Jogo de Shannon aplicado a LLMs para inferência da Entropia

Guilherme Iram Silva Araujo [1] Luiz Fernando Costa dos Santos[1]
[1]*Aluno de Graduação - Bacharelado em Ciência de Dados e Inteligência Artificial*

Universidade Federal da Paraíba
Professor Leonardo Vidal Batista
Introdução a Teoria da Informação

8 de maio de 2024

Claude Shannon, o pai da teoria da informação, desenvolveu um método de estimar entropia de textos a partir de um certo ente. No seu experimento inicial, ele utilizou humanos para jogar um pequeno jogo de adivinhação. O objetivo de Shannon era obter a distribuição de probabilidade para cada subsequência de caracteres. Entretanto, como essa tarefa era difícil de ser feita para todos os públicos, ele desenvolveu um jogo no qual se faz o jogador escolher o próximo caractere de uma dada sequência de símbolos dado um contexto (que pode variar de 0 até N). O jogador precisa tentar adivinhar qual o caractere da vez até acertar. A contagem de tentativas até o acerto é contabilizada e esses valores, para cada caractere, são utilizados no cálculo da Entropia. No nosso experimentos, fizemos do LLM GPT da OpenAI nosso jogador e avaliamos o seu desempenho no jogo com o objetivo de extrair a entropia alcançada pelo modelo de linguagem.

Dados do experimento

Para avaliar o jogador, que no caso será o LLM GPT 3.5 Turbo, 100 trechos aleatórios de tamanho 15 foram extraídas do livro "Apologia de Sócrates" por Platão. Vale ressaltar que o alfabeto utilizado foram as 26 letras de a até z minúsculas, excluindo qualquer tipo de variação envolvendo acentuação (por exemplo: “à”, “ç”, “é” etc) mais o caractere de espaço (“ “). Totalizando um alfabeto de tamanho total 27. Dado que cada frase tem tamanho 15, o contexto máximo é 14.

Aplicação do jogo no GPT

Para fazer com que o GPT consiga jogar, primeiro implementamos o jogo de uma maneira que ele consiga se comunicar com a API, depois iteramos sobre os 100 trechos aleatórios, e para cada trecho, começamos o Jogo de Shannon. A partir disso, salvamos a quantidade de tentativas para cada caractere, com o objetivo de gerar uma tabela com o número de tentativas e o contexto. A partir dai, podemos usar as seguintes equações para gerar a entropia:

Dado a sequência de caracteres $c_1, c_2, ..., c_n$, $g_1, g_2, ..., g_n$ representando a sequência de tentativas e $j \in [0, 27]$, o intervalo do alfabéto que vai de "a"a "z"mais o caractere de espaço.

0.1 Upper Bound

$$-\frac{1}{n} log(\prod_{i=1} P(g_i)) = -\frac{1}{n} \sum_{i=1}^{n} log(P(g_i)) = -\sum_{j=1}^{27} P(j) log(P(j)) \quad (1)$$

0.2 Lower Bound

$$\sum_{j=1}^{27} j \cdot [P(j) - P(j+1)] \cdot log(j) \quad (2)$$

Com relação ao upper bound, essa fórmula é a definida por Shannon no seu artigo principal "A Mathematical Theory of Communication"como sendo a forma padrão de se calcular a entropia de um texto livre. O lower bound, por outro, trabalha com os valores advindos de uma decomposição retangular com relação às probabilidades. Isto é, sua entropia é calculada em função da subtração entre uma probabilidade associada a um g_i com respeito ao seu g_{i+1}. Foi adotado como a entropia final o valor embuido no upper bound.

