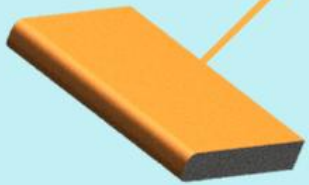
APLICAÇÃO DE MODELOS DE LINGUAGEM NATURAL PARA

COMPARAÇÃO INTERTEXTUAL

POR: L. F. LAGUARDIA

Probabilidade de uma Palavra

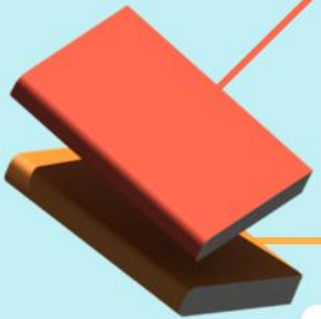


Como **eu** adoro esses fogos embelezando o céu. **Eu** quase me esqueço que você tá do meu lado exigindo atenção *[...]*

Sol da Manhã, Supercombo - 2014

$$P(eu) = \frac{count(eu)}{count(\textbf{texto})} = \frac{2}{20} = 0.1$$

Probabilidade de uma Palavra - Na Prática



Como **eu** amo esses fogos enfeitando o céu.

Como **eu** adoro esses fogos embelezando o céu. **Eu** quase me esqueço que você tá do meu lado exigindo atenção *[...]*

$$P(eu) = \frac{count(eu) + 1}{count(\textbf{texto}) + count(\textbf{novas})} = \frac{2+1}{20+2} \approx 0.14$$

Fontes dos Modelos

Fontes dos Modelos

Twitter

250 tweets do G1.

#DebateNaGlobo: Comentaristas do g1 e da GloboNews avaliam desempenho dos candidatos [...]

Fontes dos Modelos

Vingadores

Legendas dos 4 filmes.

[...] - Eu tive muita sorte.
- Sim, eu sei.
- Muita gente não. [...]

Twitter

250 tweets do G1.

Fontes dos Modelos

Receitas

100 blogs de receitas.

Do “As Minhas Receitas”: *Em uma taça misture a água morna com o açúcar e o fermento e deixe [...]*

Twitter

250 tweets do G1.

Vingadores

Legenda dos 4 filmes.

Fontes dos Modelos

Twitter
250 tweets do G1.

Vingadores
Legenda dos 4 filmes.

Receitas
100 blogs de receitas.

Supercombo
As 79 letras da banda.

[...] Eu devia sorrir mais
Abraçar meus pais
Viajar o mundo e socializar [...]

Fontes dos Modelos

Twitter
250 tweets do G1.

Vingadores
Legenda dos 4 filmes.

Receitas
100 blogs de receitas.

Wikipedia
25 artigos matemáticos.

[...] A matemática (dos termos gregos, μάθημα, transliterado máthēma, 'ciência', [...]

Supercombo
As 79 letras da banda.

Fontes dos Modelos

Brás Cubas

A obra completa.

[...] Virgília tinha agora a beleza da velhice, um ar austero e maternal; estava menos magra do que [...]

Twitter

250 tweets do G1.

Vingadores

Legenda dos 4 filmes.

Receitas

100 blogs de receitas.

Supercombo

As 79 letras da banda.

Wikipedia

25 artigos matemáticos.

Fontes dos Modelos

Twitter
250 tweets do G1.

Vingadores
Legenda dos 4 filmes.

Receitas
100 blogs de receitas.

Supercombo
As 79 letras da banda.

