Trabalho final
Mineração de dados

Tuning do parâmetro de dificuldade das questões do ENEM utilizando PLN

Augusto Souza Nunes
Lucas Garzeri de Melo
Lucas Schimidt Coelho 11913019
Lucca Baptista da Silva Ferraz 13688134
Lua Nardi Quito 11371270

Sumário

INTRODUÇÃO
- Descrição do problema

SOBRE OS DADOS
- Dados utilizados e tratamentos feitos

ANÁLISE EXPLORATÓRIA
- Wordcloud
- BERTopic

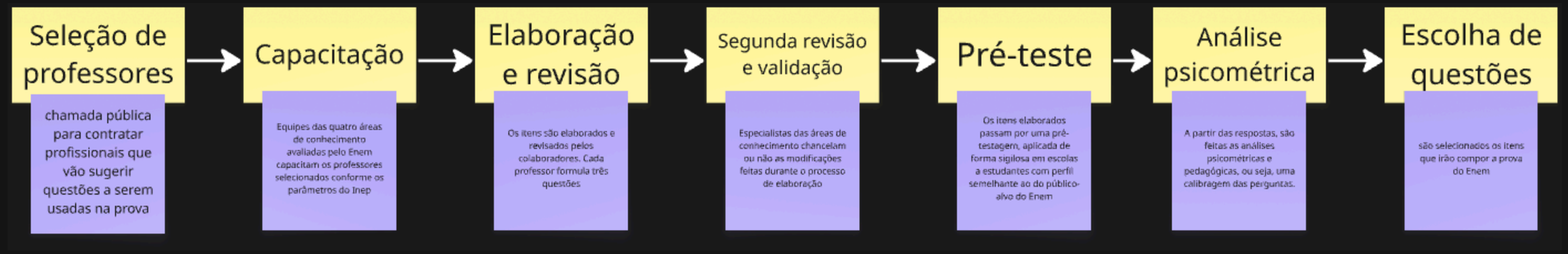
METODOLOGIAS
- Extração de features lingúisticos
- Word2Vec
- BERT
- Similaridade entre os embeddings
- Regressões

RESULTADOS

Introdução

Em 2009 o enem se tornou um exame cuja principal função é avaliar quais alunos devem ingressar nas faculdades - se tornando um vestibular nacional. E, por consequência, se tornou uma política educacional de enorme impacto social.

Processo de criação das questões:



- Alto custo do processo completo
- Custos adicionais com impressão, logística, aplicação e correção

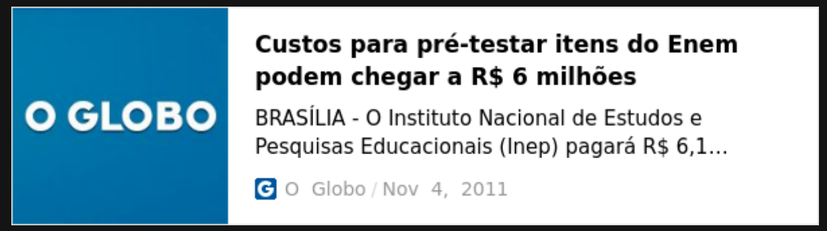
Pré-teste

2009

- Custo: R$ 939,5 mil
- 1 etapa
- 48 mil estudantes
- 10 municípios

2010

- Custo: R$ 6,1 milhões
- 4 etapas
- 100 mil estudantes
- 40 municípios



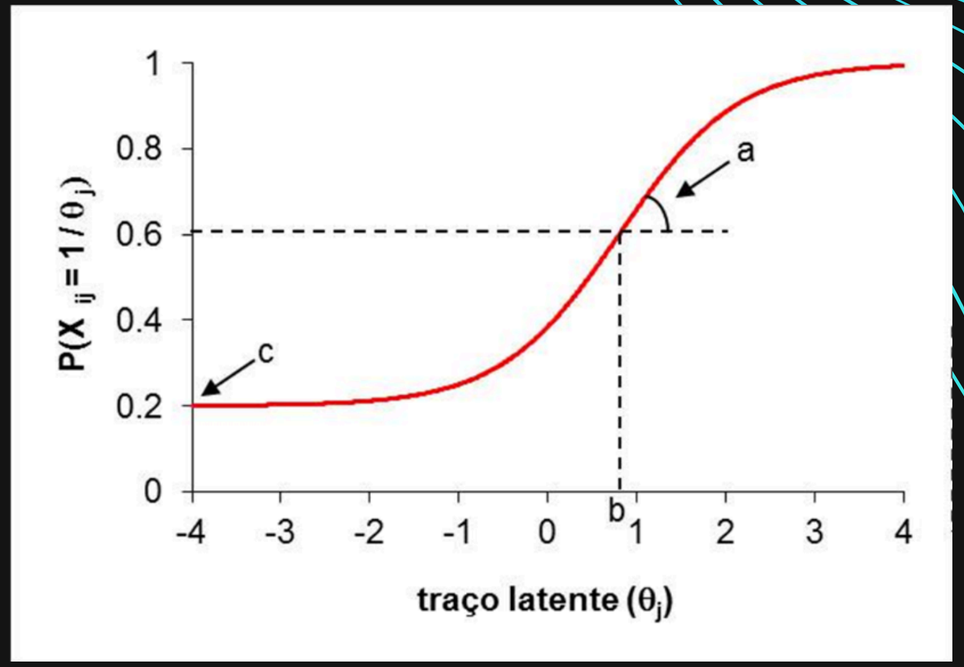
ML3: Modelo logístico de 3 parâmetros

$$P(X_{ij} = 1 \mid \theta_j, a_i, b_i, c_i) = c_i + \frac{(1 - c_i)}{1 + e^{-a_i(\theta_j - b_i)}}$$

- $i = 1, \ldots, I$ (itens)
- $j = 1, \ldots, n$ (indivíduos)
- $X_{ij} = 1$, se indiv j acerta o item i, e $X_{ij} = 0$, c.c.
- θ_j é o nível do traço latente do indiv j
- a_i parâmetro de discriminação do item i, derivada no ponto de inflexão
- b_i parâmetro de dificuldade do item i, se $b_i = \theta_j$, $P(X_{ij} = 1 \mid \theta_j, a_i, b_i, c_i) = \frac{(1+c_i)}{2}$
- c_i parâmetro de acerto ao acaso ("chute") do item i

Símbolo	Significado
X_{ij}	Resposta do aluno j ao item i: 1 = acerto, 0 = erro
θ_j	Habilidade (traço latente) do indivíduo j
a_i	Discriminação: quanto o item distingue entre alunos com diferentes níveis de habilidade
b_i	Dificuldade: nível de habilidade onde a chance de acerto é aproximadamente 50%
c_i	Probabilidade de acerto ao acaso (chute)

Table 1: Significados dos parâmetros do modelo logístico de 3 parâmetros (TRI)



Objetivo

- Verificar a capacidade de predizer a dificuldade dos itens de Ciências da Natureza a partir de suas características textuais, utilizando métodos computacionais.
- Analisar os métodos propostos no artigo e propor novos métodos
- Reduzir os custos da prova
- Aumentar a segurança do processo

É possível calibrar os itens do Enem sem pré-teste?

Sobre os dados

- Dados de questões de prova e dados dos itens das provas
- Anos de coleta 2009 - 2023
- 675 questões de CN
- 23 features

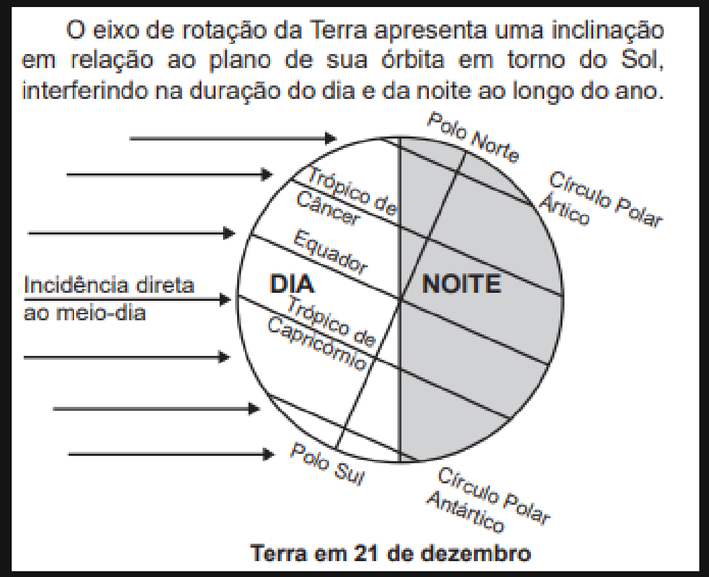
#	Column	Non-Null Count	Dtype
0	numero	675 non-null	float64
1	enunciado	675 non-null	object
2	A	675 non-null	object
3	B	675 non-null	object
4	C	675 non-null	object
5	D	675 non-null	object
6	E	675 non-null	object
7	CO_POSICAO	675 non-null	float64
8	SG_AREA	675 non-null	object
9	CO_ITEM	675 non-null	float64
10	TX_GABARITO	675 non-null	object
11	CO_HABILIDADE	630 non-null	float64
12	IN_ITEM_ABAN	675 non-null	float64
13	TX_MOTIVO_ABAN	16 non-null	object
14	NU_PARAM_A	659 non-null	object
15	NU_PARAM_B	659 non-null	float64
16	NU_PARAM_C	659 non-null	float64
17	TX_COR	675 non-null	object
18	CO_PROVA	675 non-null	float64
19	ANO	675 non-null	float64
20	TP_LINGUA	0 non-null	float64
21	IN_ITEM_ADAPTADO	450 non-null	float64
22	TP_VERSAO_DIGITAL	0 non-null	float64

