	1992
9	Günter M. Ziegler. On the difference between real and complex arrangements. 1992. arXiv: arXiv:alg-geom/9202017	Ziegler, Günter M.	On the difference between real and complex arrangements	1992
10	Claude LeBrun and Yat-Sun Poon. Twistor theory, kaehler manifolds, and bimeromorphic geometry. 1992. arXiv: arXiv:alg-geom/9202018	LeBrun, Claude and Poon, Yat-Sun	Twistors, Kaehler Manifolds, and Bimeromorphic Geometry	1992
11	Tadao Oda. The algebraic de rham theorem for toric varieties. 1992. arXiv: arXiv:alg-geom/9202019	Oda, Tadao	The algebraic de Rham theorem for toric varieties	1992
12	Walter D. Neumann and Le Van Thanh. On irregular links at infinity. 1992. arXiv: arXiv:alg-geom/9202020	Neumann, Walter D. and Thanh, Le Van	On irregular links at infinity of algebraic plane curves	1992
13	V. B. Mehta and Wilberd van der Kallen. On a grauert-riemenschneider vanishing theorem for toric varieties. 1992. arXiv: arXiv:alg-geom/9202021	Mehta, V. B. and van der Kallen, Wilberd	On a Grauert-Riemenschneider vanishing theorem for toric varieties	1992
14	John B. Little. Another relation between approaches to the schottky problem. 1992. arXiv: arXiv:alg-geom/9202022	Little, John B.	Another Relation Between Approaches to the Schottky Problem	1992
15	Peter F. Stiller. The geometry and arithmetic of elliptic surfaces. 1992. arXiv: arXiv:alg-geom/9202023	Stiller, Peter F.	The Geometry and Arithmetic of Elliptic Surfaces	1992
16	George R. Kempf. Cohomology of coherent sheaves on a proper scheme. 1992. arXiv: arXiv:alg-geom/9202024	Kempf, George R.	Cohomology of coherent sheaves on a proper scheme	1992
17	George R. Kempf. Deformations of semi-euler characteristics. 1992. arXiv: arXiv:alg-geom/9202025	Kempf, George R.	Deformations of Semi-Euler Characteristics	1992
18	Martin Schlichenmaier. Degenerations of generalized klnn algebras. 1992. arXiv: arXiv:alg-geom/9202026	Schlichenmaier, Martin	Degenerations of generalized KNN algebras on tori	1992
19	Elisabetta Colombo and Bert van Geemen. Note on curves in a jacobian. 1992. arXiv: arXiv:alg-geom/9202027	Colombo, Elisabetta and van Geemen, Bert	Note on curves in a Jacobian	1992
20	G. Mikhalkin. Congruences for real algebraic curves on an ellipsoid. 1992. arXiv: arXiv:alg-geom/9202028	Mikhalkin, G.	Congruences for real algebraic curves on an ellipsoid	1992
21	G. Mikhalkin. Extensions of rokhlin congruence for curves on surfaces. 1992. arXiv: arXiv:alg-geom/9202029	Mikhalkin, G.	Extensions of Rokhlin congruence for curves on surfaces	1992
22	Sheldon Katz. Arithmetically cohen-macaulay curves cut out by quadrics. 1992. arXiv: arXiv:alg-geom/9202030	Katz, Sheldon	Arithmetically Cohen-Macaulay Curves cut out by Quadrics	1992
23	Frank DeMeyer, Tim Ford, and Rick Miranda. The cohomological brauer group of a toric variety. 1992. arXiv: arXiv:alg-geom/9202031	DeMeyer, Frank and Ford, Tim	The cohomological Brauer group of a toric variety	1992
24	G. Niesi and L. Robbiano. Disproving hibi's conjecture with coocoa or projective Groebner bases. 1992. arXiv: arXiv:alg-geom/9202032	Niesi, G. and Robbiano, L.	Disproving Hibi's Conjecture with CoCoA or Projective Groebner bases	1992
25	Dave Bayer, Andre Galligo, and Mike Stillman. Grobner bases and extension of scalars. 1992. arXiv: arXiv:alg-geom/9202033	Bayer, Dave and Galligo, Andre and Stillman, Mike	Grobner bases and extension of scalars	1992
26	Richard Hain. Classical polylogarithms. 1992. arXiv: arXiv:alg-geom/9202034	Hain, Richard	Classical Polylogarithms	1992
27	Johan Dupont, Richard Hain, and Steven Zucker. Regulators and characteristic classes of flat bundles. 2000. arXiv: arXiv:alg-geom/9202035	Dupont, Johan and Hain, Richard and Zucker, Steven	Regulators and characteristic classes of flat bundles	2000

Metodologia experimental

B. Tratamento de Dados

- Remoção de caracteres especiais

	referencia	autor	titulo	ano
0	Sheldon Katz Rational curves on calabi-yau thr...	Katz Sheldon	Rational Curves on Calabi-Yau Threefolds	1992
1	Sheldon Katz and David R Morrison Gorenstein t... Katz Sheldon and Morrison David R	Gorenstein Threafold Singularities with Small ...	1992	
2	Yun-Gang Ye Complex contact threefolds and the...	Ye Yun-Gang	Complex Contact Threefolds and Their Contact C...	1992
3	Sheldon Katz Rational curves on calabi-yau thr...	Katz Sheldon	Rational Curves on Calabi-Yau Threefolds	1992
4	Sheldon Katz and David R Morrison Gorenstein t... Katz Sheldon and Morrison David R	Gorenstein Threafold Singularities with Small ...	1992	

Metodologia experimental

C. Modelo BERT

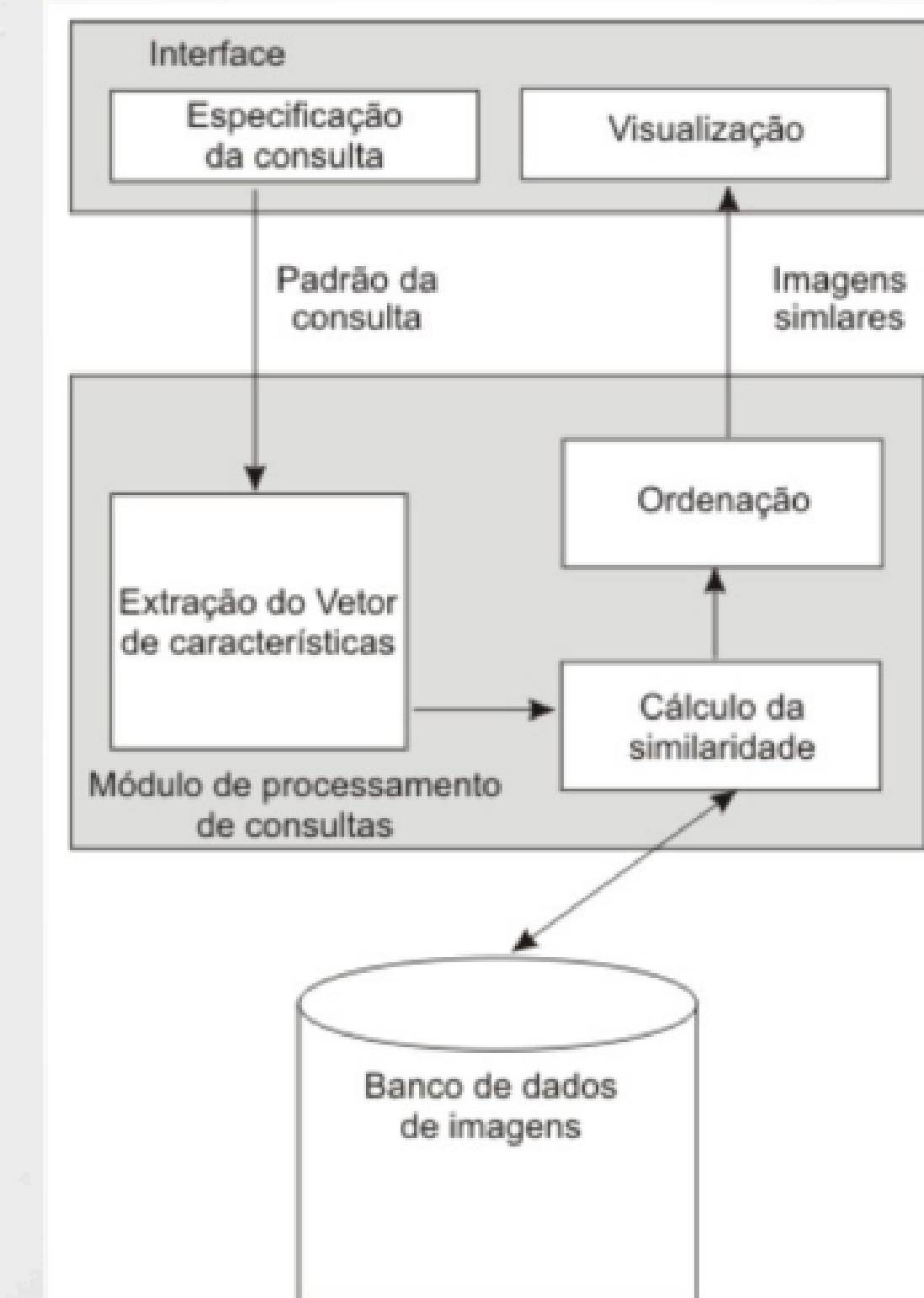
- Google
- Processamento de Linguagens Naturais
- Transformers
- Pré-treinamento com bert-base-uncased



Metodologia experimental

D. Recuperação de Informação

- Buscar, Localizar e extrair informações
- Encontrar referências bibliográficas
- Tokenização e transformação em embeddings das referências



Metodologia experimental

E. Testes Realizados

1. **Remoção de palavras:** palavras foram removidas de forma aleatória da referência de entrada;
2. **Remoção de caracteres:** caracteres foram removidos de forma aleatória da referência de entrada;
3. **Substituição de caracteres:** caracteres foram substituídos por outras letras, também de forma aleatória na referência de entrada.

Todos testes foram gradualmente aumentando de uma modificação até sete modificações na mesma referência.

F. Métricas de Avaliação

Similaridade Cosseno

$$\frac{A \cdot B}{\| A \| \| B \|}$$

- A e B - vetores de *embeddings* das referências;



$$\frac{1}{N} \sum_{q=1}^N Average\ precision \left(\frac{1}{q} \right)$$

MAP

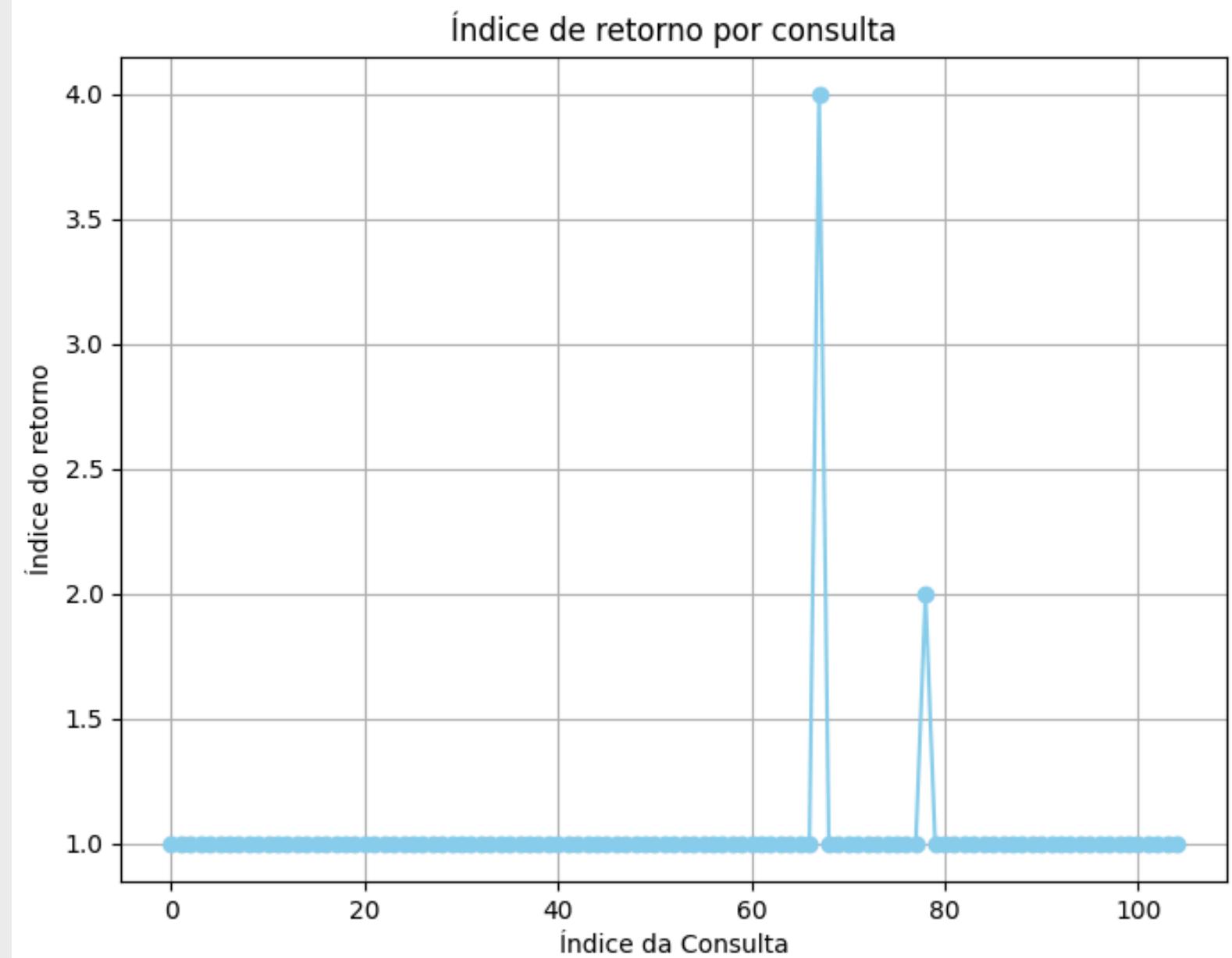
- N - Quantidade total de consultas;
- q - Número da consulta;