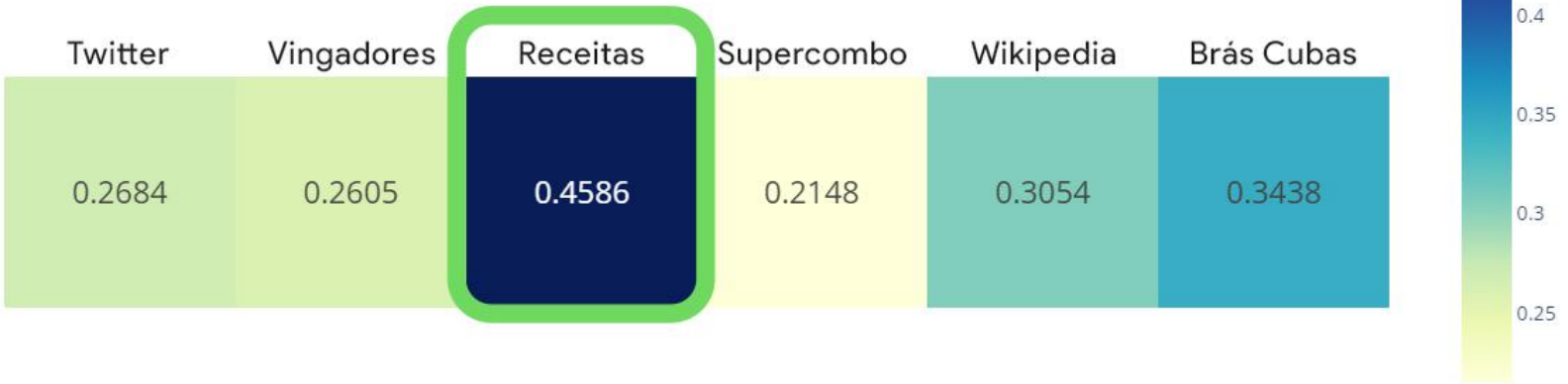
Wikipedia
25 artigos matemáticos.

Brás Cubas
A obra completa.

Comparação dos Resultados

Pontuação para:

“três colheres de sopa”

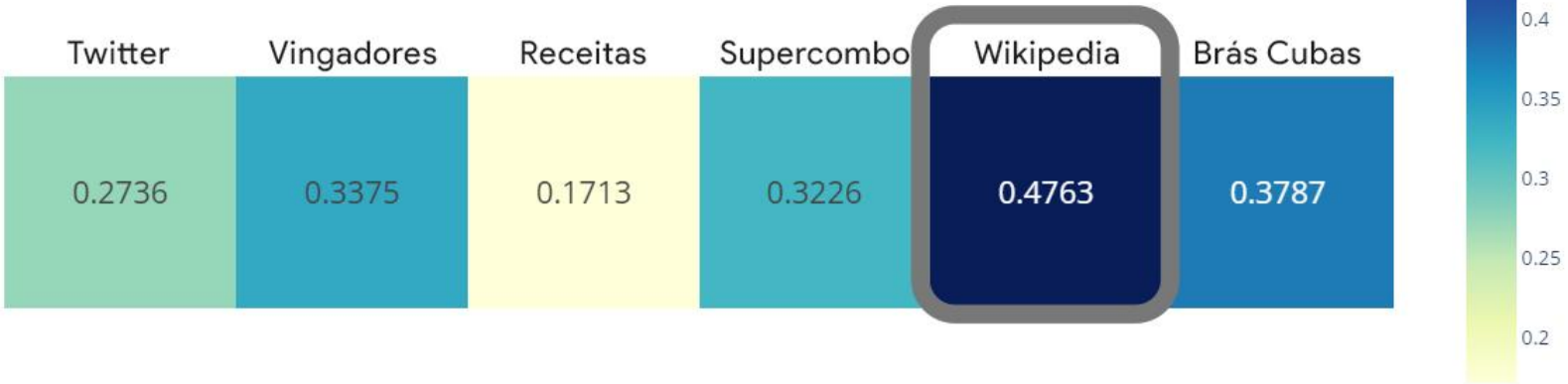


Como esperado, o modelo treinado com receitas foi o que recebeu a maior pontuação, pois é onde essas palavras são mais frequentes. É importante notar que, embora essa pontuação seja proporcional à, ela não equivale à probabilidade - isto é, não há 45% de chance de encontrar essa frase no corpus de receitas.

Comparação dos Resultados

Pontuação para:

“A prova matemática é simples e trivial.”



Como o nosso modelo da Wikipedia é baseado apenas em artigos matemáticos, o resultado mais uma vez saiu como o esperado, apontando um match maior com os textos da Wikipedia.