Desafios

Um dos maiores desafios que enfrentamos foi lidar com as respostas absurdas do modelo e tratar os resultados que ele nos enviava para selecionar apenas o palpite do GPT, que era o que nos interessava. Muita das vezes, o modelo começava a responder a frase completa, ou até mesmo perguntar qual era o próximo caractere (como se ele estivesse invertendo de jogador a quem propõe o jogo) e entre outras coisas como repetir o mesmo caractere por várias tentativas seguidas. Para lidar com isso, tivemos que deixar claro nos prompts o comportamento que gostaríamos, e também implementamos expressões regulares para tratar as respostas da API.

Um outro caso bem curioso foi que o modelo não conseguia fornecer o caractere de espaço. Então tratamos o ç como espaço (tanto nos prompts, quanto no código em sí). Assim como, ele insistia em usar vogais com acentos, mas tratamos isso bem como as outras exceções.

Resultados

O resultado do jogo de Shannon nas 100 amostras de tamanho 15 pode ser visto abaixo:

A entropia inferida foi 3.01735 de bits por símbolo, com desvio padrão de 0.31747.

Além disso, geramos a tabela com as frequências para cada contexto e tentativa:

	ctx_1	ctx_2	ctx_3	ctx_4	ctx_5	ctx_6	ctx_7	ctx_8	ctx_9	ctx_10	ctx_11	ctx_12	ctx_13	ctx_14	ctx_15
g_1	9	26	15	13	11	13	19	14	23	19	24	19	21	21	23
g_2	7	13	6	11	9	10	12	12	8	11	6	4	14	13	10
g_3	6	5	13	11	13	12	11	13	3	7	12	12	6	10	14
g_4	7	8	8	3	2	5	6	6	5	6	6	4	6	10	11
g_5	5	3	4	4	8	7	3	9	1	6	1	4	6	5	5
g_6	6	5	2	6	4	9	7	9	5	3	2	3	6	2	7
g_7	3	1	4	6	7	5	10	7	4	5	3	6	6	3	7
g_8	0	5	10	2	2	3	2	2	5	4	3	4	1	4	5
g_9	7	0	4	3	2	3	1	2	3	4	6	1	4	2	2
g_10	0	3	1	3	6	3	2	6	2	4	4	4	3	4	0
g_11	5	1	3	1	3	0	5	0	4	2	3	3	0	3	0
g_12	9	2	1	4	3	2	2	1	2	3	3	4	3	2	0
g_13	17	1	2	2	1	0	0	1	0	3	5	3	2	2	2
g_14	6	0	3	1	1	1	2	0	1	2	2	2	3	0	0
g_15	1	0	2	1	1	1	1	3	0	2	2	1	2	3	0
g_16	0	1	0	0	1	1	3	1	0	0	1	1	1	1	0
g_17	3	0	2	1	2	1	0	0	0	2	1	1	0	1	0
g_18	0	1	0	0	1	0	0	2	4	0	0	2	1	1	0
g_19	0	4	0	0	2	0	2	0	4	0	0	1	1	0	0
g_20	2	1	0	0	1	1	1	0	2	1	0	1	0	0	0
g_21	0	0	1	0	4	1	1	0	1	3	0	4	1	1	1
g_22	3	0	1	0	0	1	1	1	1	3	0	1	0	2	1
g_23	1	0	1	2	0	2	0	2	1	0	0	1	1	0	0
g_24	0	2	0	2	0	2	2	2	0	0	1	0	1	1	0
g_25	3	0	1	0	0	1	0	1	0	0	1	0	0	1	1
g_26	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
g_27	0	18	16	24	16	15	7	6	21	10	14	14	11	8	10

Figura 1: *Tabela de Frequências*

E geramos o gráfico, com os uppers e lower bounds para cada número de contexto, sendo considerado para a entropia o upper bound.