Parâmetro de dificuldade da questão

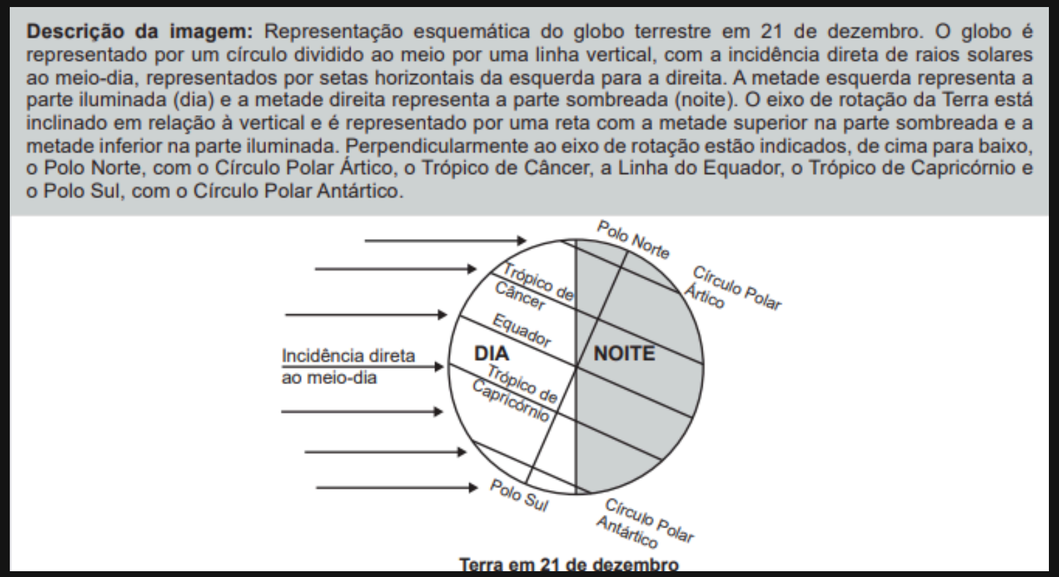
	numero	enunciado	A	B	C	D	E	CO_POSICAO	SG_AREA	CO_ITEM	...	TX_MOTIVO_ABAN	NU_PARAM_A	NU_PARAM_B	NU_PARAM_C	TX_COR	CO_PROVA	ANO
0	1.0	A atmosfera terrestre é composta pelos gases\n...	reduzir o calor irradiado pela Terra mediante ...	promover a queima da biomassa vegetal, respons...	reduzir o desmatamento, mantendo-se, assim, o\...	aumentar a concentração atmosférica de H2O,\nm...	remover moléculas orgânicas polares da atmosfe...	1.0	CN	60083.0	...	NaN	138076	-1.70677	14199.0	AZUL	49.0	2009.0
1	2.0	Considere que, no sistema de coordenadas xOy, ...	Concentração média de álcool no sangue ao long...	Variação da frequência da ingestão de álcool a...	Concentração mínima de álcool no sangue a part...	Estimativa de tempo necessário para metaboliza...	Representação gráfica da distribuição de frequ...	2.0	CN	60084.0	...	NaN	214163	0.62043	9837.0	AZUL	49.0	2009.0
2	3.0	Estima-se que haja atualmente no mundo 40\nmil...	induzir a imunidade, para proteger o organismo...	ser capaz de alterar o genoma do organismo por...	produzir antígenos capazes de se ligarem ao ví...	ser amplamente aplicada em animais, visto que ...	estimular a imunidade, minimizando a transmiss...	3.0	CN	60085.0	...	NaN	92339	2.07704	30865.0	AZUL	49.0	2009.0
3	4.0	Em um experimento, preparou-se um conjunto de\...	os genótipos e os fenótipos idênticos.	os genótipos idênticos e os fenótipos diferentes.	diferenças nos genótipos e fenótipos.	o mesmo fenótipo e apenas dois genótipos difer...	o mesmo fenótipo e grande variedade de genótipos.	4.0	CN	60086.0	...	NaN	123281	0.11500	34173.0	AZUL	49.0	2009.0
4	5.0	Na linha de uma tradição antiga, o astrônomo\n...	Ptolomeu apresentou as ideias mais valiosas, p...	Copérnico desenvolveu a teoria do heliocentris...	Copérnico viveu em uma época em que a pesquisa...	Kepler estudou o planeta Marte para atender às...	Kepler apresentou uma teoria científica que, g...	5.0	CN	60087.0	...	NaN	121819	0.21694	21795.0	AZUL	49.0	2009.0

Tratamento de questões com imagem

- Provas Ledor
- LLM’s para questões diferentes (ChatGPT, DeepSeek, Gemini)



Prova azul (sem acessibilidade)



Prova Laranja - Ledor (com acessibilidade)

Microdados

- Não utilizados em nossa análise

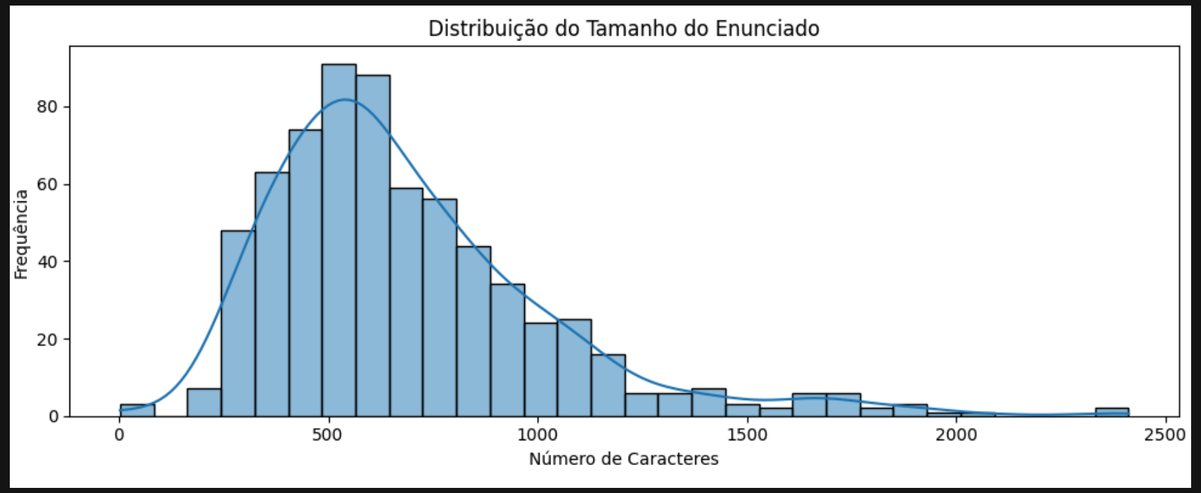
- Dados do participante (12)
- Dados da escola (7)
- Dados do local de aplicação da prova (4)
- Dados da prova objetiva (21)
- Dados da redação (7)
- Dados do questionário socioeconômico (25)