Comparação dos Resultados

Pontuação para:

Notícia do G1

Textos longos apresentam muitas palavras desconhecidas nos modelos com corpus menores. Por isso, a pontuação dos modelos da Wikipedia e Brás Cubas geralmente é maior para trechos maiores, como a notícia. Apesar disso, os tweets ainda tiveram o 3º maior match.

Notícia usada: https://g1.globo.com/politica/noticia/2022/10/11/propostas-de-mudancas-na-composicao-do-stf-sao-inconstitucionais-e-agridem-democracia-diz-entidade.ghtml

Comparação dos Resultados

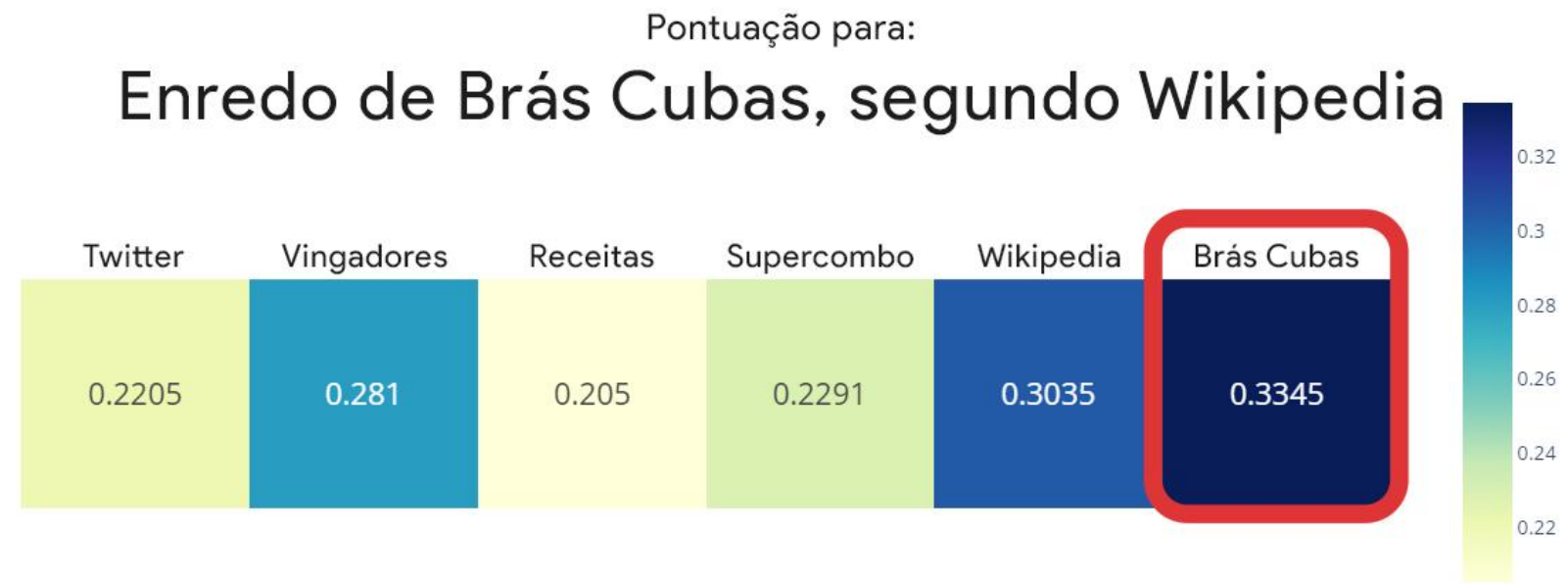
Pontuação para:

Tweet da banda Supercombo

Os tweets são geralmente feitos em uma linguagem mais informal, em tom de conversa. Por isso, o maior match foi com as legendas de filme, que representam principalmente conversas. A grande diferença com o modelo baseado em tweets se deve ao fato do G1 ter uma comunicação mais formal, mesmo nas redes.