Resultados obtidos

Remoção de palavras

Uma palavra

Foram recuperadas dentro do top 3:
104 das 105 consultas

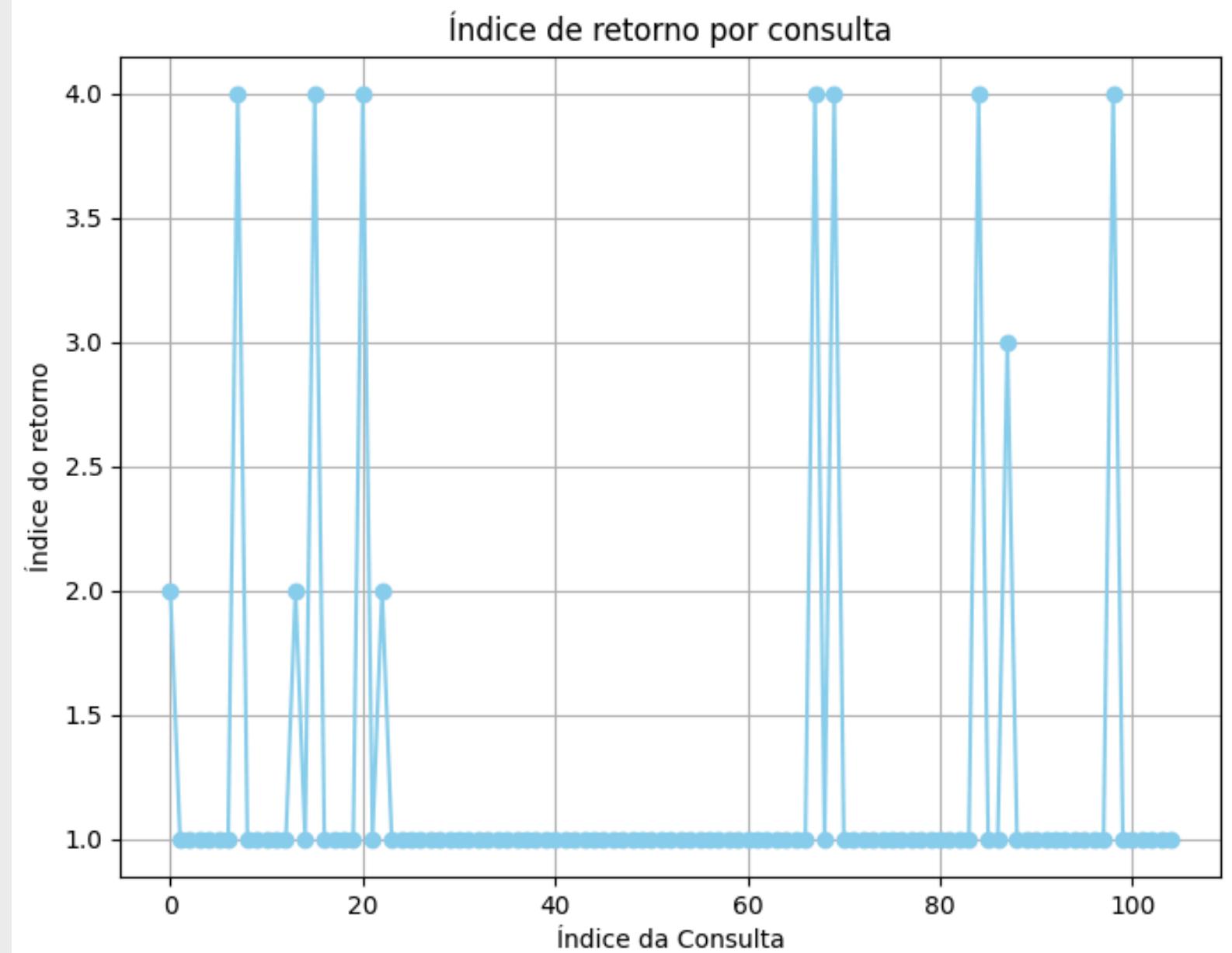


Resultados obtidos

Remoção de palavras

Duas palavras

Foram recuperadas dentro do top 3:
98 das 105 consultas

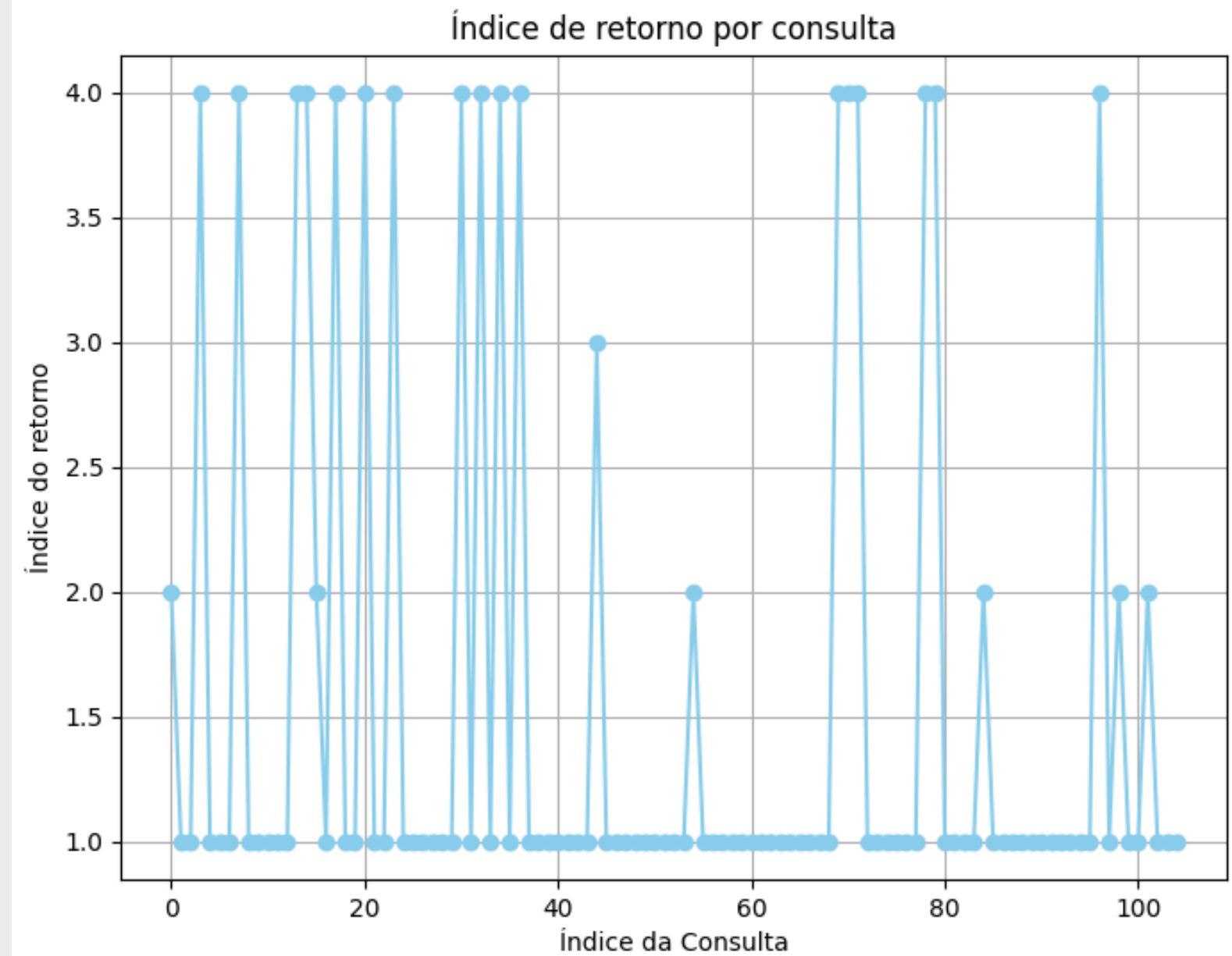


Resultados obtidos

Remoção de palavras

Três palavras

Foram recuperadas dentro do top 3:
88 das 105 consultas

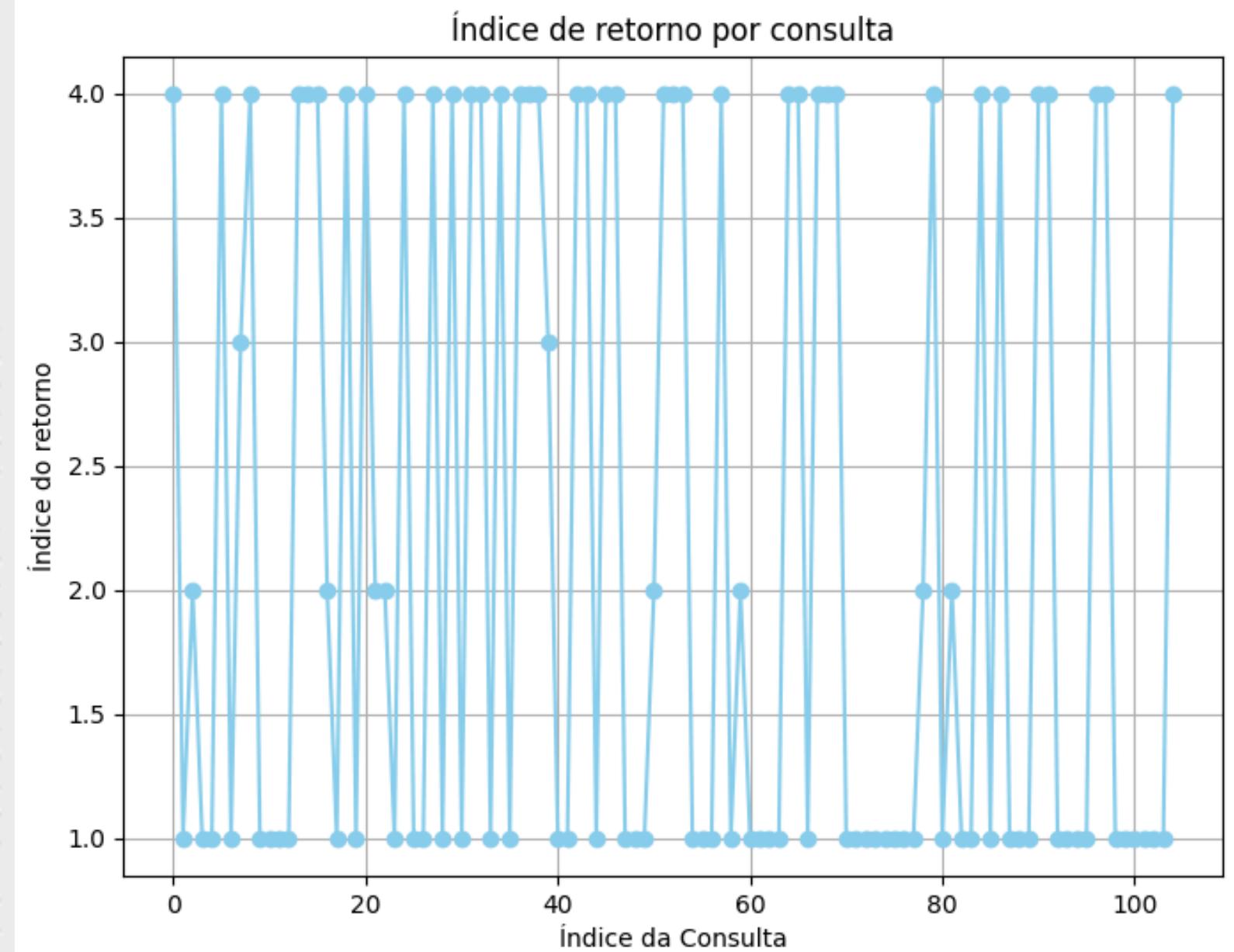


Resultados obtidos

Remoção de palavras

Quatro palavras

Foram recuperadas dentro do top 3:
67 das 105 consultas

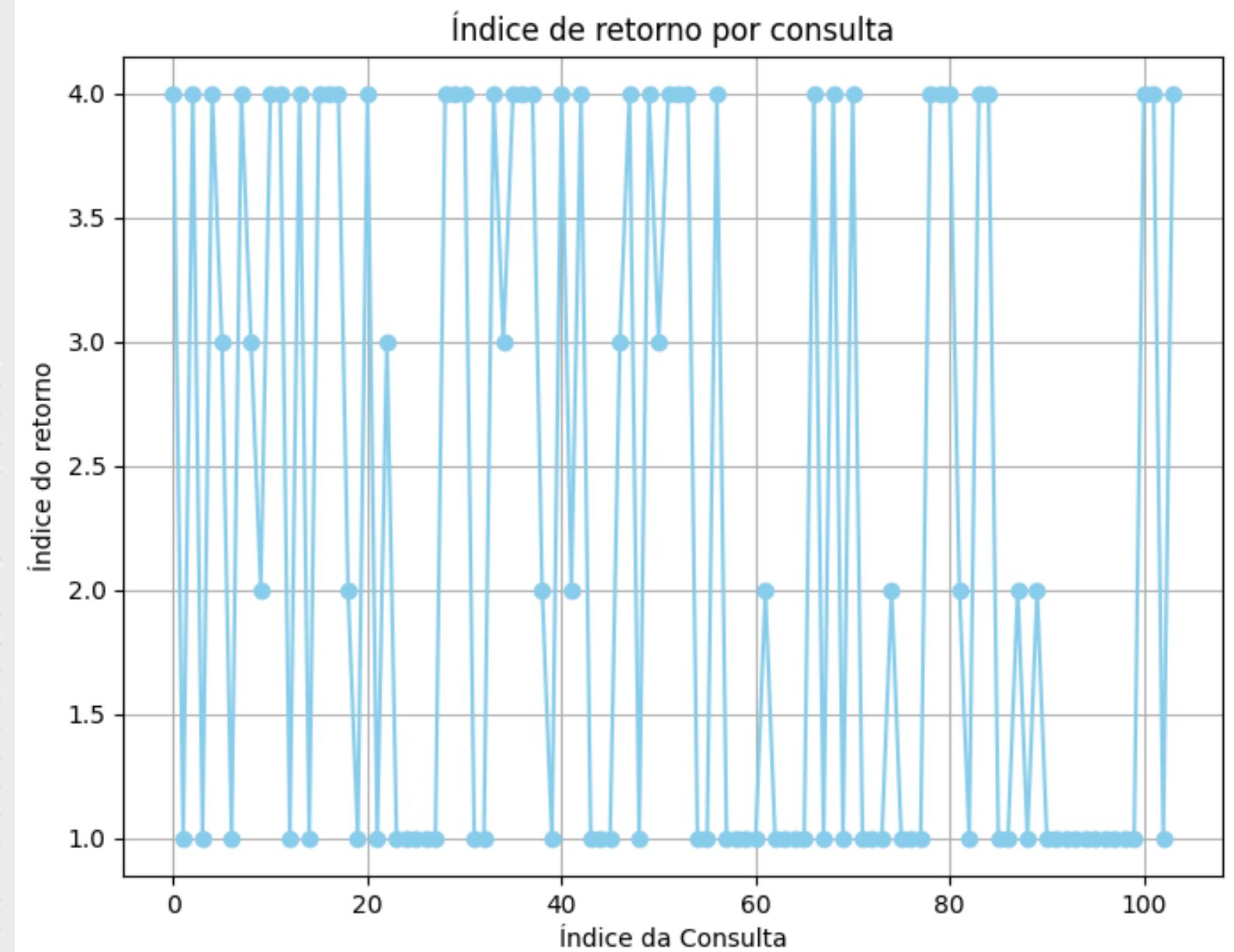


Resultados obtidos

Remoção de palavras

Cinco palavras

Foram recuperadas dentro do top 3:
67 das 105 consultas

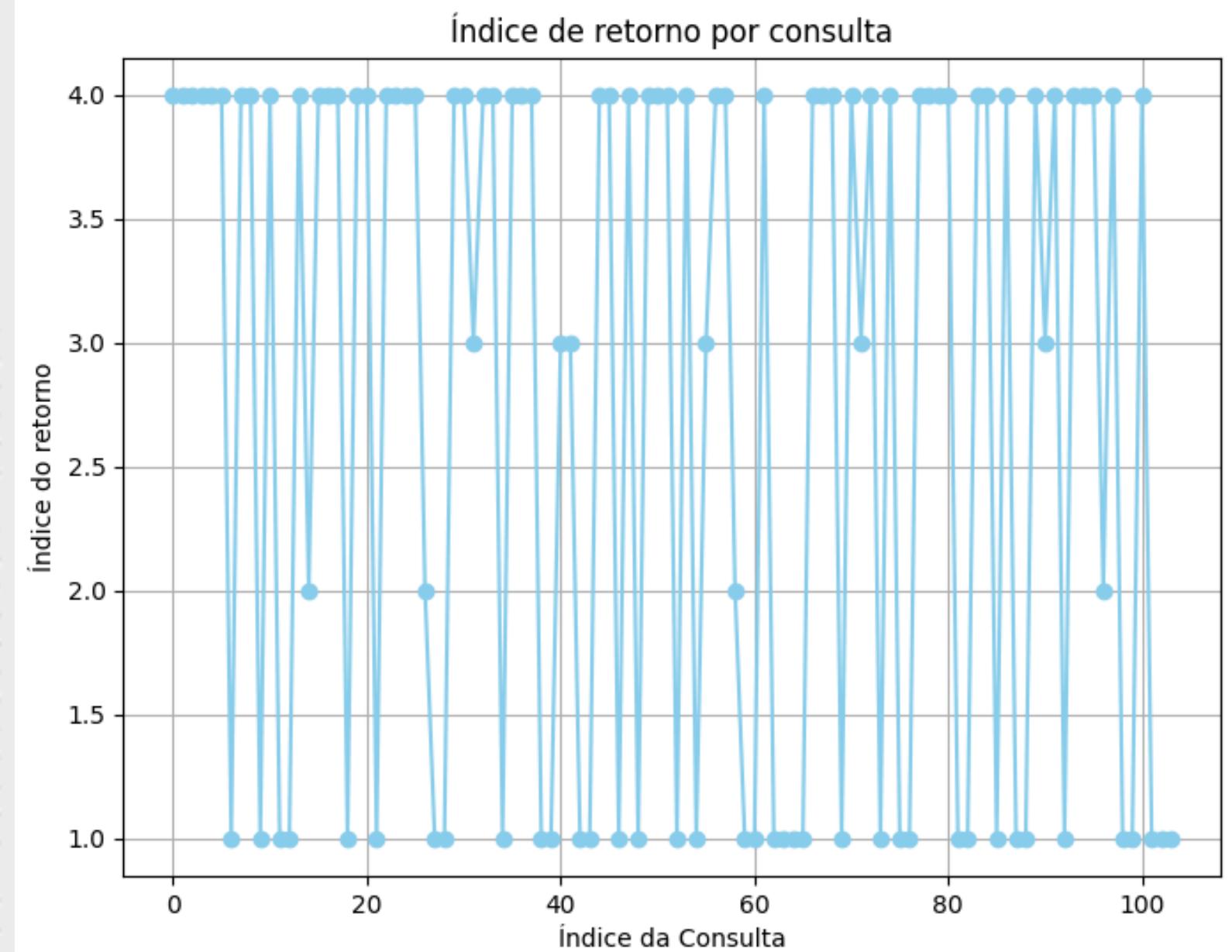