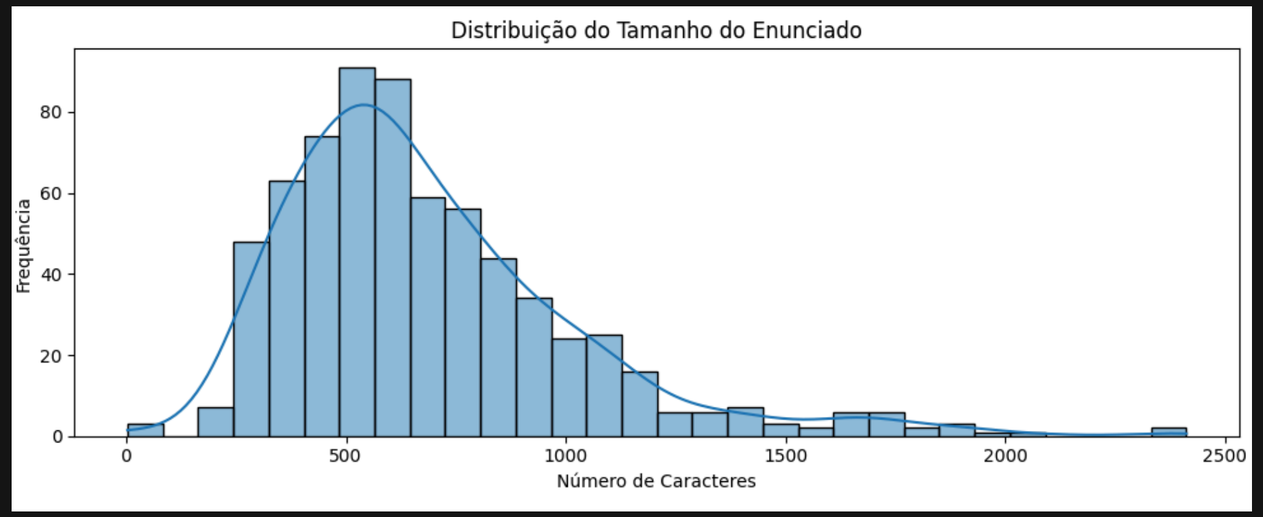
Ano	Número de inscritos (confirmados)
2024	4 325 960[1]
2023	3 933 970[2]
2022	3 409 682[3]
2021	4 004 764[4]
2020	5 783 357[5]
2019	5 095 308[6]
2018	5 513 662[7]
2017	6 731 186[8]
2016	8 627 371[8]
2015	7 792 025[8]
2014	8 722 290[8]
2013	7 173 574[9]
2012	5 791 332[9]
2011	5 380 857[9]
2010	4 626 094[9]
2009	4 148 721[9]

Estatísticas Descritivas

	Tam. Enunciado	Tam. Total Questão	NU_PARAM_B
Contagem	677.0	677.0	657.0
Média	688.35	892.96	123.27
Desvio Padrão	338.6	396.27	134.63
Mínimo	3.0	18.0	-639.0
1º Quartil (25%)	463.0	613.0	32.21
Mediana (50%)	610.0	804.0	120.77
3º Quartil (75%)	832.0	1072.0	179.36
Máximo	2412.0	3169.0	957.0

Histogramas





Teste de normalidade e Correlação

```
Teste de normalidade Shapiro-Wilk para NU_PARAM_B:
Estatística = 0.805
p-valor = 2.065e-27
→ A distribuição de NU_PARAM_B NÃO é normal.
```

```
📈 Correlação entre Tamanho Total e Dificuldade:
Correlação de Pearson: 0.027 (p-valor: 4.880e-01)
```

Questão mais fácil e mais difícil

- Parâmetro B: -1.80092
- Prova 2011

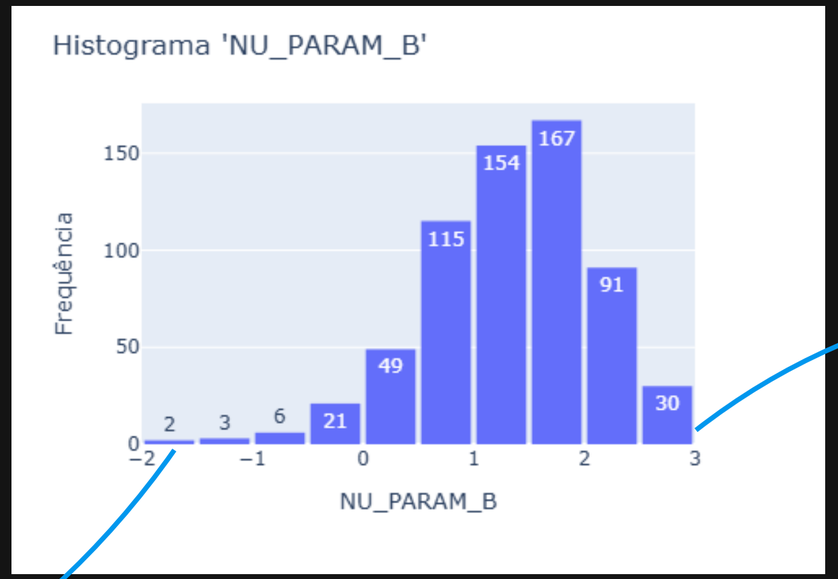
QUESTÃO 88

Certas espécies de algas são capazes de absorver rapidamente compostos inorgânicos presentes na água, acumulando-os durante seu crescimento. Essa capacidade fez com que se pensasse em usá-las como biofiltros para a limpeza de ambientes aquáticos contaminados, removendo, por exemplo, nitrogênio e fósforo de resíduos orgânicos e metais pesados provenientes de rejeitos industriais lançados nas águas. Na técnica do cultivo integrado, animais e algas crescem de forma associada, promovendo um maior equilíbrio ecológico.

SORIANO, E. M. Filtros vivos para limpar a água. **Revista Ciência Hoje**. V. 37, n° 219, 2005 (adaptado).

A utilização da técnica do cultivo integrado de animais e algas representa uma proposta favorável a um ecossistema mais equilibrado porque

A) os animais eliminam metais pesados, que são usados pelas algas para a síntese de biomassa.
B) os animais fornecem excretas orgânicos nitrogenados, que são transformados em gás carbônico pelas algas.
C) as algas usam os resíduos nitrogenados liberados pelos animais e eliminam gás carbônico na fotossíntese, usado na respiração aeróbica.
D) as algas usam os resíduos nitrogenados provenientes do metabolismo dos animais e, durante a síntese de compostos orgânicos, liberam oxigênio para o ambiente.
E) as algas aproveitam os resíduos do metabolismo dos animais e, durante a quimiossíntese de compostos orgânicos, liberam oxigênio para o ambiente.



- Parâmetro B: 2.99298
- Prova 2009

Questão 23

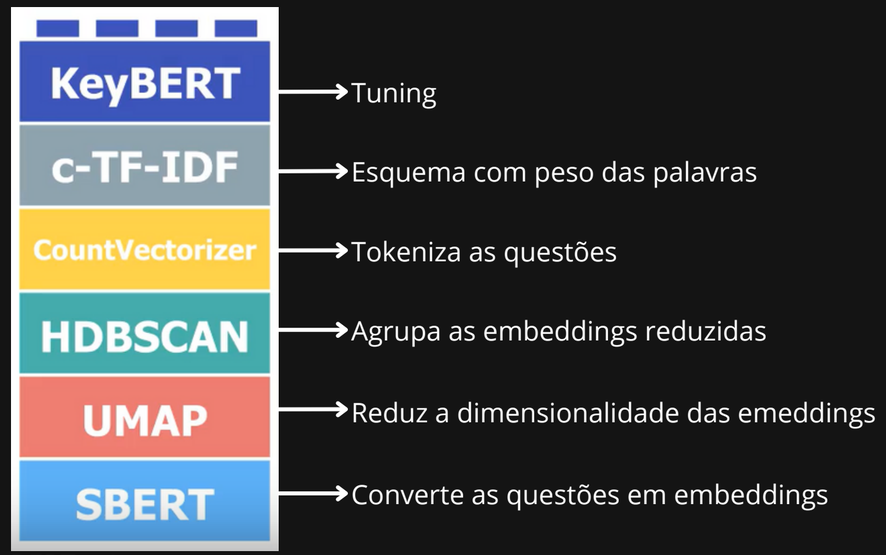
Cerca de 1% do lixo urbano é constituído por resíduos sólidos contendo elementos tóxicos. Entre esses elementos estão metais pesados como o cádmio, o chumbo e o mercúrio, componentes de pilhas e baterias, que são perigosos à saúde humana e ao meio ambiente. Quando descartadas em lixos comuns, pilhas e baterias vão para aterros sanitários ou lixões a céu aberto, e o vazamento de seus componentes contamina o solo, os rios e o lençol freático, atingindo a flora e a fauna. Por serem bioacumulativos e não biodegradáveis, esses metais chegam de forma acumulada aos seres humanos, por meio da cadeia alimentar. A legislação vigente (Resolução CONAMA n° 257/1999) regulamenta o destino de pilhas e baterias após seu esgotamento energético e determina aos fabricantes e/ou importadores a quantidade máxima permitida desses metais em cada tipo de pilha/bateria, porém o problema ainda persiste.

Disponível em: http://www.mma.gov.br. Acesso em: 11 jul. 2009 (adaptado).

