Link do repositório: https://github.com/siliciumtecnologia/Trabalho-DataScience.git

Modelos de linguagem

Modelos de linguagem são modelos desenvolvidos para realizar o processamento de linguagem natural. Eles têm o objetivo de entender ou gerar textos na mesma linguagem gerada pelos seres humanos. Essa tarefa é difícil pois a linguagem humana conta com diversas informações implícitas (dependem de contexto) (Kumar, 2013).

Alan Turing em seu trabalho *"Computing machinery and intelligence"*, propôs o "teste de Turing", onde uma máquina poderia ser considerada inteligente se um entrevistador humanos fosse incapaz de distingui-la de um ser humano (Turing, 1950).

Um dis primeiros métodos de processamento de linguagem natural é o distribucionalismo. Idéia proposta por Zellig Harris, baseia-se na obtenção do significado das palavras através de sua distribuição estatística, em vez de obtê-los através de dicionários. Segundo Harris (1954), palavras de ocorrem em contextos similares, possuem significados relacionados. Logo, é possível rastrear a evolução semântica das palavras (Yildirim e Chenaghlu, 2024).

Um dos métodos utilizados para essa abordagem é o n-grama. Baseia-se no fato de que a probabilidade da próxima palavra em um determinado texto é dependente de um conjunto fixo e palavras anteriores. Isso é conhecido como "suposição de Markov" (Jurafsky e Martin, 2025).

É possível considerar a probabilidade da próxima palavra apenas observando a própria palavra. Esse modelo é chamado de ungiram. Se for considerada uma palavra antes, então obtêm-se o digrama. Para duas palavras anteriores, o Trigrama, etc.

E comum realizar o processamento das palavras em *tokens*, que são representações simplificada das palavras, que podem ser gerenciadas pela máquina (Grefenstette, 1999).

A frase `The quick brown fox jumps over the lazy dog.` pode ser *tokenizada* a partir da separação de palavras pelo espaço em branco.

Como forma de separar as frases nas listas de *tokens*, utiliza-se marcadores de início e final

da frase `<s>` e `</s>` (Jurafsky e Martin, 2025).

```
['<s>', 'the', 'quick', 'brown', 'fox', 'jumps', 'over', 'the', 'lazy',
'dog', '</s>']
```

A partir dessa frase tokenizada pode-se gerar um digrama, por exemplo:

```
[('<s>', 'the'), ('the', 'quick'), ('quick', 'brown'), ('brown', 'fox'),
('fox', 'jumps'), ('jumps', 'over'), ('over', 'the'), ('the', 'lazy'),
('lazy', 'dog'), ('dog', '</s>')]
```

Modelo estatísticos para n-gramas

O uso de modelos estatísticos cria um problema. Muitas sequencias de linguagem não são contempladas - pois as bases de treinamento são finitas. Logo, o uso de téncnicas de suavização tendem a corrigir esse problema, atribuindo um peso para relações não visualizadas (Jurafsky e Martin, 2025).

Suavização de Laplace

É a técnica mais simples, pois atribui peso 1 para todos os n-gramas. Essa técnica não consuma desempenhar muito bem, mas serve de base para as demais técnicas (Jurafsky e Martin, 2025).

Interpolação de desconto absoluto

Essa técnica faz uso da subtração de valores absolutos às probabilidades dos n-gramas (Jurafsky e Martin, 2025). Esse métodos baseia o métodos de Kneser-Ney (Jurafsky e Martin, 2025, Ney et al.,1994).

Suavização de Witten-Bell

Essas Tecnica aplica probabilidades de n-gramas vistos uma vez para n-gramas vistos 0 vezes (Rusli, 2014).

Suavização de Kneser-Ney

É considerada a técnica de suavização mais efetiva em função de usar uma técnica de subtração dos valores de probabilidade de n-gramas com baixas frequências (Ney et al.,1994). Um exemplo dessa técnica é o caso do digrama “San Francisco”. Ele aparece várias vezes em uma base de treinamento, da mesma forma que a frequência do ungiram “Francisco” também é alta. Para evitar resultados distorcidos, o o método de Kneser-Ney faz uma correção considerando a relação deste ungiram com as possíveis palavras que o procedem (McCartney, 2005).