Resultados obtidos

Remoção de palavras

Seis palavras

Foram recuperadas dentro do top 3:
48 das 105 consultas

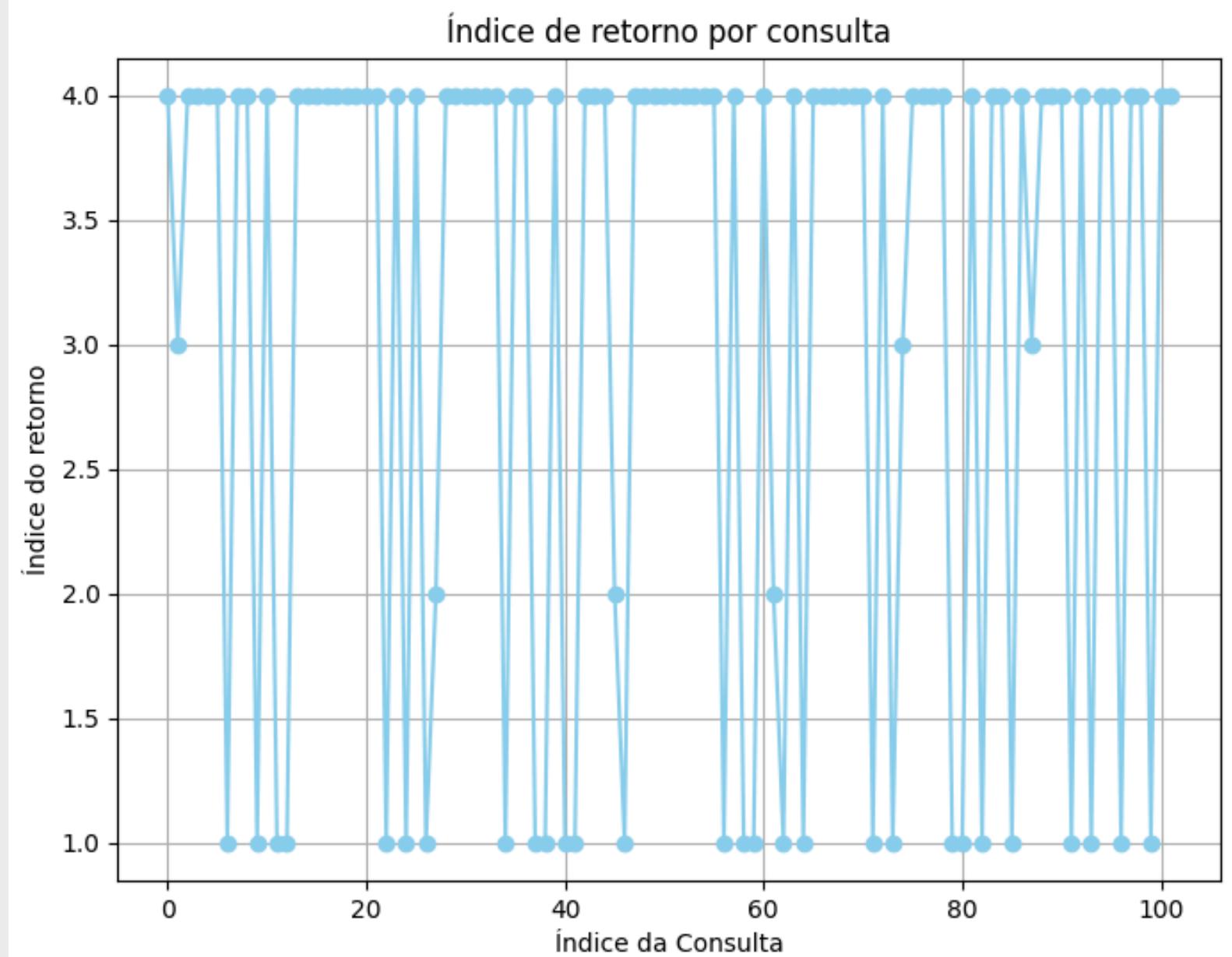


Resultados obtidos

Remoção de palavras

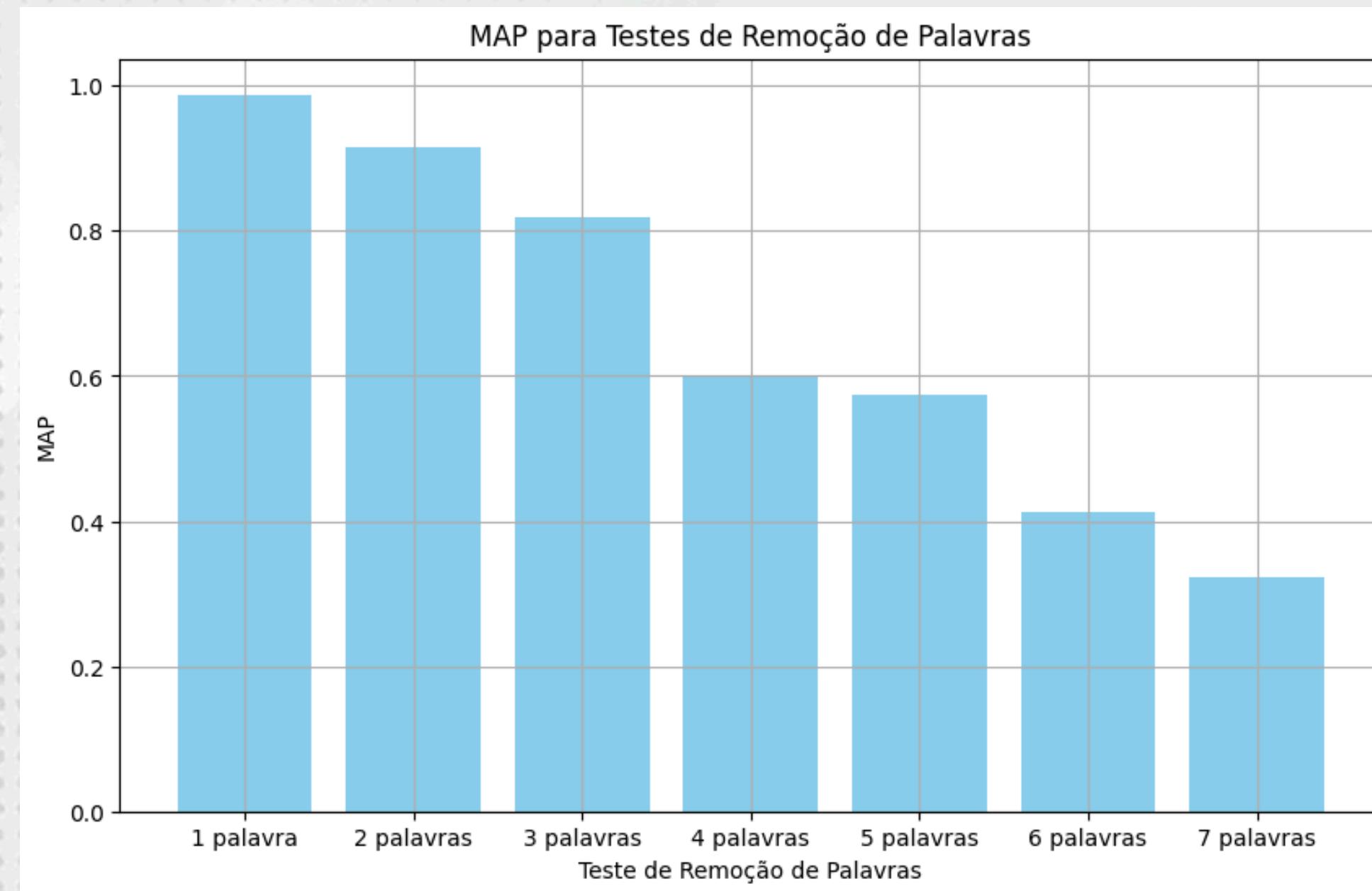
Sete palavras

Foram recuperadas dentro do top 3:
34 das 105 consultas



Resultados obtidos

Remoção de palavras

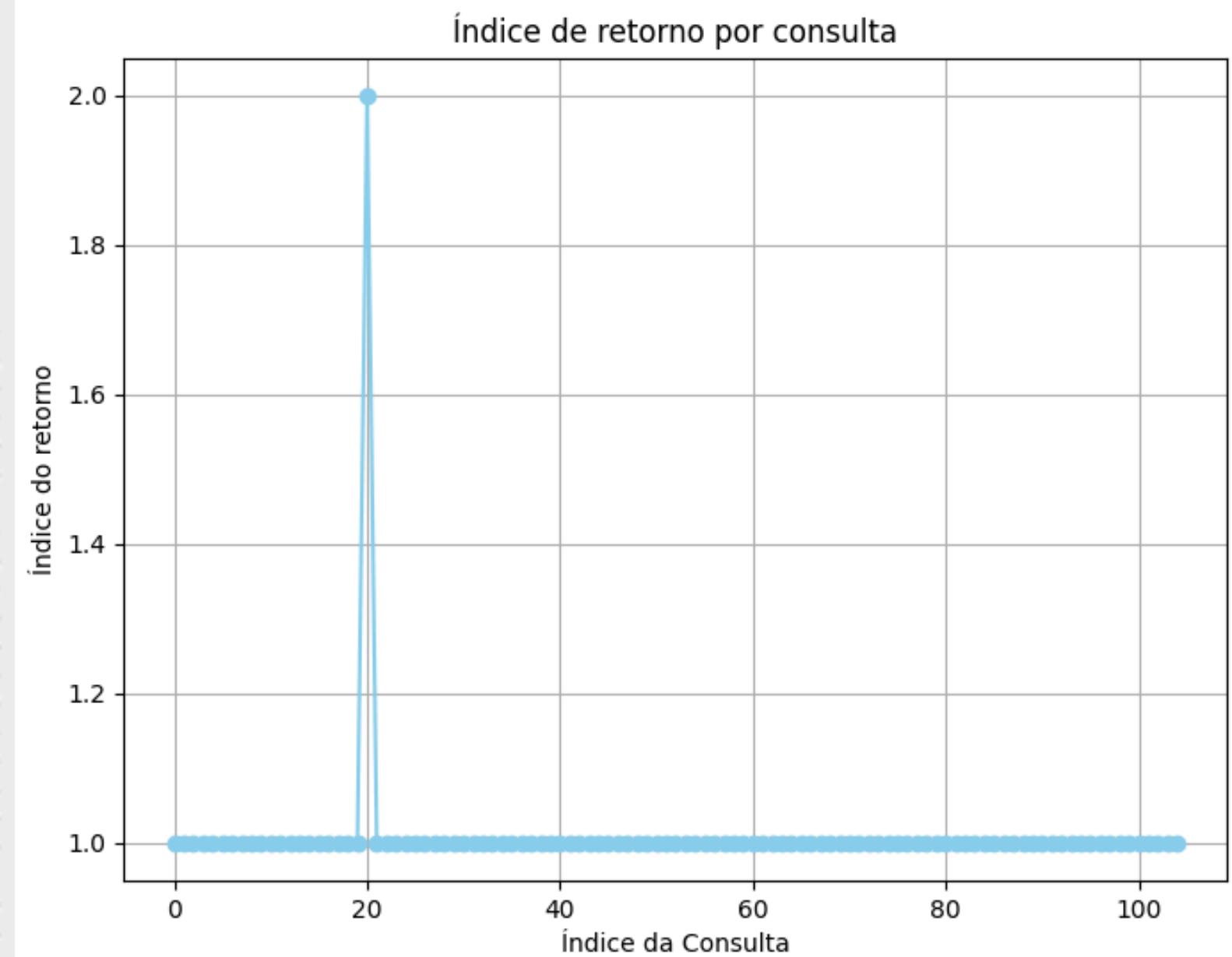


Resultados obtidos

Remoção de caracteres

Um caractere

Foram recuperadas dentro do top 3:
105 das 105 consultas

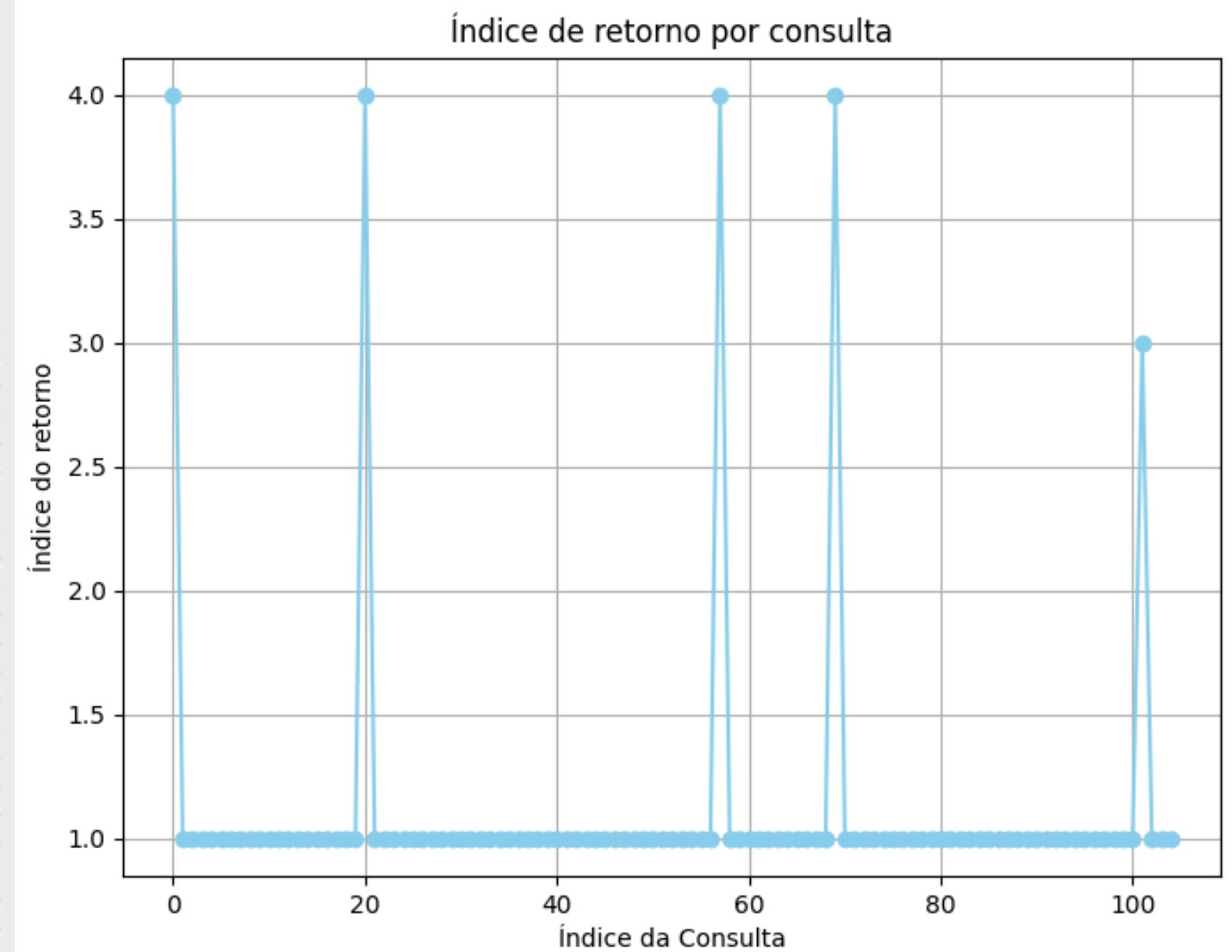


Resultados obtidos

Remoção de caracteres

Dois caracteres

Foram recuperadas dentro do top 3:
101 das 105 consultas

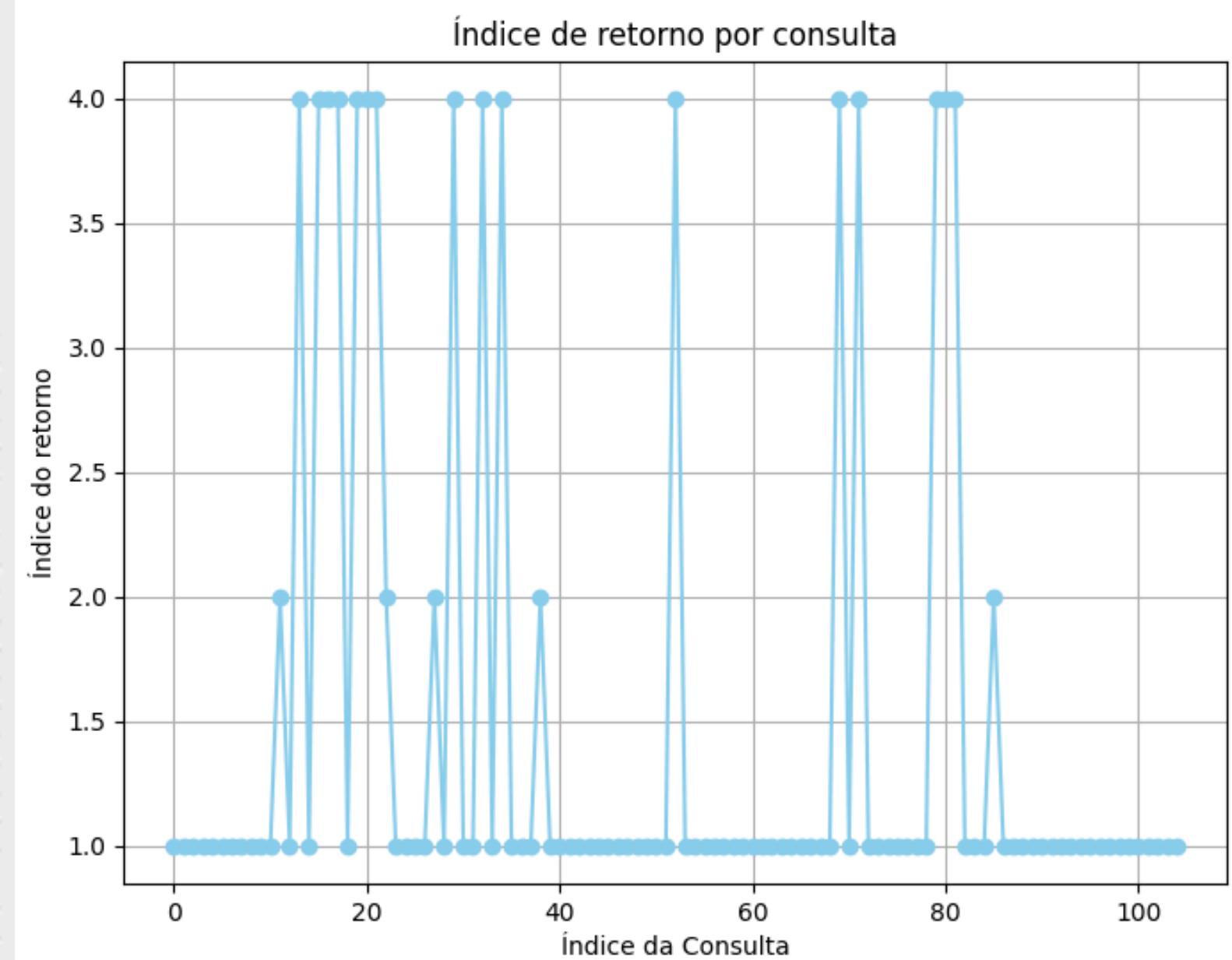


Resultados obtidos

Remoção de caracteres

Três caracteres