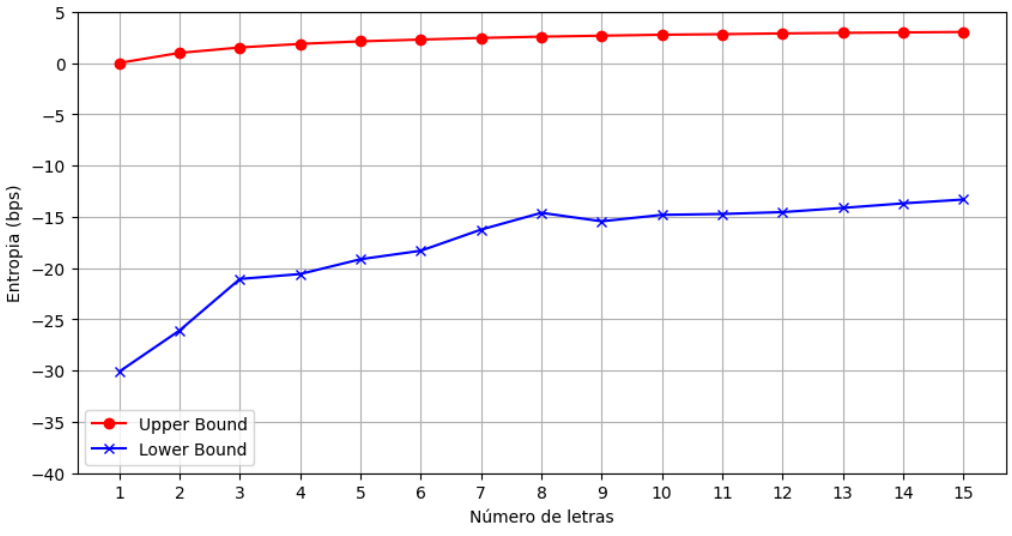

Figura 2: *Entropia de um texto em função do número de letras do contexto*

É possível perceber que, com o contexto indo para o infinito, poderíamos inferir que o upper e lower bounds iriam ser cada vez mais próximos.

Algumas análises podem ser feitas a partir dos resultados obtidos. Uma delas é sobre a indiferença ao contexto no modelo, pois, independente do tamanho do contexto o GPT não tem melhora ou piora nos acertos, o que indica que o contexto, para esses modelos, de acordo com nosso experimento não, faz diferença. Isso provavelmente está associado ao tipo de natureza probabilística desses modelos, que trabalham com contextos fixos pré-treinados. Para um ser humano, o contexto relativo da mensagem seria muito importante para essa tarefa.

Outra análise a ser feita é que o GPT teve muitas dificuldades ao lidar com o caractere de espaço, e muitas vezes ele tentou todas as letras antes de tentar o espaço (isso pode ser visto com o grande número de 27 tentativas na figura 1).

Trabalhos Futuros

Após conseguir esses resultados com o GPT 3.5, uma ideia de trabalho futuro seria comparar a entropia com outros modelos de LLM, tanto open source, quanto proprietários, mudando as configurações para controlar mais ou menos a aleatoriedade desses modelos (o que é chamado de temperatura dos modelos). Outro tópico interessante seria fazer um fine-tunning para algum contexto específico de algum modelo open source e depois, usar esse modelo para jogar o jogo de shannon com trechos desse contexto e checar a perfomance. Outra proposta seria fazer com textos em inglês, para comparar com mais precisão com o trabalho do próprio Shannon e outros.

Referências

SHANNON, C. E. Prediction and Entropy of Printed English. (Manuscrito recebido em 15 set. 1950).

GHAZVININEJAD, Marjan; **KNIGHT**, Kevin. Humans Outperform Machines at the Bilingual Shannon Game. Information Sciences Institute, University of Southern California, 4676 Admiralty Way #1001, Marina Del Rey, CA 90292, USA. Correspondência: ghazvini@isi.edu; Tel.: +1-310-822-1511. Ambos os autores contribuíram igualmente para este trabalho. Academic Editors: Kevin H. Knuth and Raúl Alcaraz Martínez. Recebido em: 3 out. 2016; Aceito em: 27 dez. 2016; Publicado em: 30 dez. 2016.