Uma medida que poderia contribuir para acabar definitivamente com o problema da poluição ambiental por metais pesados relatado no texto seria

A) deixar de consumir aparelhos elétricos que utilizem pilha ou bateria como fonte de energia.
B) usar apenas pilhas ou baterias recarregáveis e de vida útil longa e evitar ingerir alimentos contaminados, especialmente peixes.
C) devolver pilhas e baterias, após o esgotamento da energia armazenada, à rede de assistência técnica especializada para repasse a fabricantes e/ou importadores.
D) criar nas cidades, especialmente naquelas com mais de 100 mil habitantes, pontos estratégicos de coleta de baterias e pilhas, para posterior repasse a fabricantes e/ou importadores.
E) exigir que fabricantes invistam em pesquisa para a substituição desses metais tóxicos por substâncias menos nocivas ao homem e ao ambiente, e que não sejam bioacumulativas.

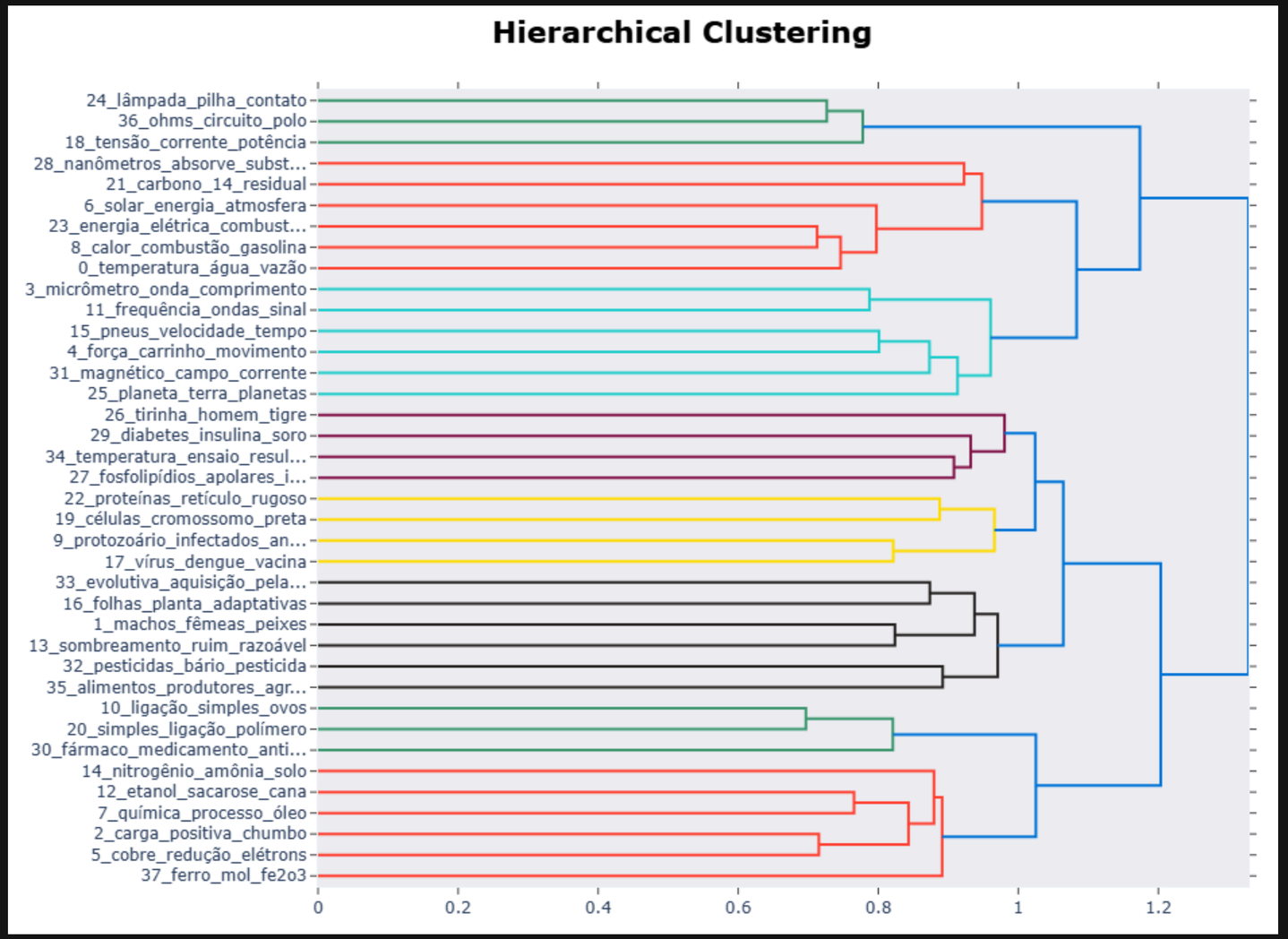
BERTopic

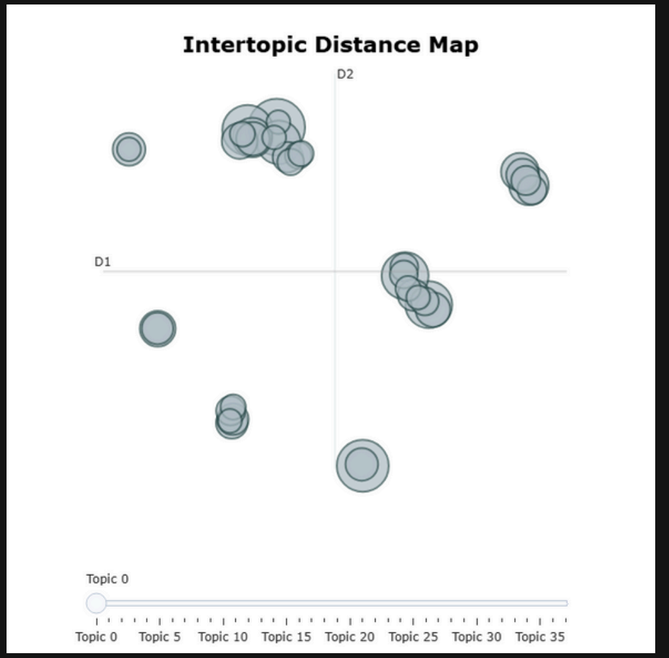


BERTopic

- Grid Search para parâmetros do UMAP e HDBSCAN
  - n_neighbors_list = [5, 10, 15, 30]
  - min_dist_list = [0.0, 0.1, 0.3]
  - min_cluster_size_list = [5, 8, 10]
  - min_samples_list = [1, 2, 5]

Topic	Count	Name	Representation	Representative_Docs
-1	59	-1_água_anticorpos_membrana_célula	[água, anticorpos, membrana, célula, glicose, ...	[moedas despertam interesse colecionadores num...
0	45	0_temperatura_água_vazão_umidade	[temperatura, água, vazão, umidade, relativa, ...	[pessoa lendo manual ducha acabou adquirir cas...
1	38	1_machos_fêmeas_peixes_animais	[machos, fêmeas, peixes, animais, espécies, vi...	[hermaphroditic demasculinized frogs after exp...
2	36	2_carga_positiva_chumbo_reação	[carga, positiva, chumbo, reação, concreto, ca...	[aplicações ambientais persulfato remediação á...


Hierarchical Clustering
24_lâmpada_pilha_contato
36_ohms_circuito_polo
18_tensão_corrente_potência
28_nanômetros_absorve_subst...
21_carbono_14_residual
6_solar_energia_atmosfera
23_energia_elétrica_combust...
8_calor_combustão_gasolina
0_temperatura_água_vazão
3_micrômetro_onda_comprimento
11_frequência_ondas_sinal
15_pneus_velocidade_tempo
4_força_carrinho_movimento
31_magnético_campo_corrente
25_planeta_terra_planetas
26_tirinha_homem_tigre
29_diabetes_insulina_soro
34_temperatura_ensaio_resul...
27_fosfolipídios_apolares_i...
22_proteínas_retículo_rugoso
19_células_cromossomo_preta
9_protozoário_infectados_an...
17_vírus_dengue_vacina
33_evolutiva_aquisição_pela...
16_folhas_planta_adaptativas
1_machos_fêmeas_peixes
13_sombreamento_ruim_razoável
32_pesticidas_bário_pesticida
35_alimentos_produtores_agr...
10_ligação_simples_ovos
20_simples_ligação_polímero
30_fármaco_medicamento_anti...
14_nitrogênio_amônia_solo
12_etanol_sacarose_cana
7_química_processo_óleo
2_carga_positiva_chumbo
5_cobre_redução_elétrons
37_ferro_mol_fe2o3
0
0.2
0.4
0.6
0.8
1
1.2


Intertopic Distance Map
D2
D1
Topic 0
Topic 0
Topic 5
Topic 10
Topic 15
Topic 20
Topic 25
Topic 30
Topic 35

Metodologias

- Análise de atributos linguísticos
- Modelos contextuais
- Similaridade de texto
- Fine tuned BERT

Extração de atributos linguísticos

A fim de iniciar a análise exploratória, foram extraídos atributos textuais dos enunciados das questões. As métricas aqui calculadas se baseiam no artigo do NILC e avaliam características como inteligibilidade, complexidade textual, coesão e coerência. Todas as métricas foram adaptadas às nuances da língua portuguesa e foram implementadas à partir da biblioteca spaCy, própria para análise morfossintática.

NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese

NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese

Sidney Evaldo Leal [1] sidleal@gmail.com
Magali Sanches Duran [1] magali.duran@gmail.com
Carolina Evaristo Scarton [2] c.scarton@sheffield.ac.uk
Nathan Siegle Hartmann [3] nathanshartmann@gmail.com
Sandra Maria Aluísio [1] sandra@icmc.usp.br

[1] Instituto de Ciências Matemáticas e de Computação - University of São Paulo, São Paulo, Brazil
[2] The University of Sheffield, Sheffield, UK
[3] Itaú Unibanco, São Paulo, Brazil

Abstract

This paper presents and makes publicly available the NILC-Metrix, a computational system comprising 200 metrics proposed in studies on discourse, psycholinguistics, cognitive and computational linguistics, to assess textual complexity in Brazilian Portuguese (BP). These metrics are relevant for descriptive analysis and the creation of computational models and can be used to extract information from various linguistic levels of written and spoken language. The metrics in NILC-Metrix were developed during the last 13 years, starting in 2008 with Coh-Metrix-Port, a tool developed within the scope of the PorSimples project. Coh-Metrix-Port adapted some metrics to BP from the Coh-Metrix tool that computes metrics related to cohesion and coherence of texts in English. After the end of PorSimples in 2010, new metrics were added to the initial 48 metrics of Coh-Metrix-Port. Given the large number of metrics, we present them following an organisation similar to the metrics of Coh-Metrix v3.0 to facilitate comparisons made with metrics in Portuguese and English. In this paper, we illustrate the potential of NILC-Metrix by presenting three applications: (i) a descriptive analysis of the differences between children's film subtitles and texts written for Elementary School I[1] and II (Final Years)[2]; (ii) a new predictor of textual complexity for the corpus of original and simplified texts of the PorSimples project; (iii) a complexity prediction model for school grades, using transcripts of children's story narratives told by teenagers. For each application, we evaluate which groups of metrics are more discriminative, showing their contribution for each task.

1 Introduction

A set of metrics called NILC-Metrix was developed both in funded projects, involving multiple re-

Descrição das métricas

Distribuição e estrutura lexical

Type-Token Ratio
→ Diversidade lexical (palavras únicas/total).

Nº de sentenças
→ Fragmentação do texto.

Tamanho médio das sentenças
→ Indicador de complexidade sintática.

Distribuição morfossintática

Verb-Token Ratio
→ Proporção de verbos.

Noun-Token Ratio
→ Proporção de substantivos.

Adjective-Token Ratio
→ Proporção de adjetivos.

Pronoun-Token Ratio
→ Proporção de pronomes

Fluidez e formalidade

Stopword-Token Ratio
→ Palavras sem conteúdo semântico.

Pausality
→ Proporção de pontuação (indica pausas).

Informalidade
→ Estimada por erros ortográficos

Complexidade e semântica textual

Índice de Brunet
→ Avalia diversidade lexical ajustada ao tamanho.

Gunning-Fog Index
→ Mede inteligibilidade; >12 indica alta complexidade.

Especificidade
→ Proporção de entidades nomeadas (tempo/espaço).

Modelos de linguagem

Arquiteturas estáticas

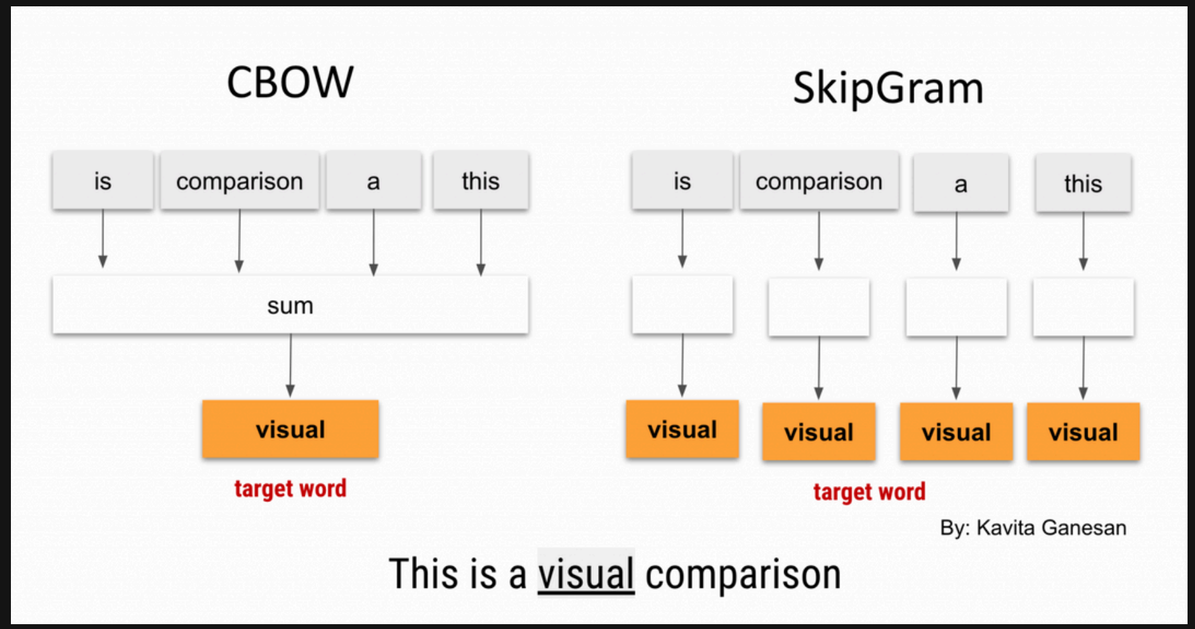
Geram uma representação vetorial fixa para as palavras, independentemente do contexto em que são inseridas.

Arquiteturas contextuais

Baseadas em mecanismos de autoatenção, geram diferentes representações vetoriais das palavras de acordo com o contexto.