Foram recuperadas dentro do top 3:
89 das 105 consultas

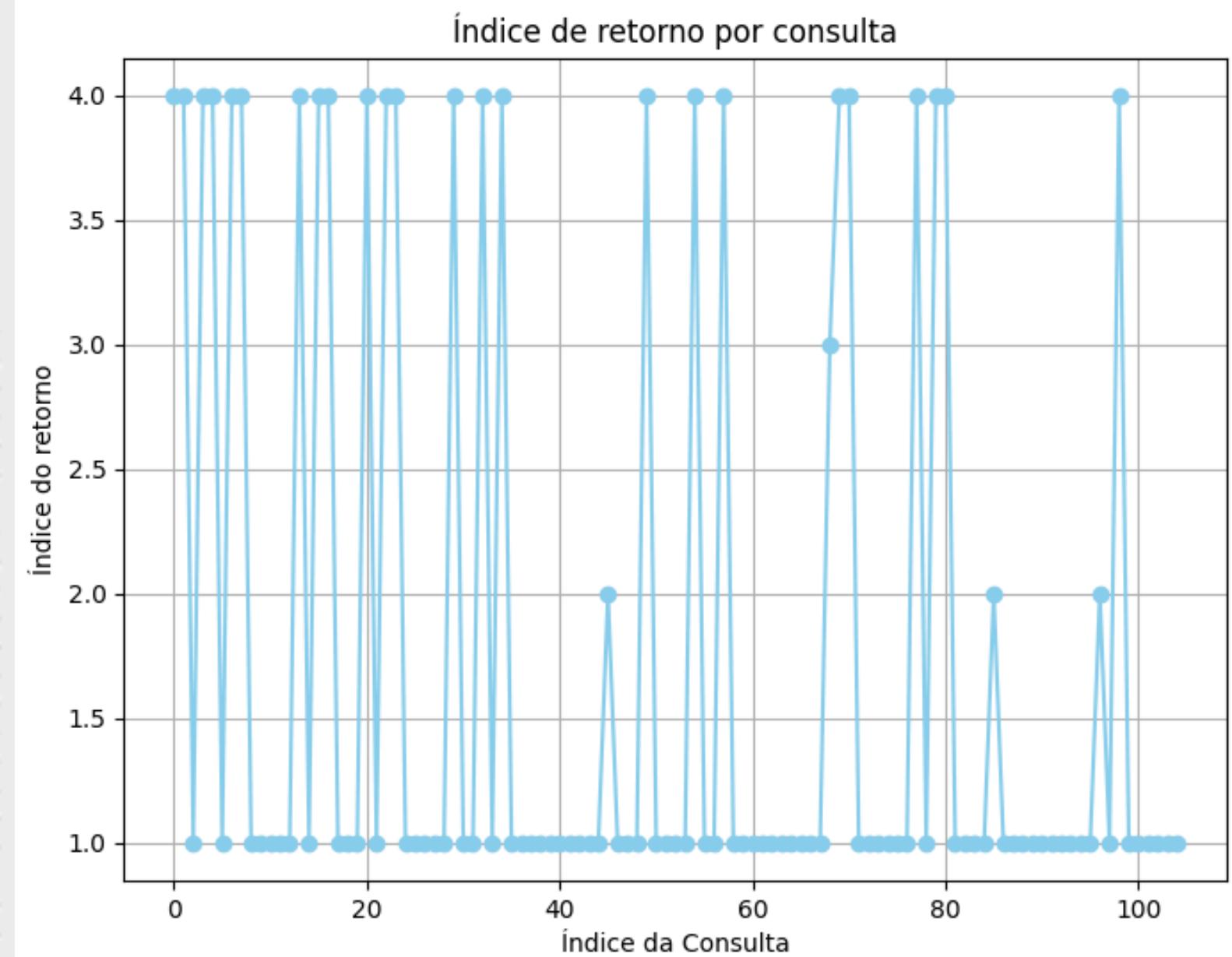


Resultados obtidos

Remoção de caracteres

Quatro caracteres

Foram recuperadas dentro do top 3:
81 das 105 consultas

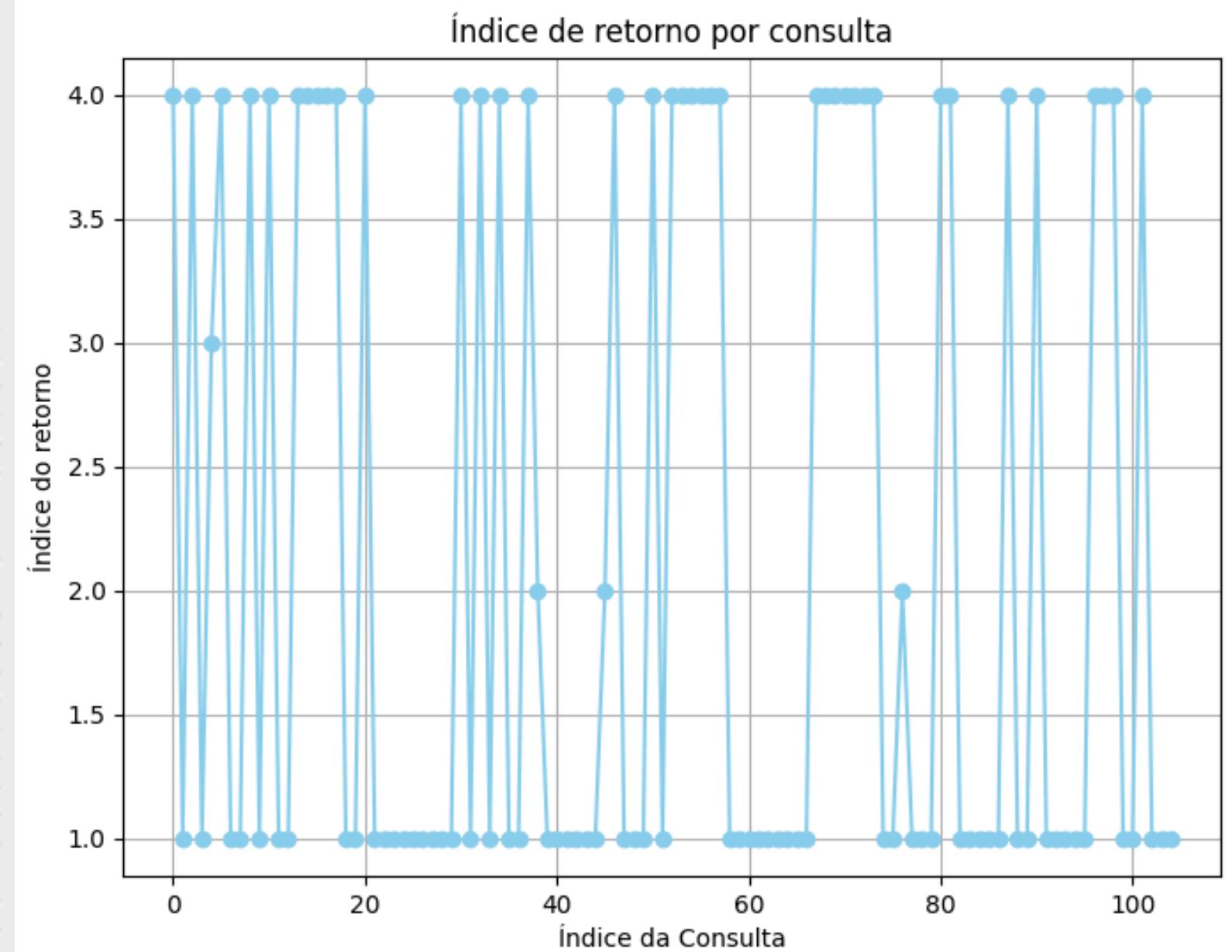


Resultados obtidos

Remoção de caracteres

Cinco caracteres

Foram recuperadas dentro do top 3:
67 das 105 consultas

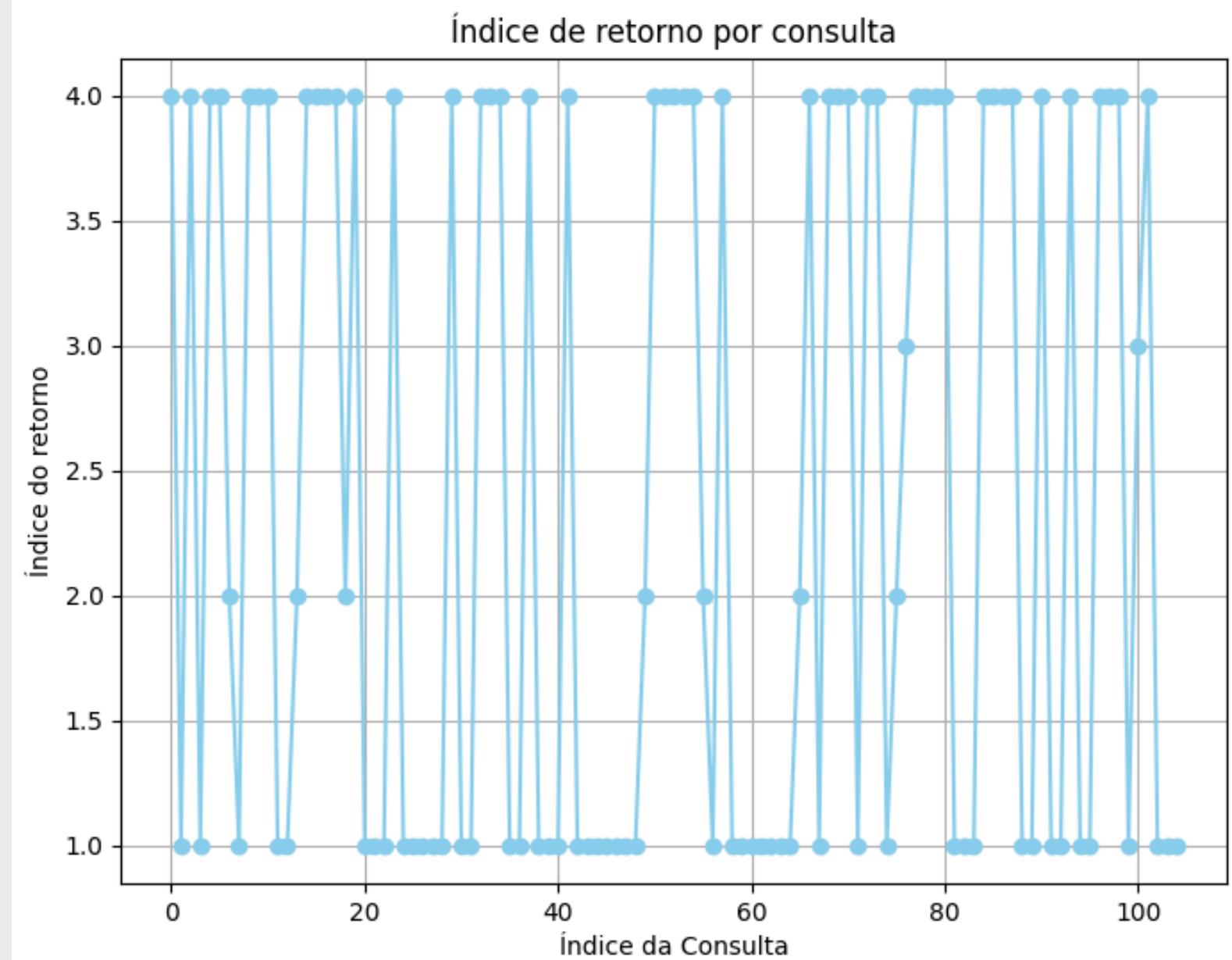


Resultados obtidos

Remoção de caracteres

Seis caracteres

Foram recuperadas dentro do top 3:
60 das 105 consultas

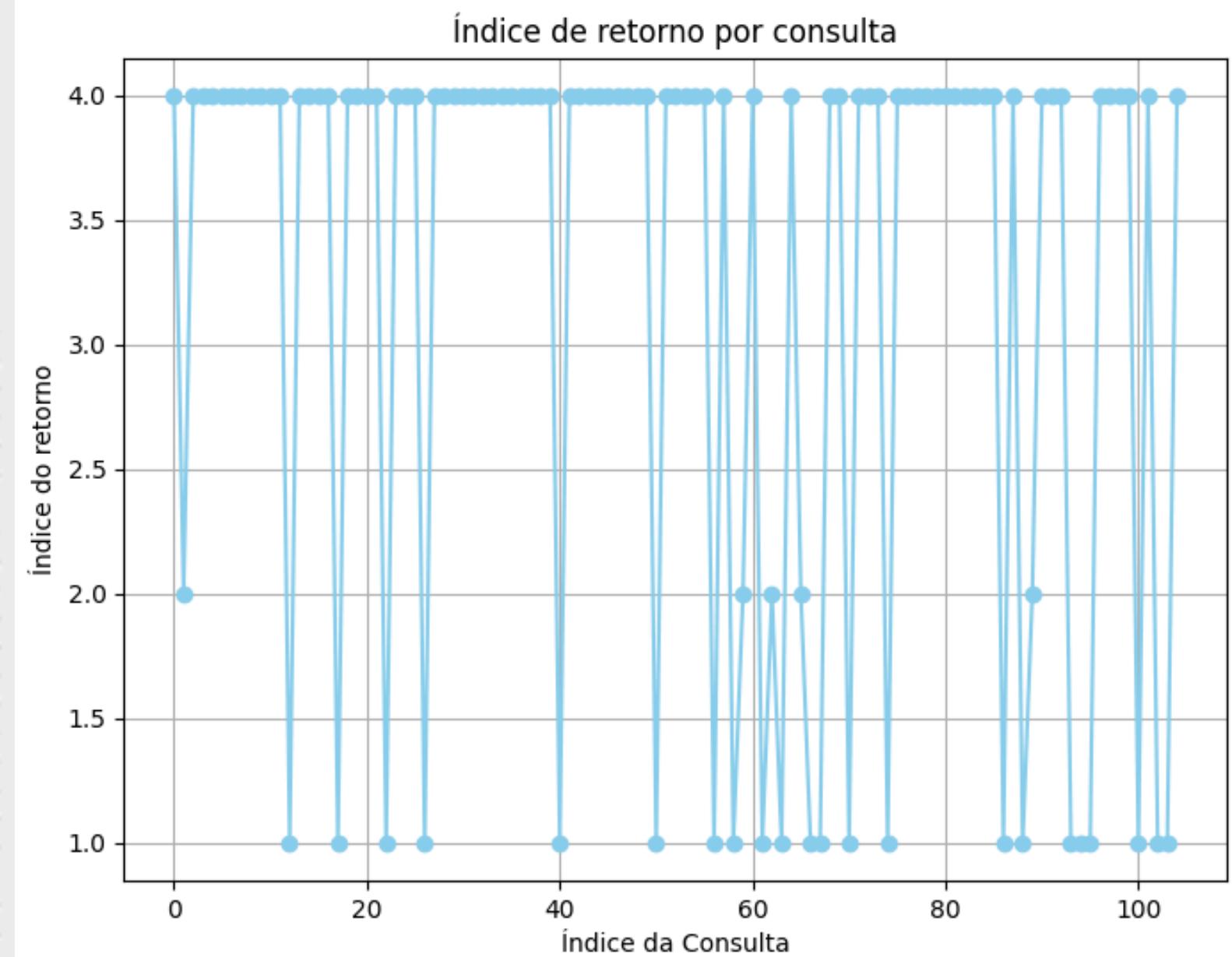


Resultados obtidos

Remoção de caracteres

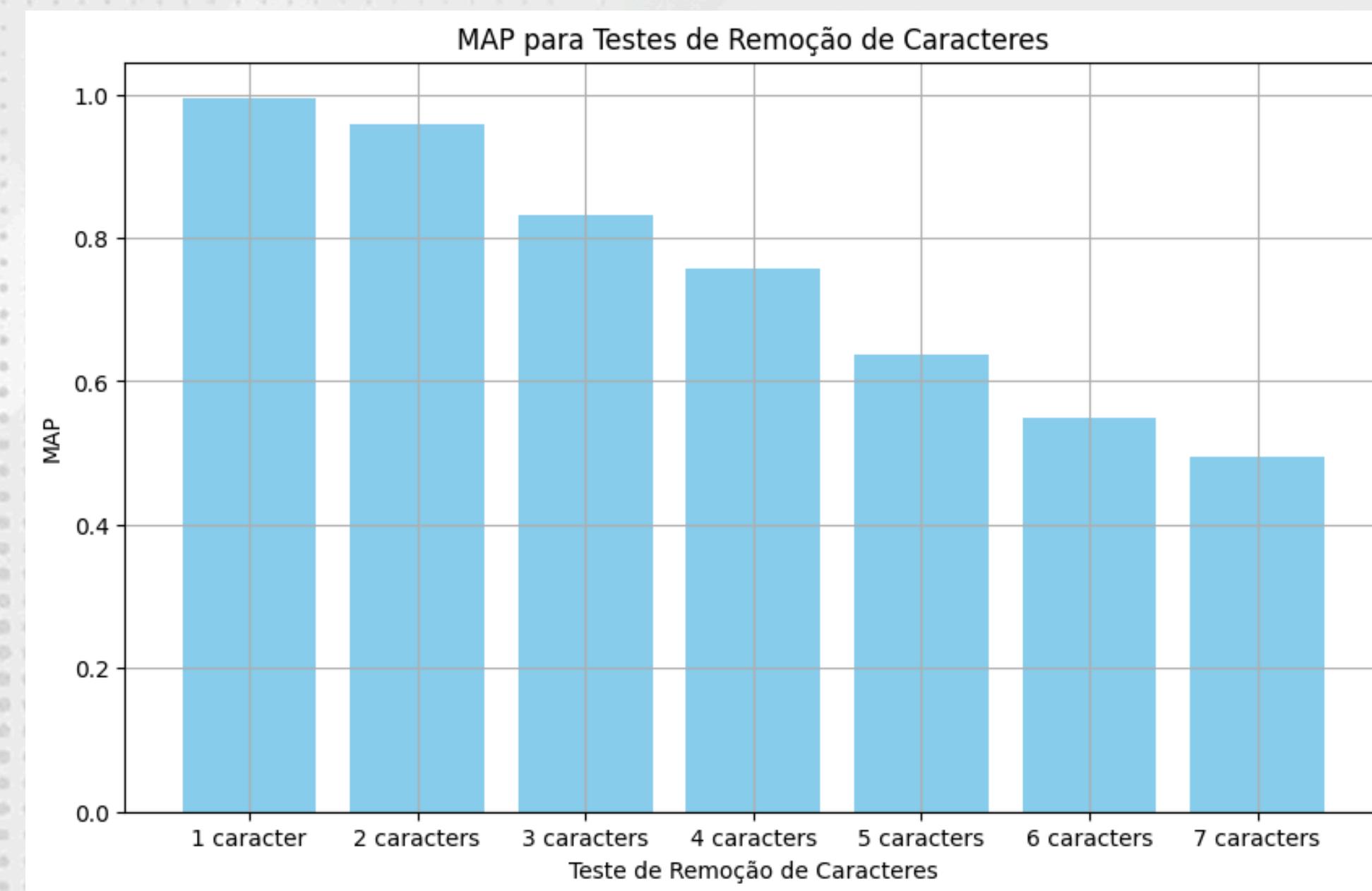
Sete caracteres

Foram recuperadas dentro do top 3:
27 das 105 consultas



Resultados obtidos

Remoção de caracteres