Perplexidade

Para a avaliação do desempenho dos modelos, não é utilizado a probabilidade estatística, pois ela é dependente do comprimento do texto. Em vez disso, uma outra métrica - a perplexidade, é utilizada, pois ela é normalizada por *token* Essa métrica é a mais importante em processamento de linguagem natural. A perplexidade é o inverso da probabilidade, normalizada pelo número de *tokens*. Essa inversão indica que, quanto menor a perplexidade, melhor é o modelo (Jurafsky e Martin, 2025).

Análise de perplexidade de frases geradas por modelos estatísticos utilizando n-grama

Através do treinamento de 4 modelos estatísticos, utilizando várias combinações de n-grama, foi realizada a análise de perplexidade de frase geradas a partir desses modelos treinados com um corpo de texto e, então comparados para verificar quanto esses modelos conseguem gerar um texto consistente com o corpo de treinamento.

Dataset utilizado

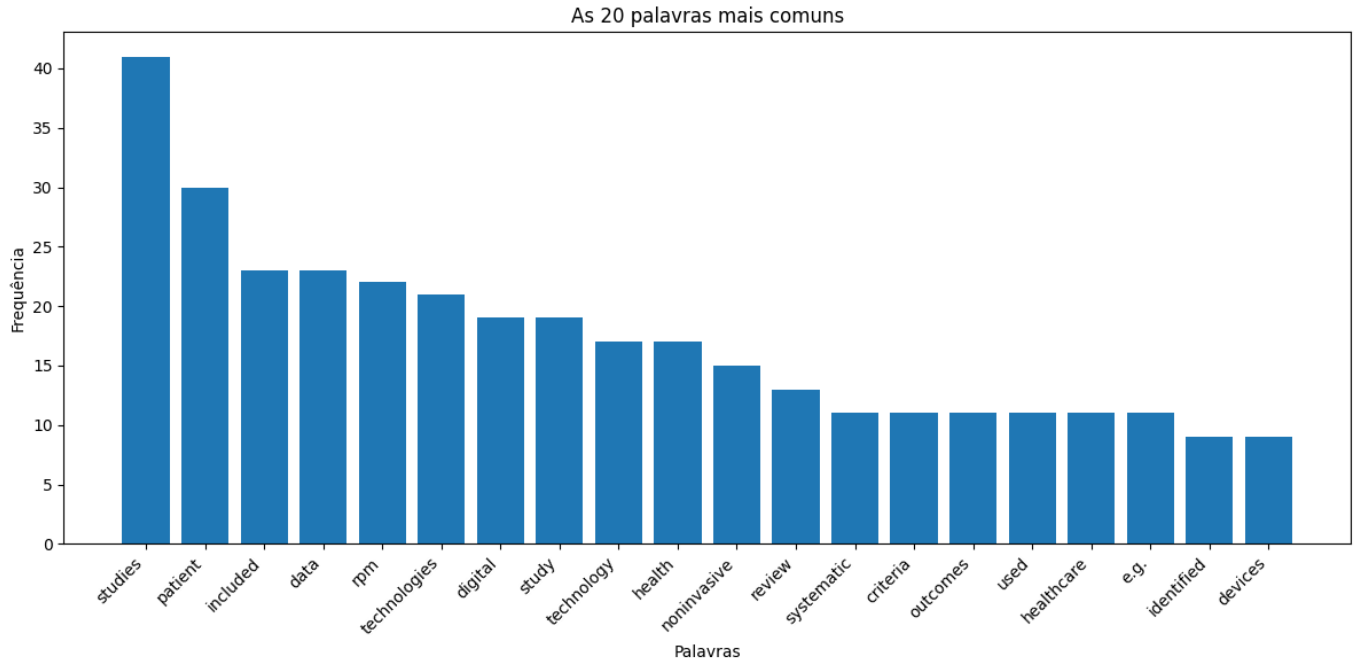
Para o treinamento, foi utilizado o texto do seguinte artigo:

Título: Remote Patient Monitoring via Non-Invasive Digital Technologies: A Systematic Review
Autores: Ashok Vegesna, PharmD Melody Tran, PharmD Michele Angelaccio and Steve

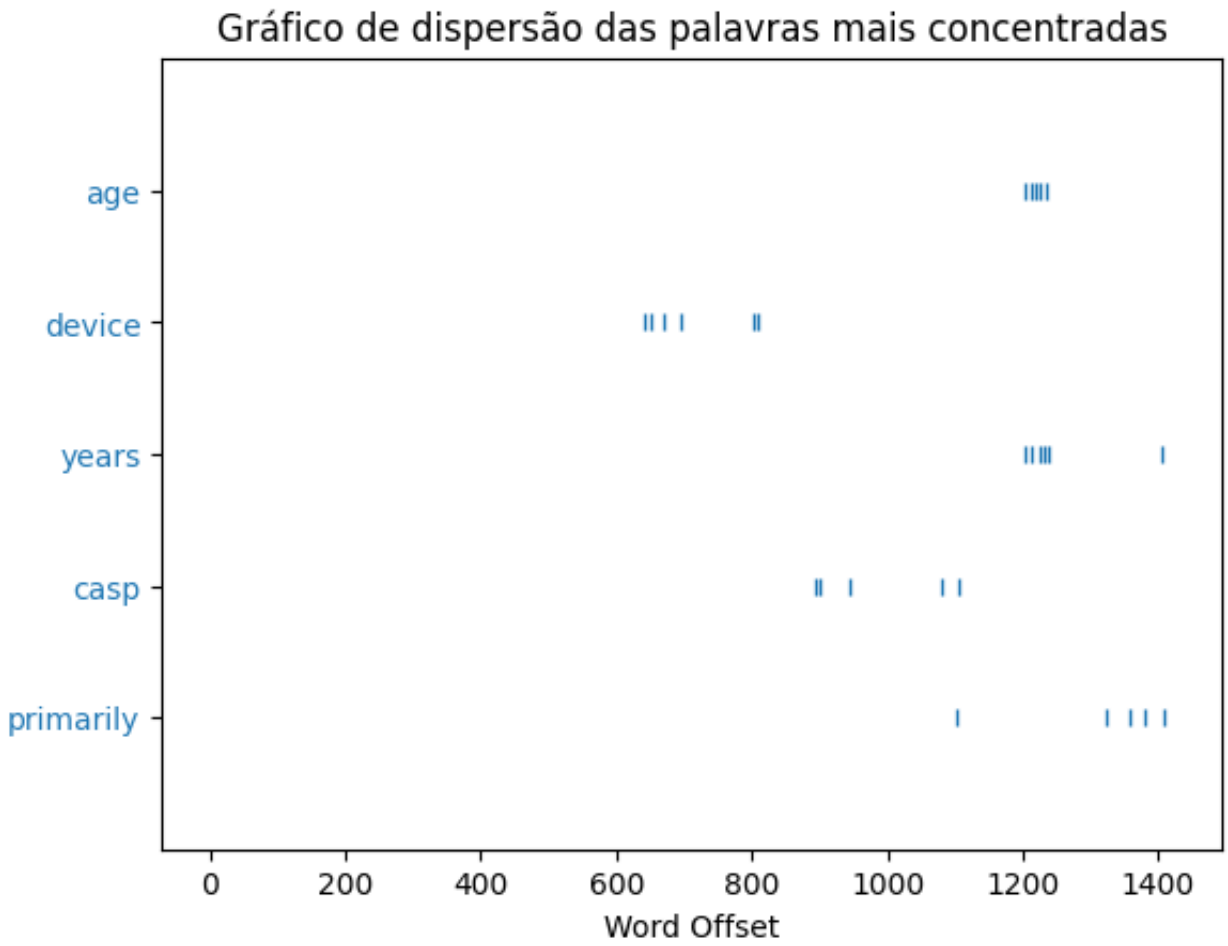
Arcona, PhD
Link: [https://www.liebertpub.com/doi/10.1089/tmj.2016.0051]

Esse artigo foi processado por Inteligência Artificial, para extrair do PDF o texto contido nele, eliminando nomes, notas de rodapé, descrição de imagens, fórmulas, referências e qualquer tipo de formatação.