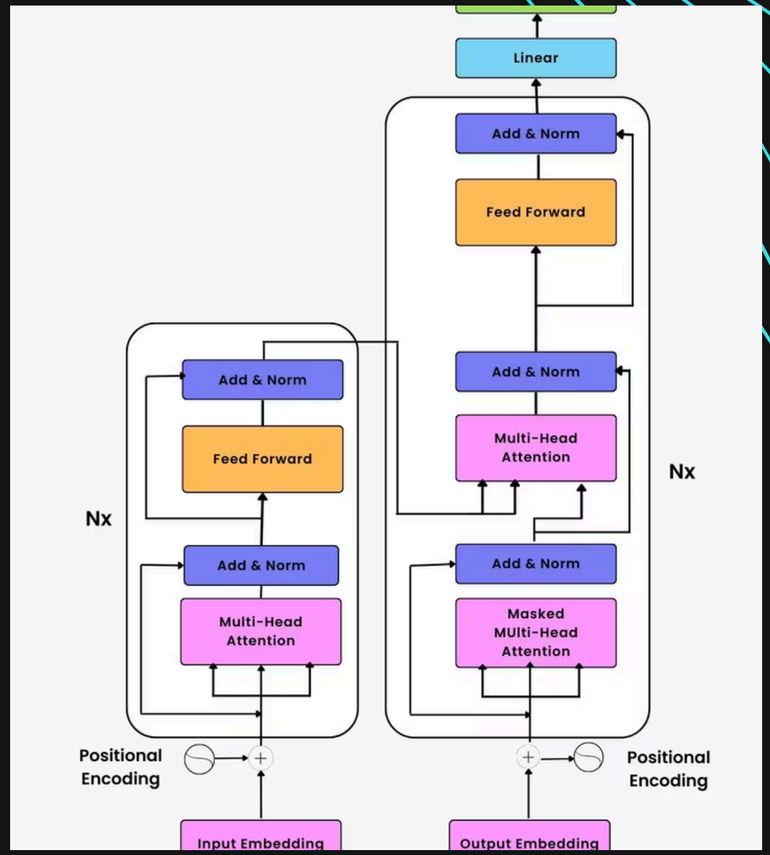
Word2VEC

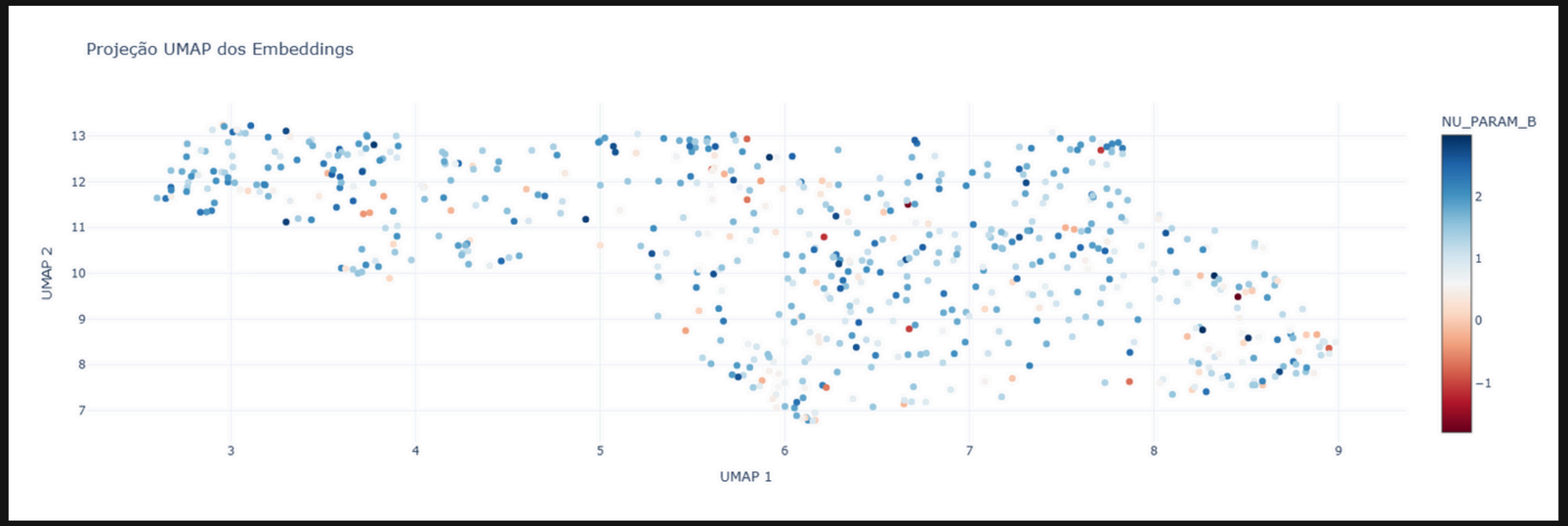
NILC embeddings



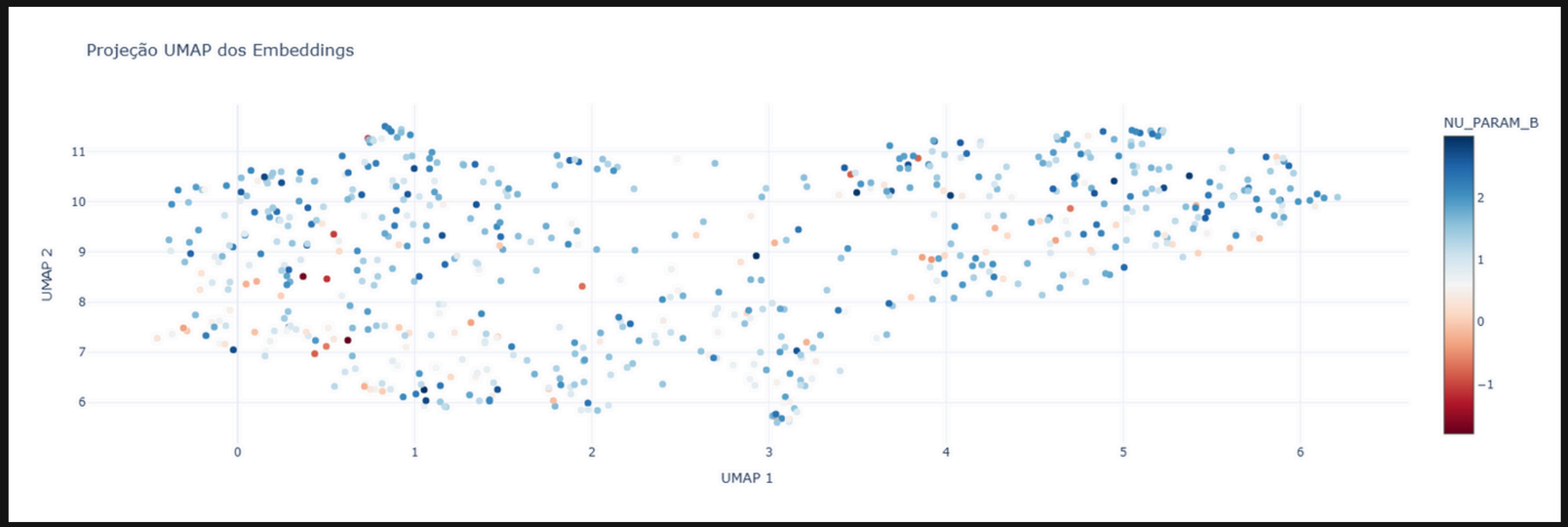
Kavita-ganesan

BERT - Bidirectional encoding representations from transformers



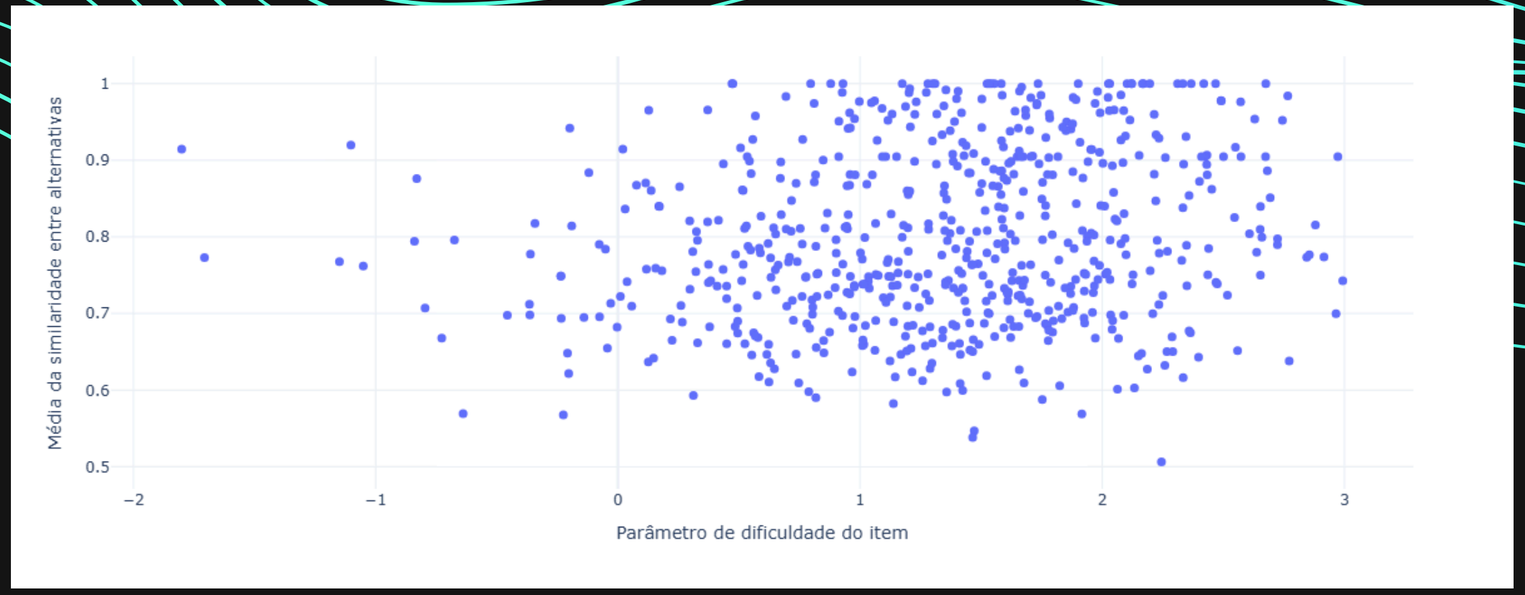


Word2VEC embeddings



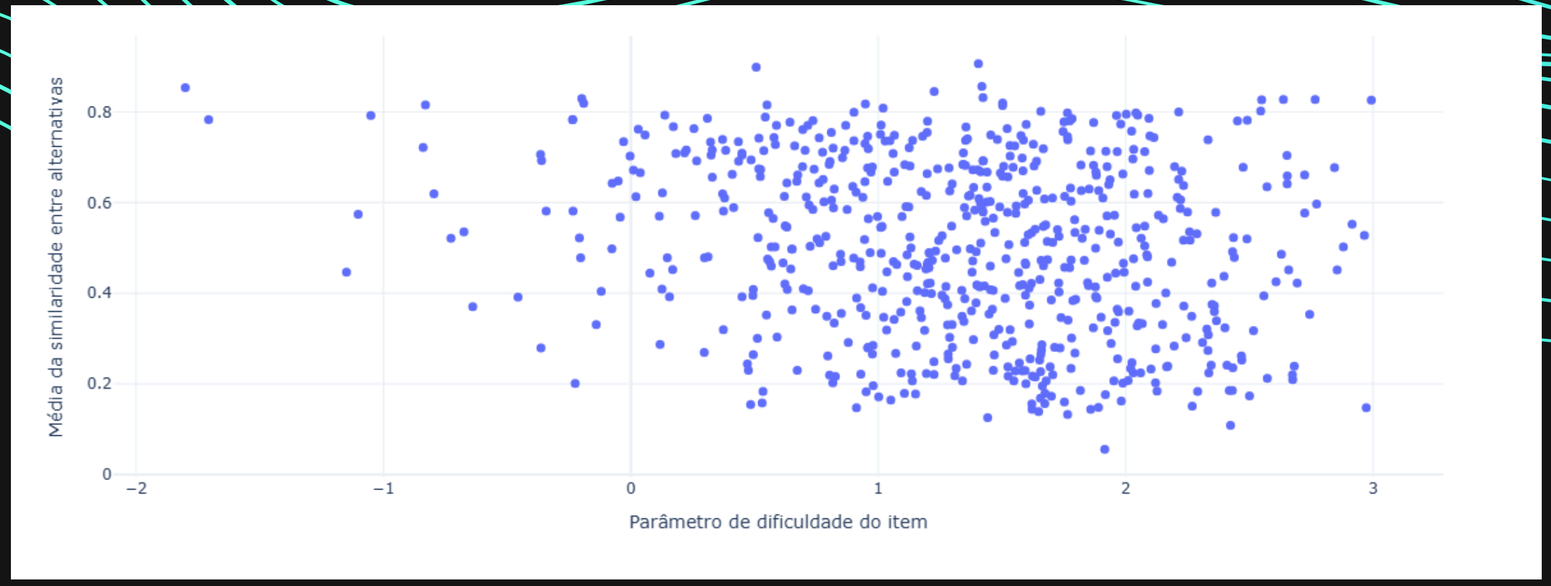
BERT embeddings