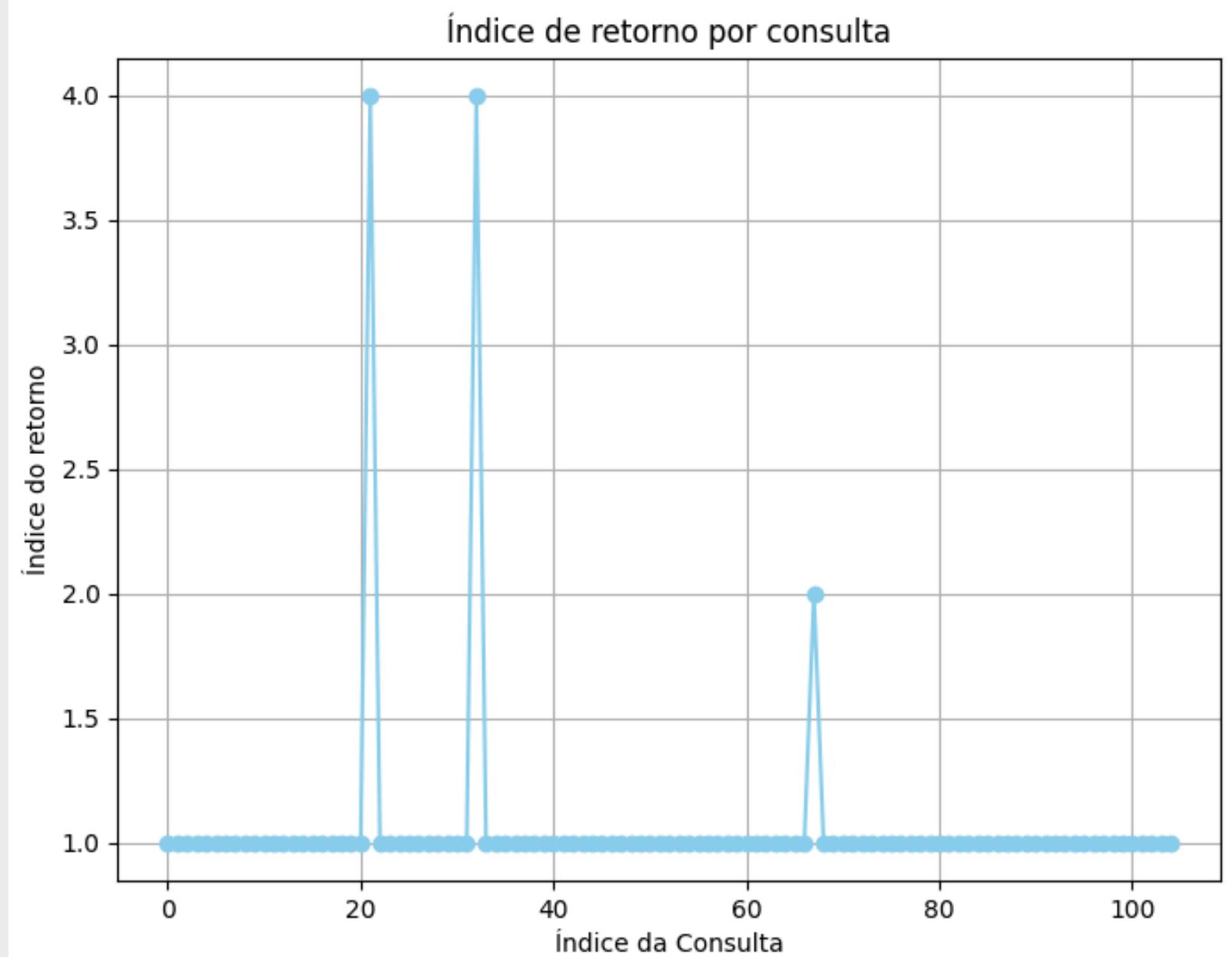


Resultados obtidos

Substituição de caracteres

Um caractere

Foram recuperadas dentro do top 3:
103 das 105 consultas

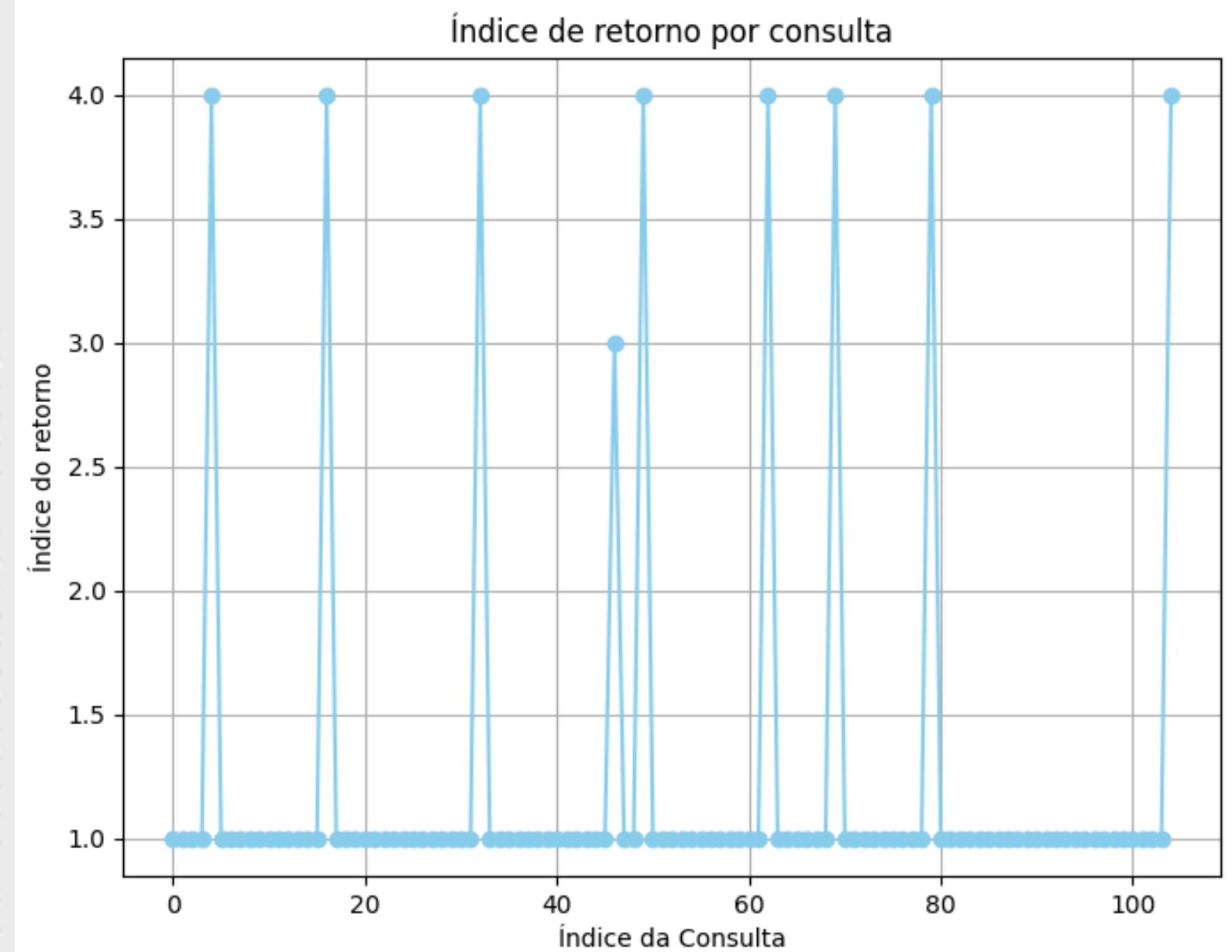


Resultados obtidos

Substituição de caracteres

Dois caracteres

Foram recuperadas dentro do top 3:
97 das 105 consultas

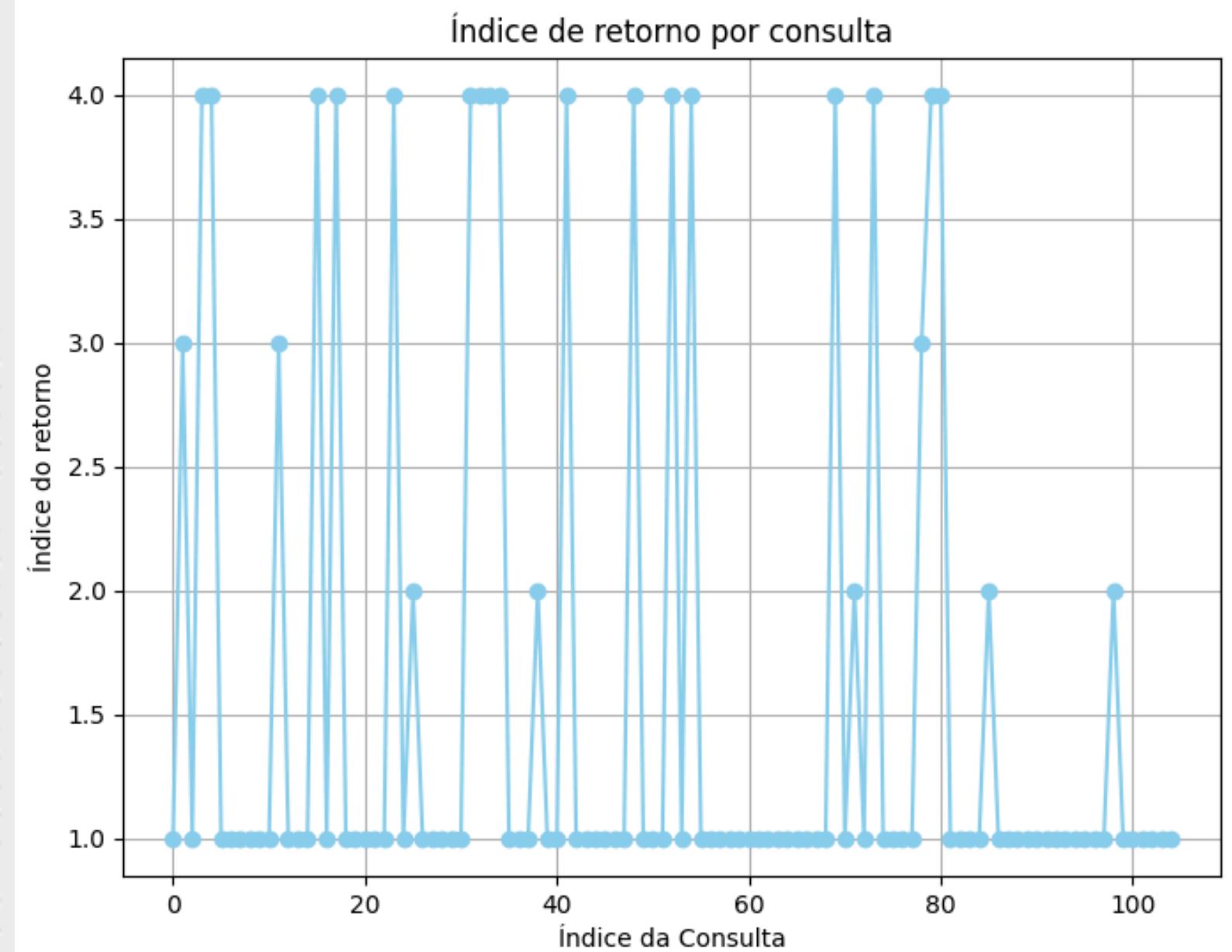


Resultados obtidos

Substituição de caracteres

Três caracteres

Foram recuperadas dentro do top 3:
88 das 105 consultas

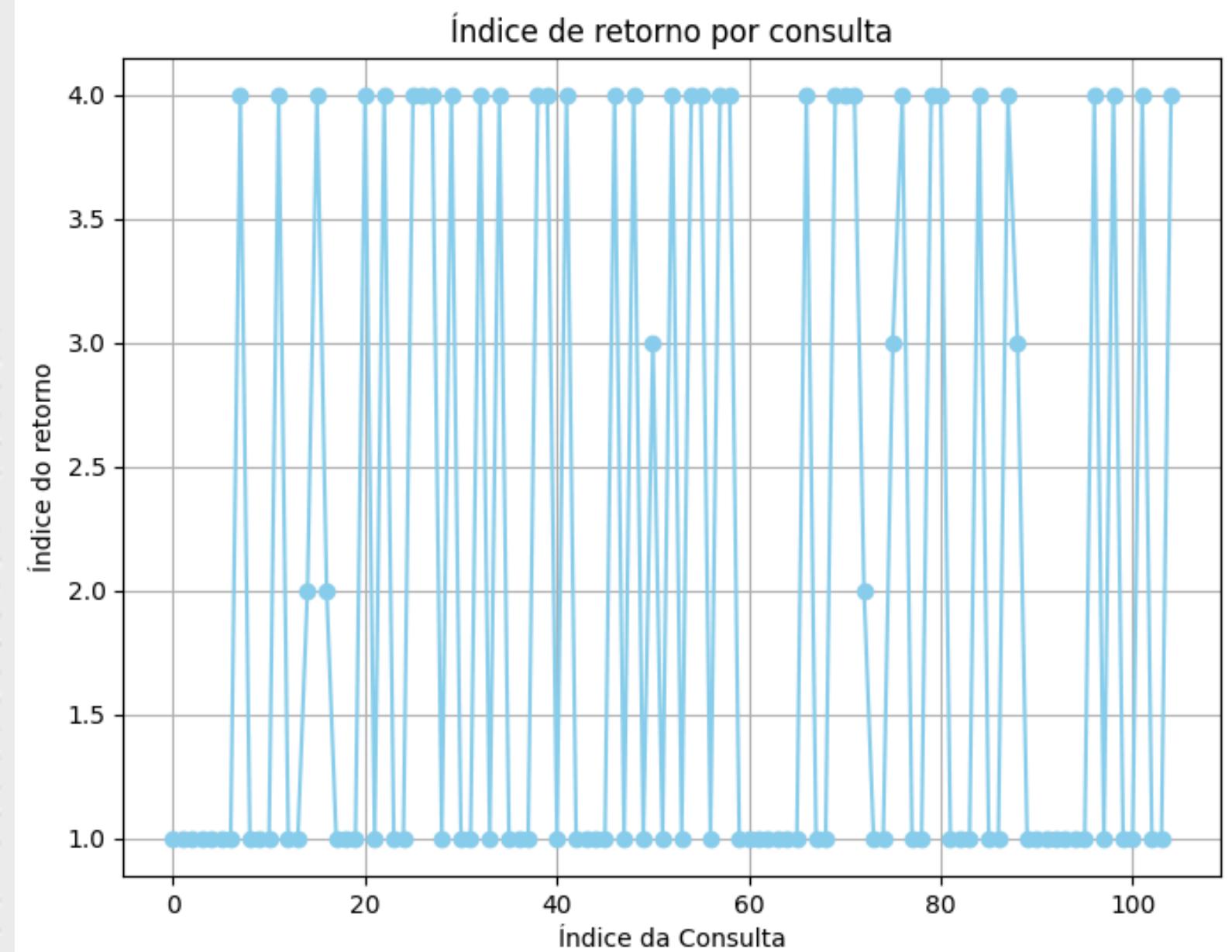


Resultados obtidos

Substituição de caracteres

Quatro caracteres

Foram recuperadas dentro do top 3:
71 das 105 consultas

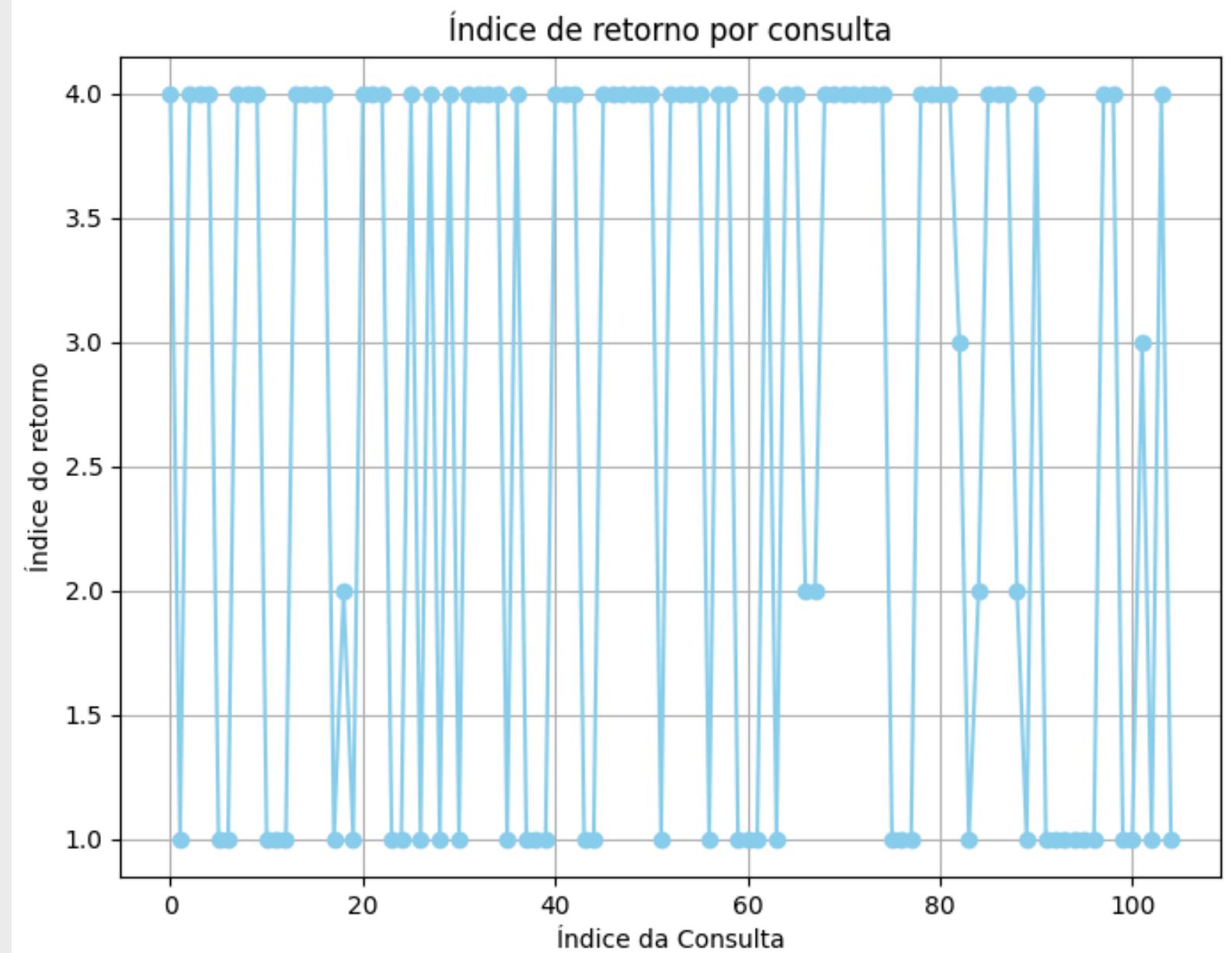


Resultados obtidos

Substituição de caracteres

Cinco caracteres