O texto foi *tokenizado*, de onde se removeu a pontuação e resultou em 2109 *tokens*. Desses tokens, foram removidas palavras de parada, que não possuem significado semântico, e então obteve-se as 20 palavras mais comuns no texto.



Uma outra observação importante é o gráfico de dispersão das palavras mais concentradas (que aparecem mais juntas).



Pode-se observar que essas palavras mais concentradas costumam localiza-se no final to texto. Isto indica que essa parte do texto contém uma conclusão, onde se faz o uso desses termos mais restritos a esse tipo e contexto.

Modelos testados

Foi realizado o treinamento de 4 modelos pra n-grama: Laplace, AbsoluteDiscountingInterpolated, WittenBellInterpolated, KneserNeyInterpolated, da biblioteca NLTK. Também foram treinados considerando bigramas, trigramas, 4-gramas e 5-gramas. Após cada treinamento, é solicitada a geração de uma frase de até 20 palavras usando como semente o seguinte texto: “common clinical data captured by these technologies”. Este trecho encontra-se no texto de treinamento.

Após a geração, é realizado o cálculo da perplexidade desta frase com o modelo. No final, um comparativo das perplexidades de cada modelo, comparando os diferente n-gramas é apresentado.

Resultados do processamento

Abaixo segue o resultado do processamento para os modelos:

Processando modelos para 2-grama
Modelo: Laplace
Vocabulário: 684
Texto gerado:common clinical data captured by these technologies was conducted by these variances this study variables including the search terms of studies were randomized controlled trials rcts cohort
Perplexidade do texto gerado: 446.6582079315395

Modelo: AbsoluteDiscountingInterpolated
Vocabulário: 684
Texto gerado:common clinical data captured by these technologies was conducted by the systematic review of the trustworthiness and systematic review noted the systematic review used to determine whether
Perplexidade do texto gerado: 57.562752408412116

Modelo: WittenBellInterpolated
Vocabulário: 684
Texto gerado:common clinical data captured by these technologies was conducted by these variances this study validity bias study validity bias study selection process utilized wearable devices.
Perplexidade do texto gerado: 18.618300327760952

Modelo: KneserNeyInterpolated
Vocabulário: 684
Texto gerado:common clinical data captured by these technologies was conducted by these variances this study validity bias study validity bias study selection process utilized wearable devices.
Perplexidade do texto gerado: 18.640698526661915

Processando modelos para 3-grama
Modelo: Laplace
Vocabulário: 684
Texto gerado:common clinical data captured by these technologies was consistent with our definition of rpm via noninvasive rpm in various patient populations of the systematic review of the
Perplexidade do texto gerado: 495.93162099658196

Modelo: AbsoluteDiscountingInterpolated
Vocabulário: 684
Texto gerado:common clinical data captured by these technologies was consistent with our definition of rpm via noninvasive digital technology smartphones/personal digital assistants pdas wearables biosensors computerized systems or
Perplexidade do texto gerado: 10.524982436848905

Modelo: WittenBellInterpolated
Vocabulário: 684
Texto gerado:common clinical data captured by these technologies was consistent with our definition of rpm via noninvasive rpm in various patient populations of the systematic review of the
Perplexidade do texto gerado: 30.502501157740305

Modelo: KneserNeyInterpolated
Vocabulário: 684
Texto gerado:common clinical data captured by these technologies was consistent with our definition of rpm via noninvasive rpm in various patient populations of the systematic review of the
Perplexidade do texto gerado: 71.56401891177504

Processando modelos para 4-grama
Modelo: Laplace
Vocabulário: 684
Texto gerado:common clinical data captured by these technologies was consistent with our definition of rpm via noninvasive digital technology studies by technology category included in the systematic review
Perplexidade do texto gerado: 509.0652726442193