Tweet usado: https://twitter.com/Supercombo/status/1496120917503655941?s=20&t=ES2AZVNHfIGZob9SIUofZA

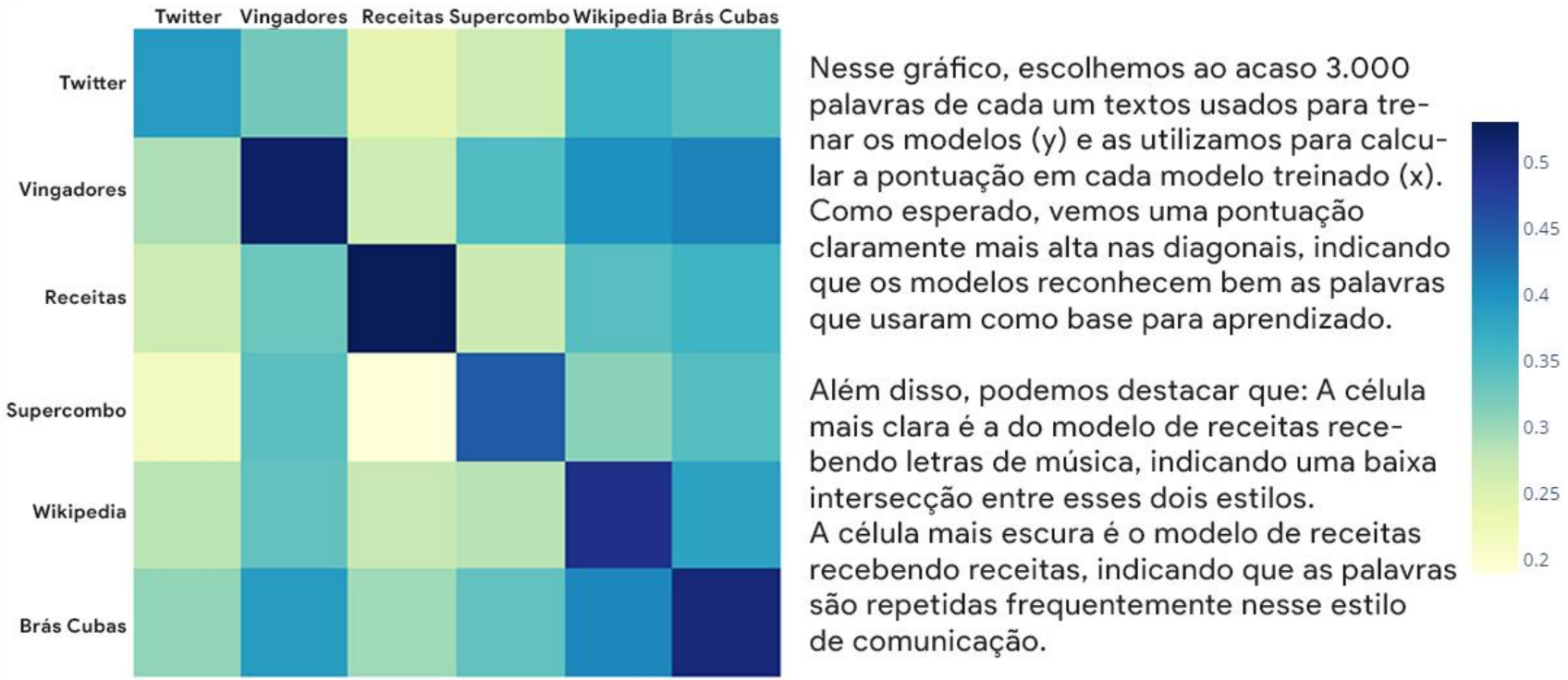
Comparação dos Resultados



Para resumir a história, o artigo da Wikipedia utiliza frequentemente os nomes das personagens. Como os nomes só aparecem no corpus do livro, a pontuação desse modelo é beneficiada. Além disso, é importante lembrar que o modelo da Wikipedia só conhece artigos de matemática, o que pode prejudicá-lo aqui.

Fonte: https://pt.wikipedia.org/wiki/Mem%C3%B3rias_P%C3%B3stumas_de_Br%C3%A1s_Cubas

Comparação Intertextual



Nesse gráfico, escolhemos ao acaso 3.000 palavras de cada um textos usados para treinar os modelos (y) e as utilizamos para calcular a pontuação em cada modelo treinado (x). Como esperado, vemos uma pontuação claramente mais alta nas diagonais, indicando que os modelos reconhecem bem as palavras que usaram como base para aprendizado.