Foram recuperadas dentro do top 3:
47 das 105 consultas

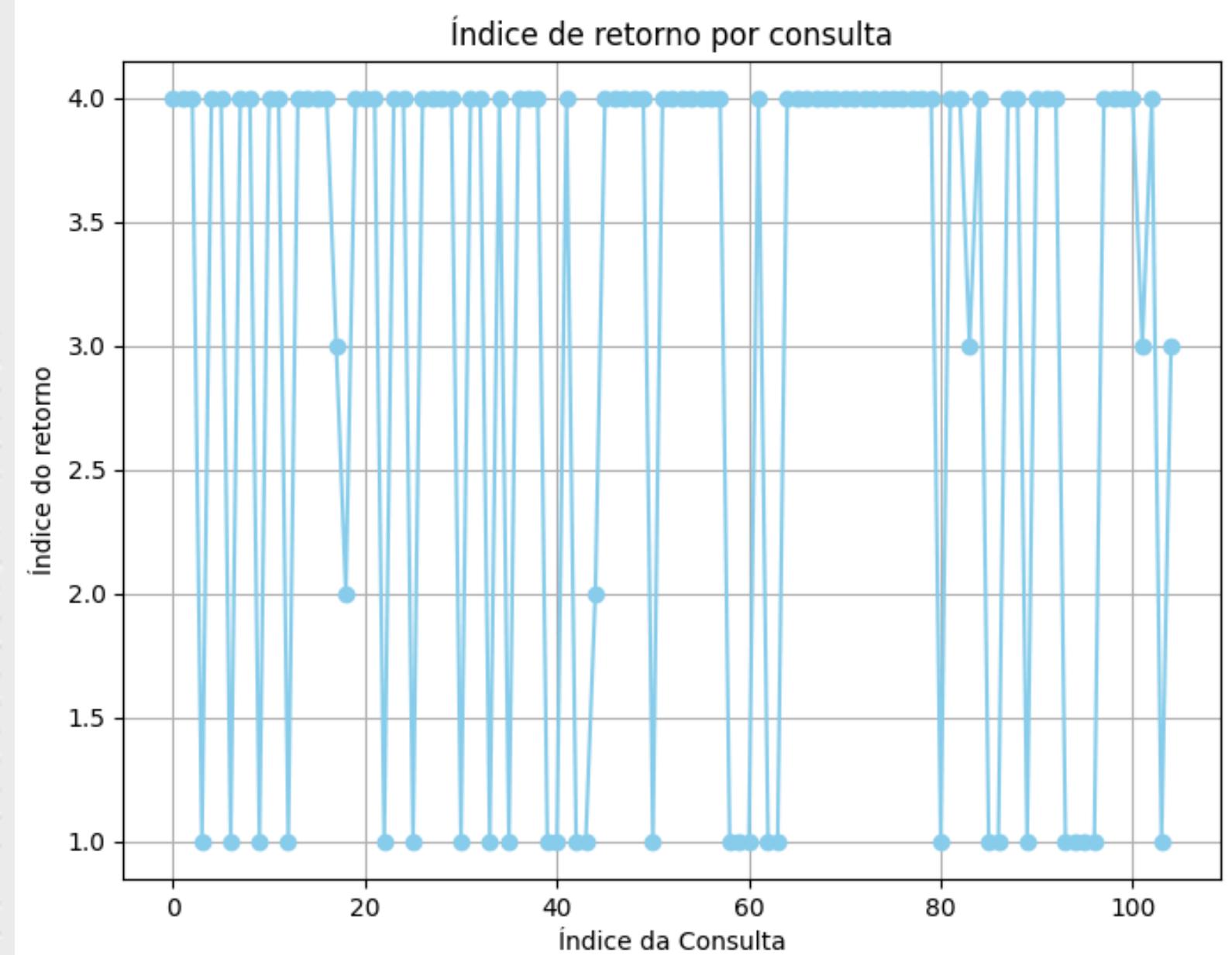


Resultados obtidos

Substituição de caracteres

Seis caracteres

Foram recuperadas dentro do top 3:
34 das 105 consultas

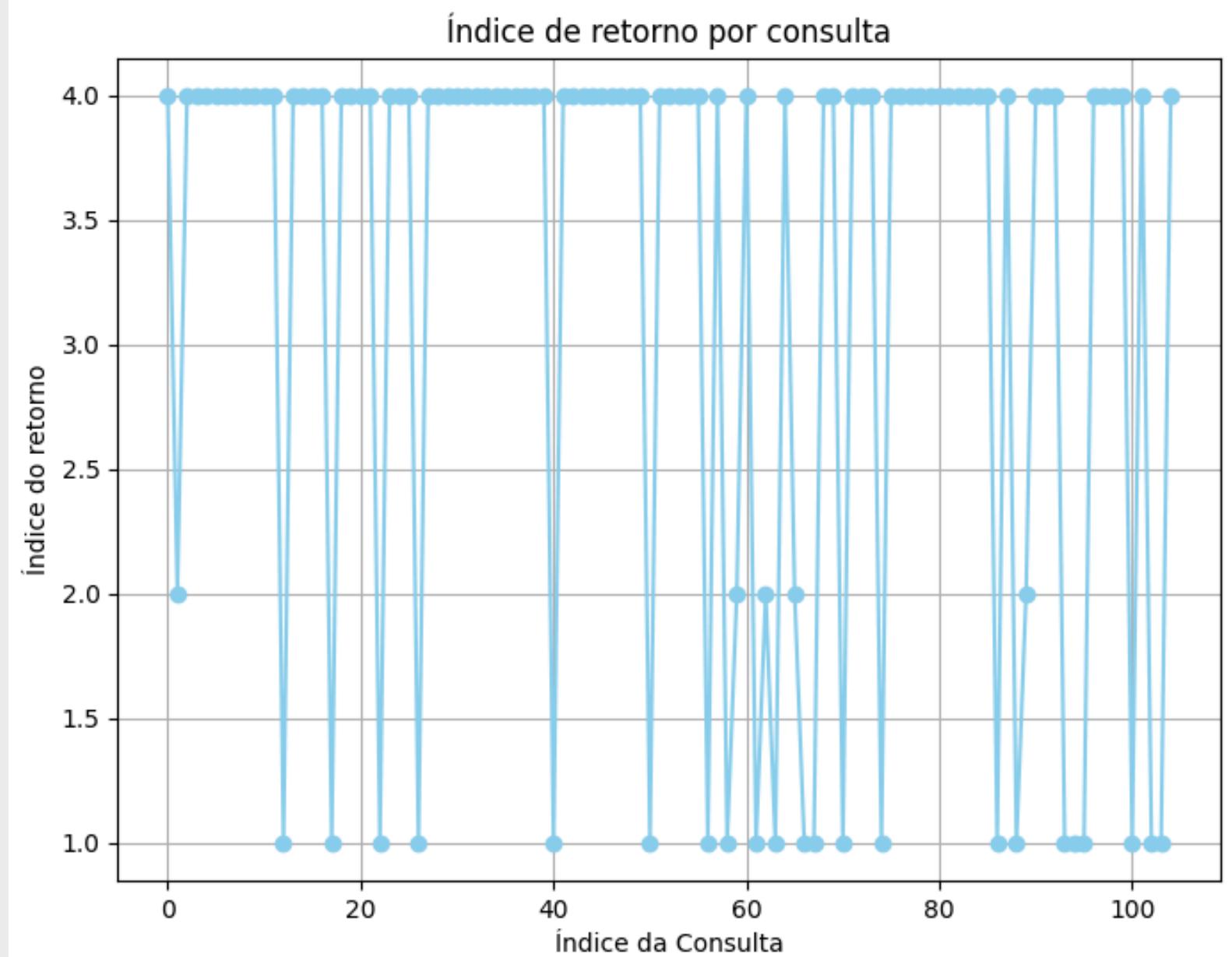


Resultados obtidos

Substituição de caracteres

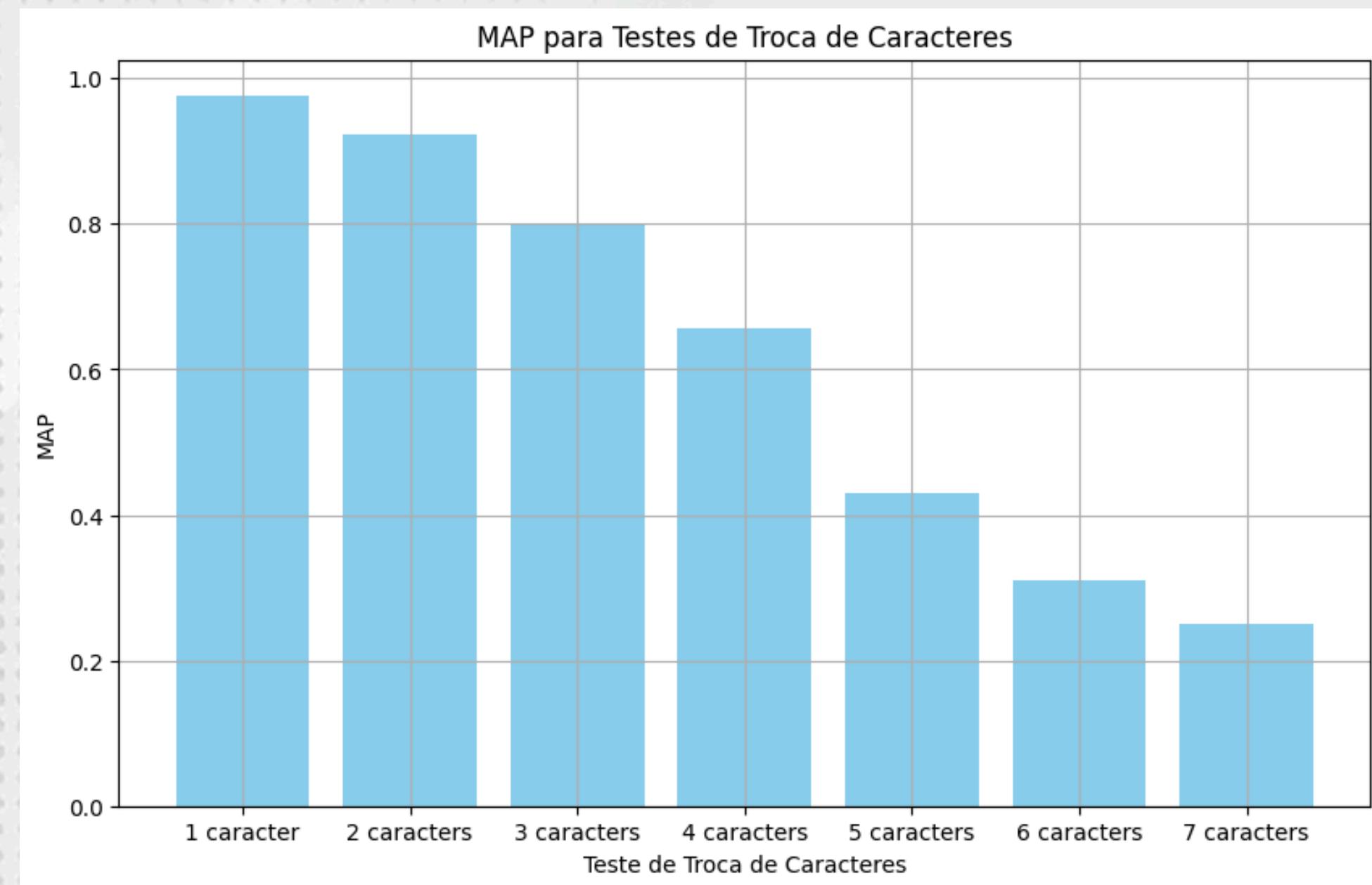
Sete caracteres

Foram recuperadas dentro do top 3:
27 das 105 consultas



Resultados obtidos

Substituição de caracteres



Conclusão

- Considerável robustez em face de pequenas perturbações
- Trocas de caracteres afeta menos o desempenho mas se torna prejudicial em grandes quantidades
- Ajustes adicionais podem ser necessários

Trabalhos Futuros

- Continuar desenvolvimento para o Dr. Thiago Macedo
- Treinar o BERT gerando um novo modelo

