Análise das Correlações entre Frequências de Palavras no CNMAC e Disponibilidade do ChatGPT-3.5

Osmar Cardoso Lopes Filho
osmarclopesfilho@gmail.com
GitHub

Introdução

A aplicação de inteligências artificiais generativas na escrita é capaz de acelerar e facilitar a produção de materiais científicos. Contudo, o uso dessa ferramenta pode arcar com a introdução de artefatos no texto que podemos analisar estatisticamente. Nesse trabalho, faremos essa analise sobre a frequência de palavras dos artigos completos submetidos ao CNMAC nos anos 2017, 2018, 2019, 2021, 2022 e 2023.

Inteligências artificiais generativas (GAI do inglês *generative artificial intelligence*) utilizam modelos estatísticos generativos para criar conteúdo, como texto, imagens, audio, etc. Essa geração é resultado do treinamento da GAI, onde os parâmetros de seu modelo são ajustados de acordo com o conteúdo para treinamento. Por exemplo, uma GAI treinada sobre uma base de código de uma certa linguaguem de programação pode ser capaz de gerar outros códigos na mesma linguagem baseados nas estruturas que ela absorveu do conteúdo de treinamento.

O Congresso Nacional de Matemática Aplicada e Computacional (CNMAC) possui edições desde 2014 e, a partir de 2017, seu site apresenta-se sob uma única padronização, facilitando a obtenção através de *web scrapping* dos artigos submetidos. No total, o CNMAC possui 390 artigos completos submetidos nos anos de 2017, 2018 e 2019 e 415 artigos nos anos de 2021, 2022 e 2023.

A separação dos anos antes de depois de 2020 é feita devido ao seguinte marco: a disponibilização ao público do ChatGPT-3.5. Essa AI, baseada em *large language models* (LLMs), *aprende* relações estatísticas dentre o conteúdo de treino e utiliza-as para prever continuações de sequências de texto, gerando material escrito.

Análise Exploratória

Os dados coletados (aproximadamente 20% dos artigos) totalizam 24164 palavras distintas ao longo dos 6 anos supracitados. Dentre elas, procuramos por palavras que exibem um baixo uso nos anos anteriores à disponibilização do ChatGPT-3.5 e um aumento consideravel nos anos seguintes. Tal comportamento pode constituir um artefato deixado no corpus pela geração de trechos de texto por AI.

Para garantir que não existem interferências entre as frequências e o número de artigos publicados em um certo ano, trabalhamos com valores escalados w das contagens de palavras c:

$$w = \left\lfloor \frac{1000c}{y} \right\rfloor$$

Onde y é o número de artigos no respectivo ano. O multiplicador 1000 tem o único propósito de manter os números representáveis computacionalmente como inteiros.

Dada a extensa quantidade de palavras, uma verificação manual dos dados é impossível. Consideramos então alguns subconjuntos do *dataset* que possuem palavras de maior valor para nossa análise. Seja w_1 a média das frequências anteriores ao ChatGPT-3.5 de uma certa palavra w e w_2 a média das posteriores. A partir disso, construimos um filtro sobre o *dataset*:

$$S_1 = \{w \mid f(w_1, w_2) > k\}$$

$$f(w_1, w_2) = \begin{cases} \left\lfloor 100\frac{w_2}{w_1} \right\rfloor & \Longleftarrow w_1 \neq 0 \\ 100w_2 & \Longleftarrow w_1 = 0 \end{cases}$$

Onde k é um parametro de nossa escolha.

Esse subconjunto possui palavras que sofreram algum aumento na frequência média entre os anos pré- e pós-ChatGPT-3.5.

Row	word String	countprev Int64	countpost Int64	countratio Int64	count17 Int64	count18 Int64	count19 Int64	count21 Int64	count22 Int64	count23 Int64
1	equação	0	685	68500	0	0	0	64	200	421
2	parâmetros	0	598	59800	0	0	0	23	295	280
3	possível	0	517	51700	0	0	0	0	219	298
4	equações	0	504	50400	0	0	0	40	333	131
5	variáveis	0	502	50200	0	0	0	0	371	131
6	urs	0	482	48200	0	0	0	0	0	482
7	população	0	472	47200	0	0	0	0	209	263
8	grãos	0	469	46900	0	0	0	0	285	184
9	sir	0	447	44700	0	0	0	163	276	8
10	funções	0	441	44100	0	0	0	0	257	184
11	média	0	406	40600	0	0	0	0	152	254
12	proposição	0	368	36800	0	0	0	0	342	26
13	cultivares	0	364	36400	0	0	0	29	28	307
14	aproximação	0	362	36200	0	0	0	0	47	315
15	condições	0	355	35500	0	0	0	5	219	131
16	ffm	0	353	35300	0	0	0	35	161	157
17	até	0	344	34400	0	0	0	0	152	192
18	st	0	338	33800	0	0	0	58	9	271
19	cálculo	0	337	33700	0	0	0	0	171	166
20	irregularity	0	333	33300	0	0	0	0	0	333
21	mensagem	0	328	32800	0	0	0	52	276	0
22	vírus	0	318	31800	0	0	0	0	266	52
23	canal	0	314	31400	0	0	0	23	28	263
24	disk	0	312	31200	0	0	0	76	0	236
25	pandemia	0	307	30700	0	0	0	116	95	96

Figure 1: Palavras com maiores aumentos na média. A ordenação é feita pela coluna `countratio`.