Modelo: AbsoluteDiscountingInterpolated
Vocabulário: 684
Texto gerado:common clinical data captured by these technologies was consistent with our definition of rpm via noninvasive digital technology smartphones/personal digital assistants pdas wearables biosensors computerized systems or
Perplexidade do texto gerado: 8.252430716788215

Modelo: WittenBellInterpolated
Vocabulário: 684

Texto gerado:common clinical data captured by these technologies was consistent with our definition of rpm via noninvasive digital technology studies by technology category included in the systematic review
Perplexidade do texto gerado: 20.602233834668553

Modelo: KneserNeyInterpolated
Vocabulário: 684
Texto gerado:common clinical data captured by these technologies was consistent with our definition of rpm via noninvasive digital technology studies by technology category included in the systematic review
Perplexidade do texto gerado: 56.77831355678958

Processando modelos para 5-grama
Modelo: Laplace
Vocabulário: 684
Texto gerado:common clinical data captured by these technologies was consistent with our definition of rpm via noninvasive digital technology smartphones/personal digital assistants pdas wearables biosensors computerized systems or
Perplexidade do texto gerado: 464.892957252438

Modelo: AbsoluteDiscountingInterpolated
Vocabulário: 684
Texto gerado:common clinical data captured by these technologies was consistent with our definition of rpm via noninvasive digital technology.
Perplexidade do texto gerado: 12.02251881499256

Modelo: WittenBellInterpolated
Vocabulário: 684
Texto gerado:common clinical data captured by these technologies was consistent with our definition of rpm via noninvasive digital technology smartphones/personal digital assistants pdas wearables biosensors computerized systems or
Perplexidade do texto gerado: 6.684848588371884

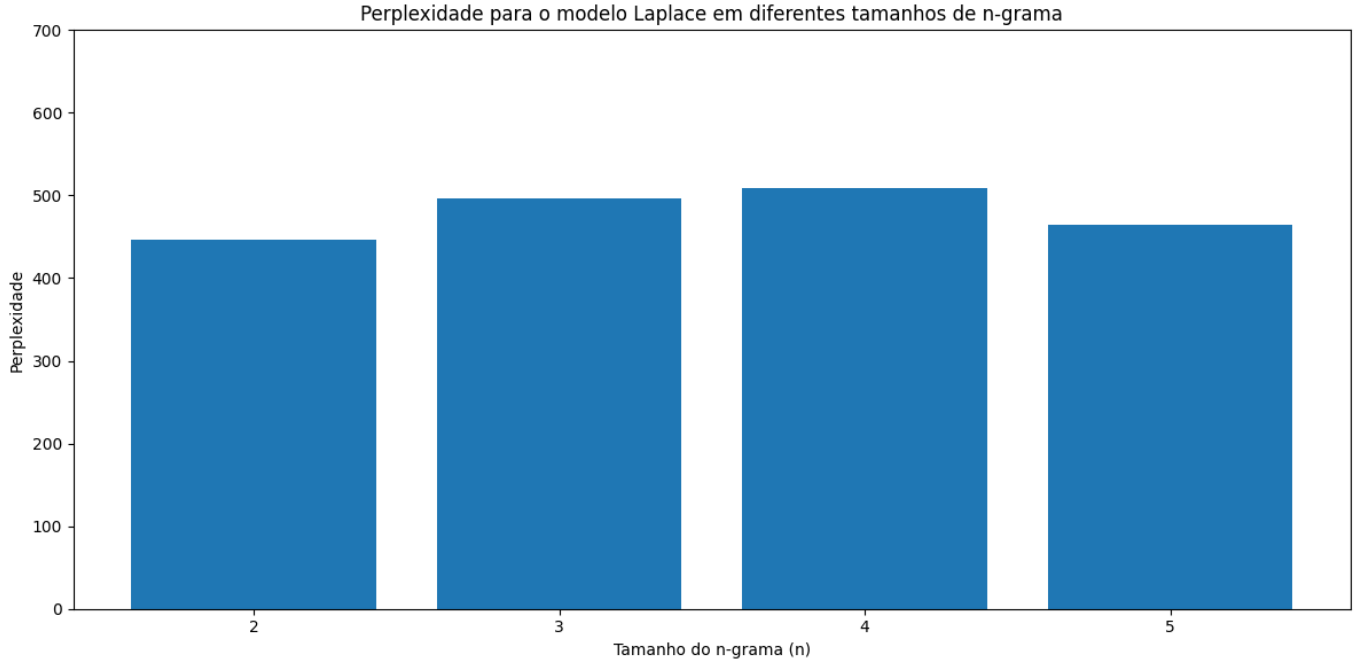
Modelo: KneserNeyInterpolated
Vocabulário: 684
Texto gerado:common clinical data captured by these technologies was consistent with our definition of rpm via noninvasive digital technology smartphones/personal digital assistants pdas wearables biosensors computerized systems or