Similaridade entre os embeddings



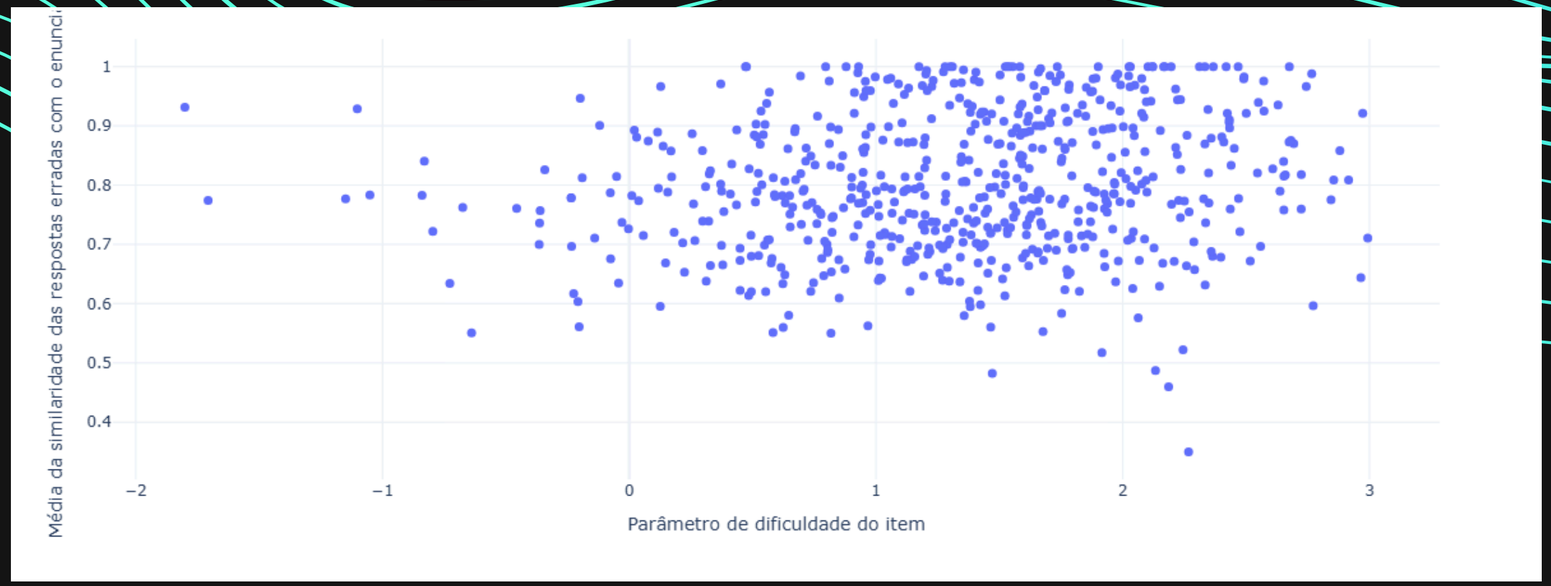
Correlação entre similaridade média dos Embeddiings das alternativas e o parâmetro de dificuldade: 0.17752

Similaridade entre os embeddings



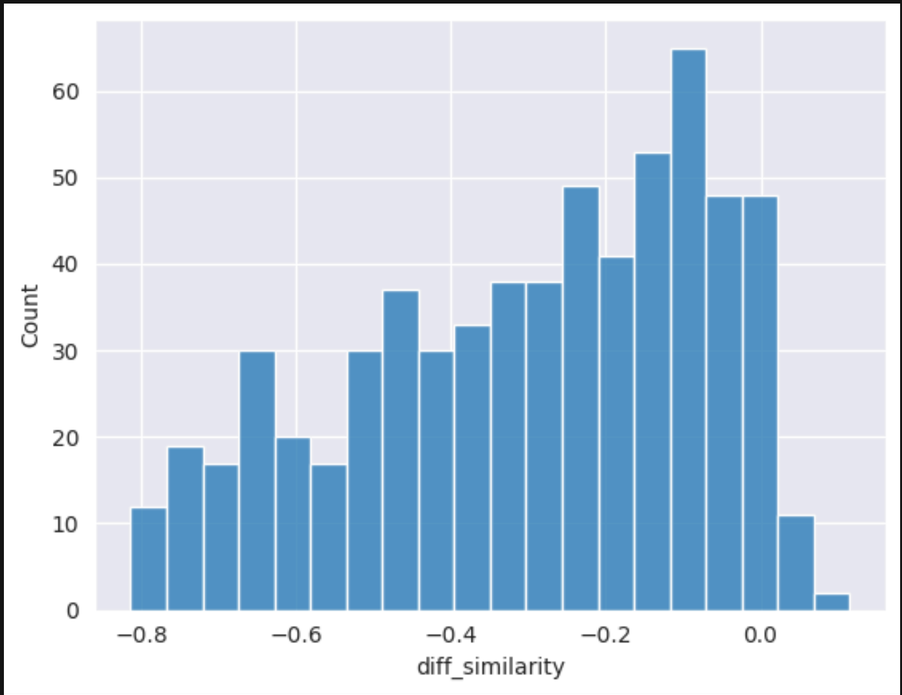
Correlação entre similaridade dos embeddings gabarito/enunciado e o parâmetro de dificuldade: -0.2052