Obtemos também um conjunto onde são ignorados os casos em que a média anterior ao ChatGPT-3.5 é 0.

Row	word String	countprev Int64	countpost Int64	countratio Int64	count17 Int64	count18 Int64	count19 Int64	count21 Int64	count22 Int64	count23 Int64
1	também	8	615	7687	0	8	0	5	295	315
2	distribuição	8	543	6787	0	8	0	0	114	429
3	matemática	16	1079	6743	0	16	0	5	504	570
4	três	8	447	5587	0	8	0	0	228	219
5	orientada	5	276	5520	5	0	0	0	276	0
6	solução	31	1643	5300	23	8	0	70	828	745
7	sector	8	412	5150	0	8	0	0	0	412
8	agricultural	7	334	4771	0	0	7	17	19	298
9	alunos	15	694	4626	0	0	15	146	171	377
10	aluno	5	230	4600	5	0	0	46	123	61
11	números	16	734	4587	0	16	0	0	647	87
12	está	8	353	4412	0	8	0	0	161	192
13	através	8	340	4250	0	8	0	0	200	140
14	ρ	7	294	4200	0	0	7	17	85	192
15	além	16	671	4193	0	16	0	5	333	333
16	virtual	8	305	3812	0	8	0	251	28	26
17	após	8	283	3537	0	8	0	0	161	122
18	curves	12	382	3183	5	0	7	11	371	0
19	lagrangian	5	156	3120	5	0	0	23	133	0
20	ms	8	238	2975	0	8	0	0	28	210
21	métodos	8	236	2950	0	8	0	0	114	122
22	online	11	314	2854	11	0	0	40	38	236
23	distributions	8	226	2825	0	8	0	23	19	184
24	atraso	5	141	2820	5	0	0	0	133	8
25	convexa	5	141	2820	5	0	0	46	95	0

Figure 2: Palavras com `countprev` não nulo ordenadas por `countratio`.

Além disso, criamos também um subconjunto com palavras que decairam em uso:

$$S_2 = \{w \mid g(w_1, w_2) > k\}$$

$$g(w_1, w_2) = \begin{cases} \left\lfloor 100\frac{w_1}{w_2} \right\rfloor & \Longleftarrow w_2 \neq 0 \\ 100w_1 & \Longleftarrow w_2 = 0 \end{cases}$$

A partir disso, podemos ordenar as palavras de duas formas. Desenhamos, então, os gráficos das contagens de ambas as médias de cada palavras; um deles com as palavras ordenadas de acordo com $f(w_1, w_2)$ e o outro com $g(w_1, w_2)$:

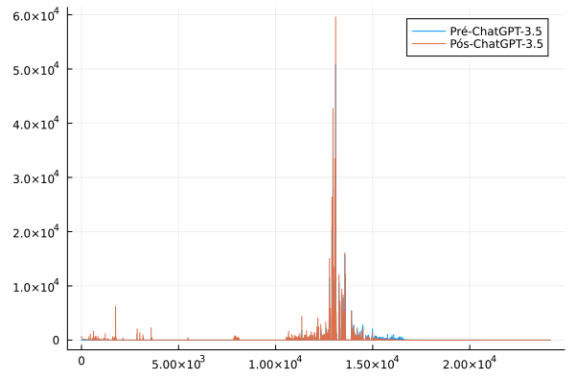

Figure 3: Contagens ordenadas por f.

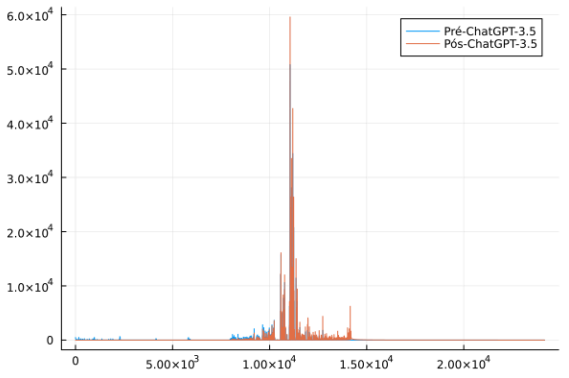

Figure 4: Contagens ordenadas por g.

Nota-se com isso a existência de picos a esquerda das contagens ordenadas por f. Tomamos isso como indicativo de que possuimos apenas palavras com crescimento extremo, nenhuma com decrescimento extremo.

Prosseguimos, então, analisando o crescimento do uso das palavras.

Metodologia

Uma vez que temos um subconjunto de palavras de interesse, ajustamos um modelo linear nas três primeiras contagens e analisaremos os resíduos obtidos nas últimas contagens. Isso nos informa de maneira mais refinada quais palavras cresceram em uso.

Para realizar esse processo, tomamos as mil primeiras palavras ordenadas por f e ajustamos os modelos. A partir deles, ordenamos elas pelo somatório dos resíduos das três últimas contagens. Caso seja positivo, esse somatório indíca o crescimento do uso da palavra em relação a melhor predição linear que obtivemos.