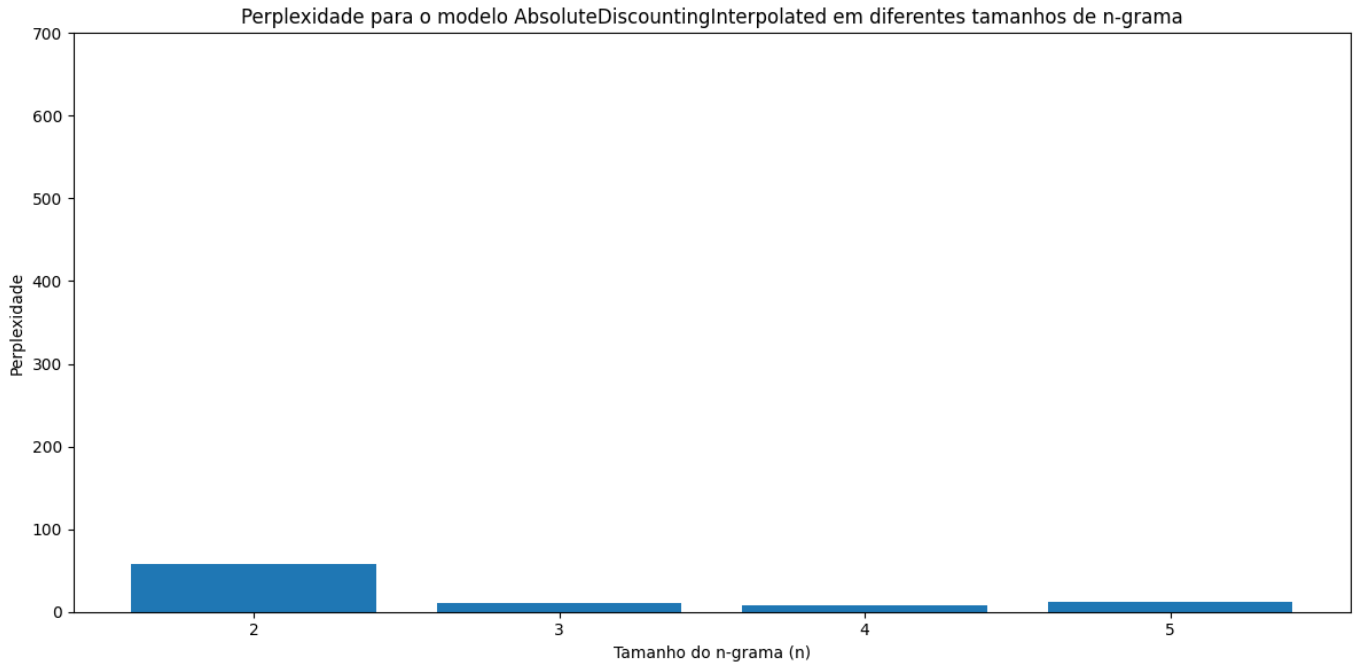
Perplexidade do texto gerado: 13.385168658579921