Além disso, podemos destacar que: A célula mais clara é a do modelo de receitas recebendo letras de música, indicando uma baixa intersecção entre esses dois estilos.
A célula mais escura é o modelo de receitas recebendo receitas, indicando que as palavras são repetidas frequentemente nesse estilo de comunicação.

Aplicações + Extras

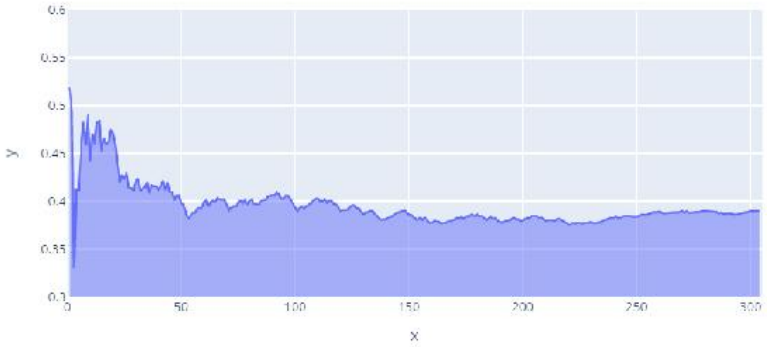


Gráfico mostrando a taxa de convergência para um artigo matemático não aprendido no modelo adequado. Vemos que 30 palavras foram suficientes para convergir para a pontuação nesse exemplo.

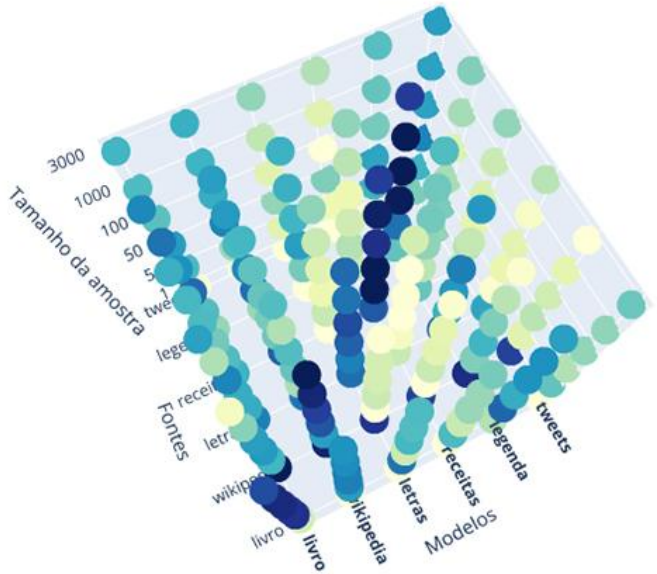


Gráfico mostrando que a convergência para os resultados reais acontecia por volta das 50 palavras. A versão interativa desse gráfico se encontra no notebook.

Aplicações + Extras

Análise de estilo de texto (tipo sentimentos): Muitas empresas usam ferramentas de software capazes de avaliar o sentimento predominante em um texto (desculpa, ordem, convite, etc) para avaliar se o texto passa a ideia desejada pela corporação.
Uma ferramenta baseada nessa experiência seria capaz de fazer uma "análise de modelo", e avaliar se o estilo escrito condiz o suficiente com o objetivado.

Avaliação de redações: Um dos blocos do Enem que mais evidencia a desigualdade é o da redação, pois a autavaliação dos resultados neste quesito é muito mais difícil que nas questões objetivas. Nessas situações, alunos com acesso à um corretor de redações particular têm uma vantagem substancial.
Uma solução para esse problema seria treinar nossa ferramenta com redações de diferentes pontuações, criando assim um método autmatizado de avaliar textos e tornando a correção muito mais acessível.

Ferramenta criativa: Com a recente abertura da I.A. de geração de imagens DALL-E, artistas do mundo todo viram na tecnologia um apoio criativo para suas produções.
De modo similar, podemos usar essa tecnologia para avaliar textos e proporcionar a escritores uma lista de estilos ou autores semelhantes. Dessa forma, ele será capaz de checar outros materiais similares aos que ele almeja atingir.