Referências

- ALVES, Maria Bernadete Martins; ARRUDA, Susana Margareth. Como fazer referências bibliográficas. Biblioteca Universitária. UFSC, Florianópolis, sd, 2008.
- MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. Sistemas inteligentes-Fundamentos e aplicações, v. 1, n. 1, p. 32, 2003.
- PADOVA, Ted. Adobe acrobat 9 PDF bible. John Wiley & Sons, 2008.
- GONZALEZ, Marco; LIMA, Vera Lúcia Strube. Recuperação de informação e processamento da linguagem natural. In: XXIII Congresso da Sociedade Brasileira de Computação. sn, 2003. P. 347-395.
- CASELI, Helena de Medeiros; NUNES, Maria das Graças Volpe; PAGANO, Adriana Silvina. O que é PLN?. Processamento de linguagem natural: conceitos, técnicas e aplicações em português, 2023.
- FALCÃO, João Vitor Regis et al. Redes neurais deep learning com tensorflow. RE3C-Revista Eletrônica Científica de Ciência da Computação, v. 14, n. 1, 2019.
- CARDOSO, Olinda Nogueira Paes. Recuperação de Informação. INFOCOMP Journal of Computer Science, v.2, n.1, p.33-38, 2000.
- KUMAR, Sanit et al. Web scraping using Python. International Journal of Advances in Engineering and Management, v. 4, n. 9, p. 235-237, 2022. <https://sbert.net/index.html>
- GOLLAPALLI, S. D., Li, Z., & Mitra, P. (2011). CiteSeerX: A Scholarly Big Dataset for Literature Search and Beyond. ACM Transactions on Information Systems (TOIS). WHITE, W. E., & Hernandez, P. (2016). Automated Reference Checking in Scholarly Manuscripts. Journal of Scholarly Publishing



UNIVERSIDADE FEDERAL
DE SÃO PAULO

MUITO OBRIGADO!