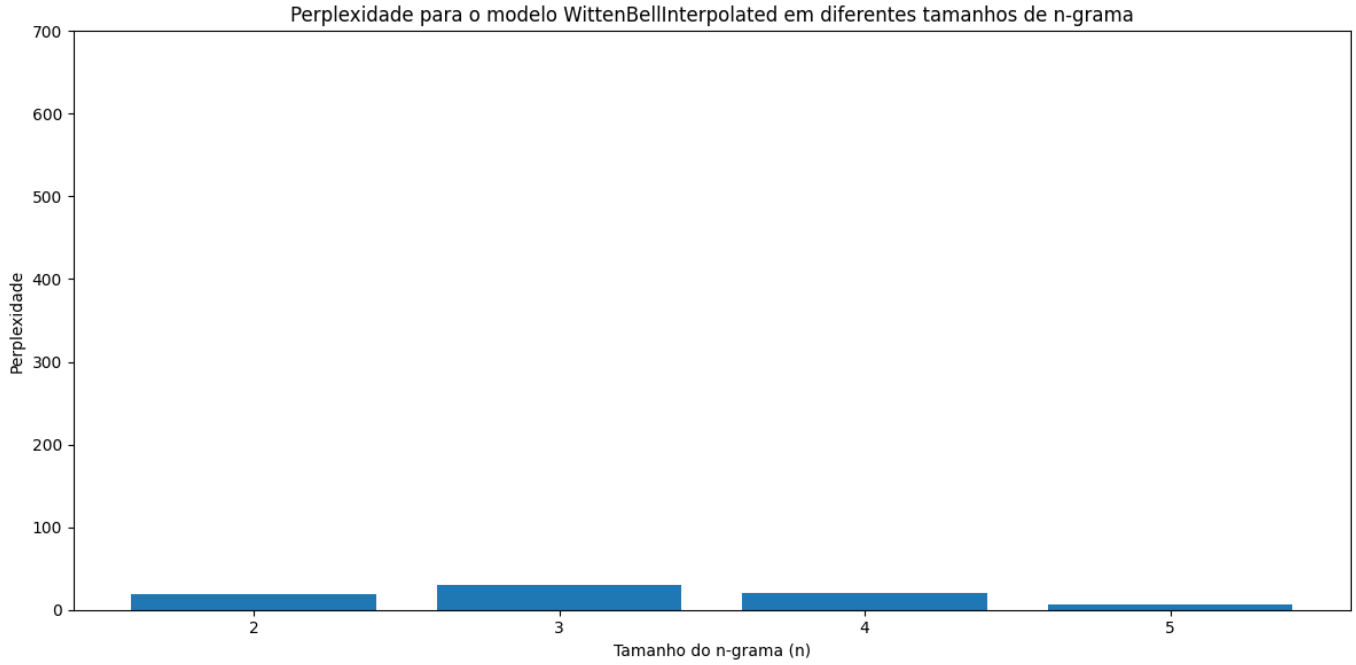
Concluído

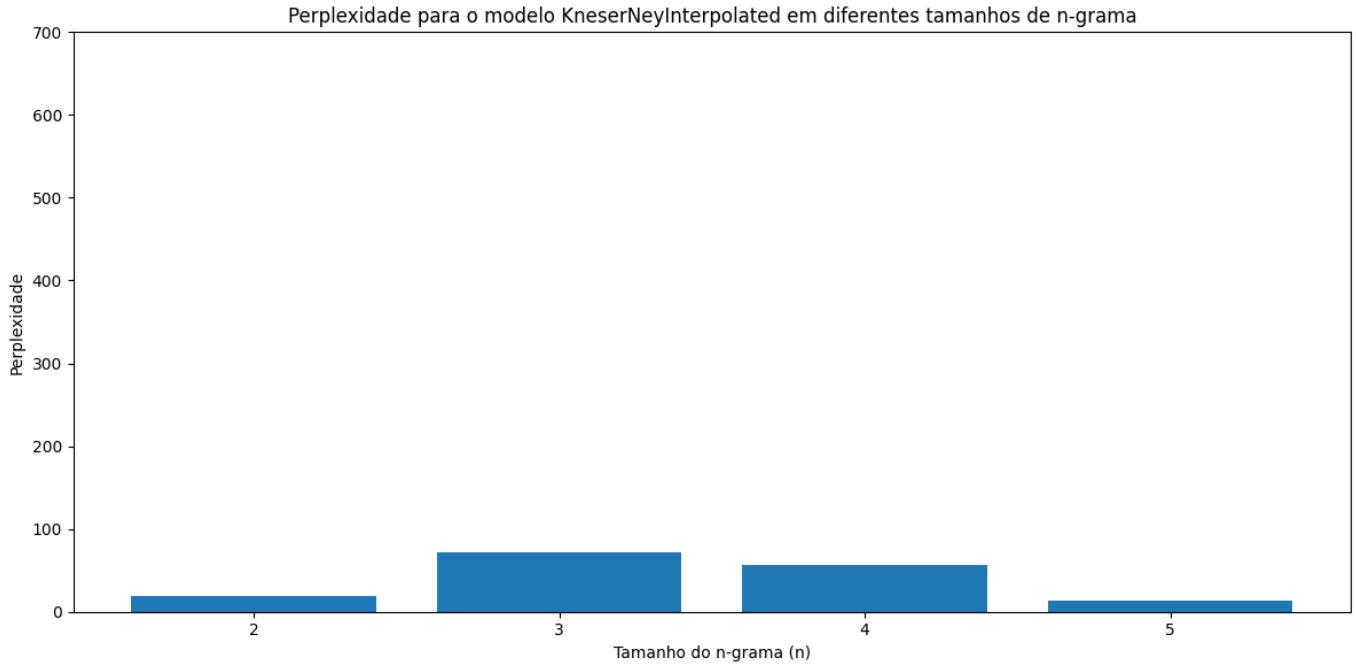
Gráficos comparativos

Os gráficos abaixo apresentam o desempenho de perplexidade para os 4 modelos considerados os diversos n-gramas.









O dsempenho do modelo de Laplace foi o pior, pois ele usa uma técnica simplória. Já os outros três modelos tiveram desempenhos mistos. O modelo de Witten-Bell e o de desconto absoluto apresentaram a menor perplexidade, o que indica que esses modelos teviveram um desempenho melhor. O modelo Kneser-Ney teve um desempenho bom também, embora pior que os outros dois.

Referências

Grefenstette, G. (1999). Tokenization. In: van Halteren, H. (eds) Syntactic Wordclass Tagging. Text, Speech and Language Technology, vol 9. Springer, Dordrecht. https://doi.org/10.1007/978-94-015-9273-4_9

Harris, Z. S. (1954). Distributional Structure, WORD 10(2-3): 146–162. URL: http://www.tandfonline.com/doi/full/10.1080/00437956.1954.11659520

Jurafsky, D. e Martin, J. H. (2025). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edn. URL: https://web.stanford.edu/jurafsky/slp3/oldjan25/ed3bookJan25.pdf