Tomando o resultado disso, construimos o seguinte gráfico:

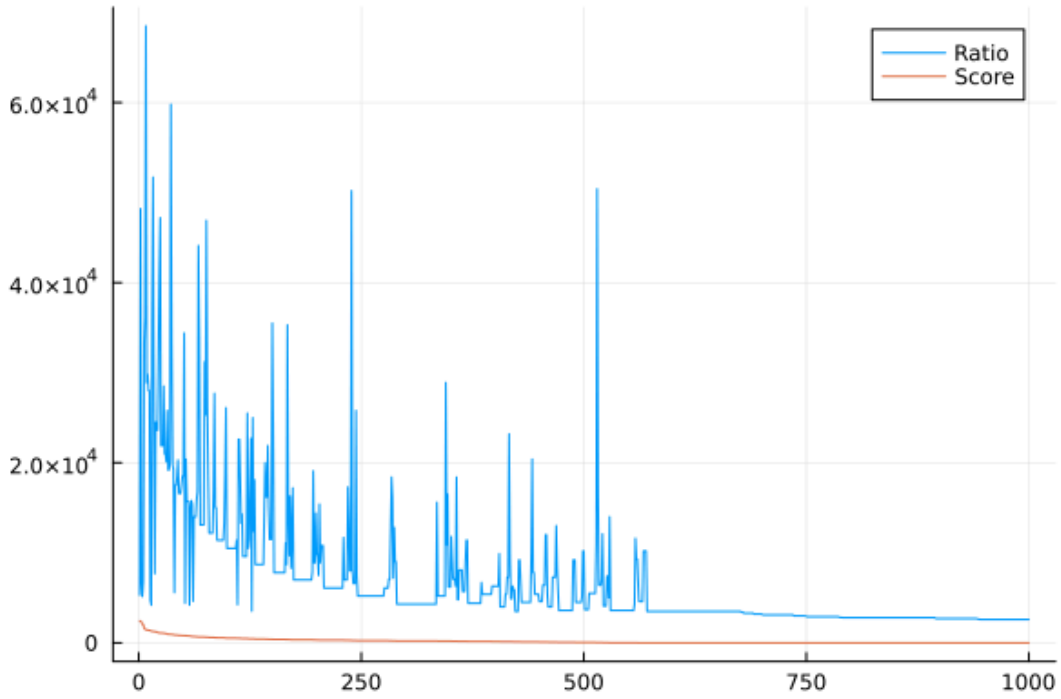

Figure 5: Somatório dos últimos resíduos (score) e valores dados por f (ratio).

Nota-se que os valores de `ratio` seguem o aumento dos valores de `score` salvo casos espontâneos.

Feito isso, tomamos as 250 palavras com maiores resíduos e ajustaremos uma curva logística sobre cada uma das contagens. O objetivo disso é determinar o quão bem essa curva, cujo centro coincide com o ano de lançamento do ChatGPT-3.5, modela nossos dados. O primeiro passo para alcançar isso é determinar os "valores de 0 e 1" que a nossa variável resposta deve tomar. Faremos isso tomando a média das primeiras e últimas contagens de cada palavra e escalando os valores de contagem de modo que a média das últimas contagens seja 1 e das primeiras seja 0.

Feito isso, selecionamos as palavras com as menores somas dos quadrados das diferênças entre suas contagens e os valores previstos pela logística. A seguir, listamos as 20 palavras com menor soma, todas com um somatório dos quadrados entre contragem e previsão menor que 1.6:

Alunos, solução, será, parâmetros, equação, nível, considerações, também, além, matemática, ótima, cálculo, três, matemáticos, versão, ruído, compreensão, dinâmica, estão, produção

Limitações

O trabalho possui duas principais limitações. A primeira vem do número de documentos usados, apenas 20% dos artigos publicados no CNMAC. Isso cria a possibilidade de artigos ou tópicos em específico terem aumentado as frequências de certas palavras, afetando a análise. A segundo vem da coleta e tratamento dos dados. Devido a estrutura do site do CNMAC, apenas obtemos acesso a arquivos PDFs, os quais tiveram que ser convertidos para texto. A ferramenta usada para isso não é perfeita e introduziu algumas palavras sem sentido, além de falhar em coletar outras.

Por fim, é importante salientar que o atual trabalho não justifica qualquer ideia de causalidade entre os dados usados. O foco é apenas análisar quais palavras possuem algum indício de serem favorecidas por ferramentas de geração artificiais.

Conclusões

Tendo em vista os dados e respostas obtidas, podemos afirmar que existem palavras genéricas, sem vínculo a um tópico em específico que sofreram um crescimento consideravel e demarcado pelos anos 2020 e 2021. Não podemos fazer afirmativas sobre a causalidade disso e vínculo com AIs generativas, mas dado que não observamos termos referêntes a eventos ocorridos nesses anos, como a pandemia do Covid, tomamos esse crescimento como possível motivação para futuras pesquisas sobre artefatos deixados por AIs em textos na forma de frequências de palavras.