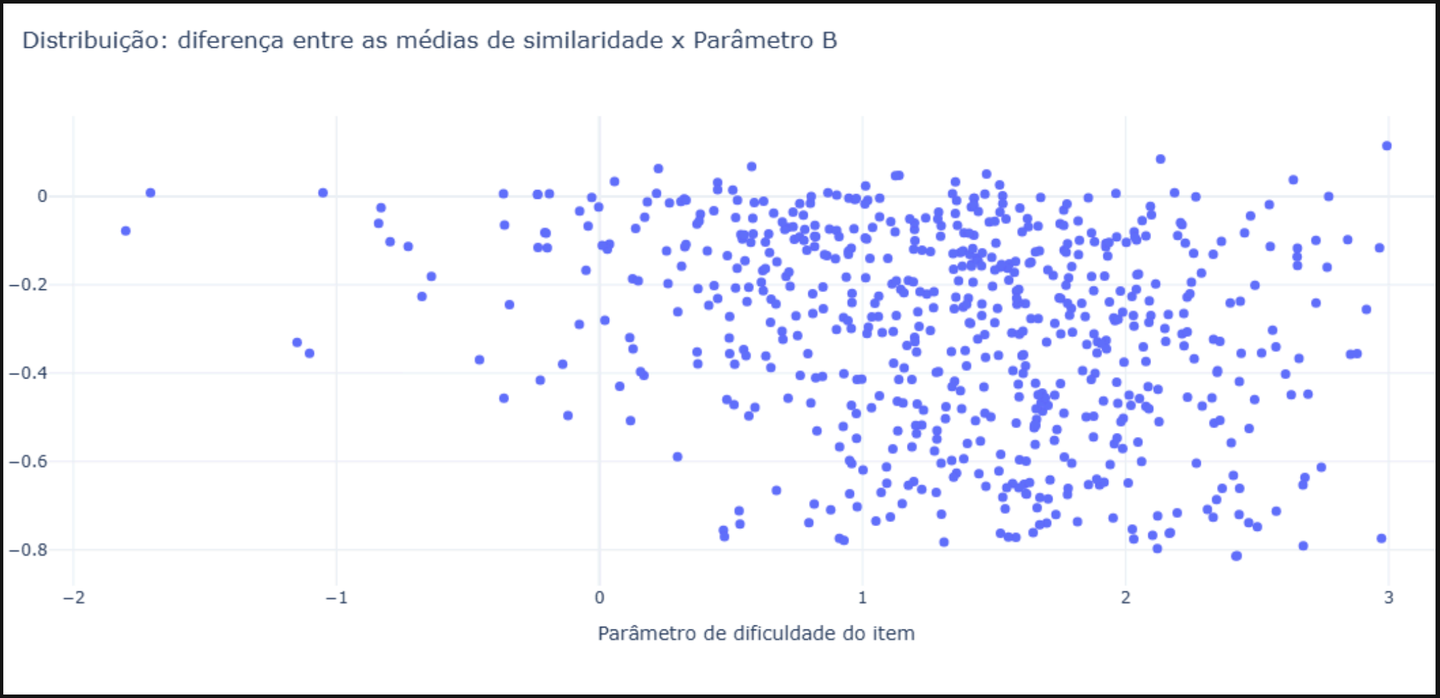
Similaridade entre os embeddings



Correlação entre similaridade dos embeddings respostas erradas/ enunciado e o parâmetro de dificuldade: 0.1598

Diferença entre as similaridades enunciado/gabarito e enunciado/respostas erradas





Regressores

Metodologia

Extrair as representações vetoriais das palavras do enunciado e “mapeá-lo’ semanticamente, utilizando das dimensões dos embeddings como variáveis independentes de um modelo de regressão.

Modelos de linguagem

Foram utilizados os modelos Word2vec CBOW e Skip-Gram com 50, 100 e 300 dimensões disponibilizados no repositório de embeddings do NILC. Utilizamos, também, para efeito de comparação, os embeddings gerados pela arquitetura BERT nos regressores.

Modelos de regressão

Linear e Lasso.

Pré-processamento

Treino e teste

Para cada ajuste de regressão, foram usados divisões treino-teste padronizadas, permitindo que monitorássemos as métricas de avaliação.

Resultados parciais

CBOW - 300 dimensões

Seguindo o protocolo de vetorização

Regressão linear

Dados de validação:
MSE: 1,208
Correlação: 0,265

Regressão lasso

Dados de validação:
MSE: 0,581
Correlação: 0,4192

CBOW - 300 dimensões

Sem o protocolo de vetorização

Regressão linear

Dados de validação:
MSE: 1,101
Correlação: 0,295

Regressão lasso

Dados de validação:
MSE: 0,610
Correlação: 0,389

BERT embeddings

Regressão linear

Dados de validação:
MSE: 1.899
Correlação: 0,12

Regressão lasso

Dados de validação:
MSE: 0,594
Corerlação: 0,41

Fine tuned BERT

Reinicialização das últimas camadas

Otimização de hiperparâmetros

Diferentes modelos

Resultados

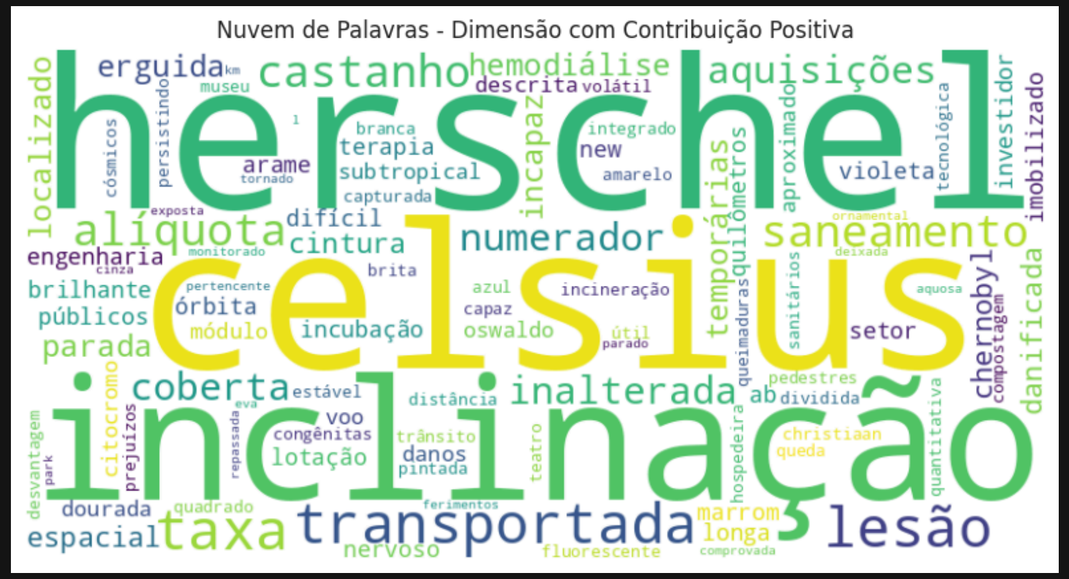
BERTimbau

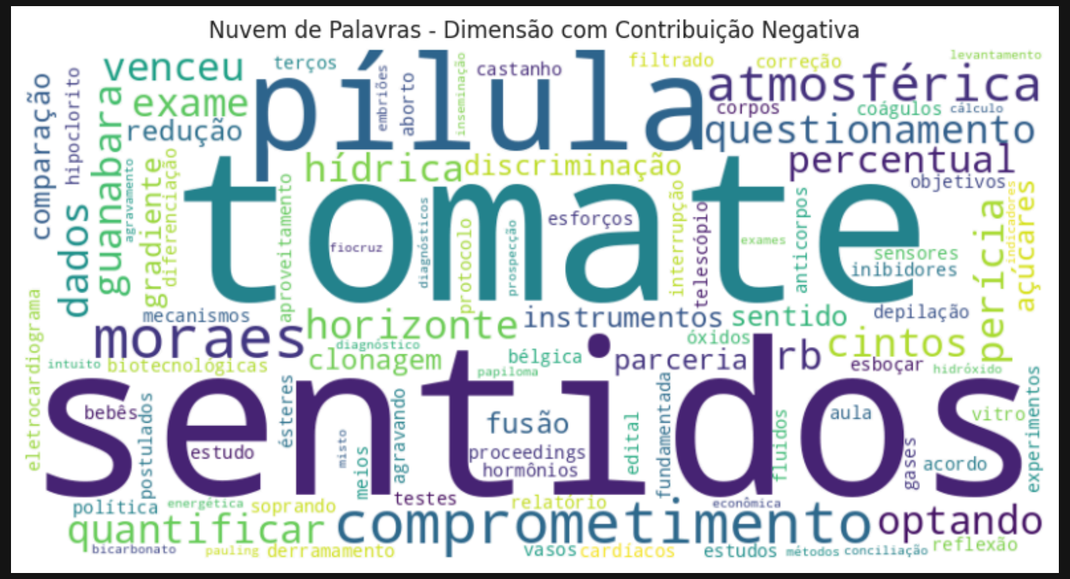
Dados de validação:
MSE: 0,537
Corerlação: 0,473

Multilingual BERT

Dados de validação:
MSE: 0,552
Corerlação: 0,447

Análise das dimensões da regressão





Próximos passos

Utilizar métodos computacionais para outras etapas do processo?

Muito obrigado!