Kumar, E. (2013). Natural Language Processing, I.K. International Publishing House Pvt. Limited. URL: https://books.google.com.br/books?id=FpUBFNFuKWgC

McCartney B. (2005). NLP Lunch Tutorial: Smoothing. Stanford. https://nlp.stanford.edu/~wcmac/papers/20050421-smoothing-tutorial.pdf

Ney H., Essen U., Kneser R. (1994). On structuring probabilistic dependences in stochastic language modelling. Computer Speech & Language, Volume 8, Issue 1. Elsevier. https://doi.org/10.1006/csla.1994.1001.

Rusli, Ismail. (2014). Comparison of Modified Kneser-Ney and Witten-Bell Smoothing Techniques in Statistical Language Model of Bahasa Indonesia. 2014 2nd International Conference on Information and Communication Technology, ICoICT 2014. 10.1109/ICoICT.2014.6914097.

TURING, A. M. (1950). I.COMPUTING MACHINERY AND INLIGENCE, Mind LIX(236): 433–460. _eprint: https://academic.oup.com/mind/article-pdf/LIX/236/433/61209000/mind_lix_236_433.pdf. URL: https://doi.org/10.1093/mind/LIX.236.433

Yildirim, S. e Chenaghlu, M. (2024). Mastering Transformers: The Journey from BERT to Large Language Models and Stable Diffusion, Packt Publishing. URL: https://books.google.com.br/books?id=M_wJEQAAQBAJ