Esse nosso terceiro passo, nossa terceira parte, digamos assim, falando sobre a Atenção. Sim, literalmente, aqui a gente está olhando ainda dentro do contexto de tokens, a probabilidade que a gente vai ter de acontecer alguma coisa. Lembra que a gente estava falando no início da aula sobre codificador e decodificador? Eles entram aqui, literalmente com toda a força. Por quê? Porque a gente vai ter um modelo de transformador em cima das várias camadas de dados. A nossa rede neural, que era aquele exemplo que a gente tinha olhado anteriormente.

Então, a gente precisa pensar em cada detalhe. Percebam que, para que tudo isso venha a acontecer, são várias etapas e essa é uma delas.

Então, a atenção, ela vai ser uma técnica que a gente vai utilizar para examinar essa sequência de tokens, ou seja, as nossas probabilidades. Lembre-se, eu preciso saber qual que é probabilidade de uma palavra aparecer depois da outra. Eu preciso entender, também, se aquilo faz sentido, se eu estou criando frases com sentido. E depois, literalmente uma pontuação, que a gente vai falar com uma atenção. Mas considerar esses tokens, quais influenciam os próximos. Então, aqui a gente vai ter blocos codificadores. Por exemplo, eu ouvi o cachorro latir. Faz sentido. Agora, se eu disser, eu ouvi o cachorro miar. Não, isso não está relacionado a cachorro, está relacionado a gato.

Então, obviamente, nós sabemos disso, mas como que a máquina vai conseguir mensurar isso? Então, aqui a gente precisa literalmente ter uma flag. Como que a gente consegue trabalhar esse tipo de informação vai ser no nosso treinamento. Então quando a gente estiver prevendo o que vai ser o nosso próximo passo. Pense que o treinamento na verdade vai trazer as probabilidades, as sequências e aquilo que é esperado. Então eu estou treinando a minha máquina, para que ela venha me trazer a resposta que faça sentido.

Então, obviamente, a gente vai ter aqui uma taxa de erros e acertos. É como um jogador de futebol, ele vai chutar várias vezes, tentando acertar um determinado ângulo, mas a forma como ele consegue chutar a bola, a direção do vento, entre outras coisas, podem influenciar.

Aqui a gente está falando de uma coisa extremamente lógica. Então, se isso é verdadeiro então tal coisa, mais ou menos por aí. Então, quando a gente fala sobre a questão do treinamento, por exemplo, aqui a nossa meta é prever o toque após o cachorro. O que eu sei que o cachorro faz? Eu sei que o cachorro late, ele não mia, ok? Então, o nosso está representando aqui, ouvir o cachorro, opa! Se eu ouvi o cachorro, então ele estava fazendo barulho. Então, o que faz mais sentido, né? Ouvir e cachorro. Então, percebam que dentro da frase sempre vai ter algumas palavras-chave né? Que vão determinar o nosso próximo passo.

A missão aqui é a gente conseguir elencar isso e com base nas nossas top aqui, né? As palavras top, conseguir determinar o que vem depois.

Então... vários tokens possíveis podem vir depois do cachorro. O cachorro caminha, o cachorro corre, ele late, ele pula. Os meus aqui fazem outras coisas também, fazem muita arte em casa, para não dizer outra coisa. Nesse caso, aqui a gente está trazendo que a sequência mais provável, com base no cenário é que ele está latindo. Então, como que a gente vai trabalhar isso na nossa máquina?

Aqui, a gente vai fazer o nosso teste. Então, eu treinando a minha máquina. Eu escutei um cachorro latir. Ok? Para cada uma dessas palavras, a gente está determinando o nosso número, a nossa probabilidade. E estamos fazendo essa atribuição. Então, aqui a gente vai ter uma sequência de inserções do nosso token. Que ele vai estar alimentando essa camada de atenção.

Então, o nosso token vai estar representando valores numéricos como vocês podem ver aqui, cada uma dessas palavras está representando. E a nossa meta, na verdade, é que o nosso decodificador consiga prever, de uma forma assertiva, qual é o próximo token, o nosso próximo passo, a próxima palavra aqui da nossa sequência.

Então, como que ele vai calcular isso? Durante o nosso treinamento eu sei qual que é sequência real. Ou seja, eu alimento a minha máquina com a informação que faça sentido.

Então, eu estou trazendo para ela, obviamente, olha, o que que o cachorro faz, ele faz no seu dia a dia. Ele late, ele corre, ele pula, mas eu ouvi e o ouvir está relacionado a um barulho. Então, entre ele correr e ele pular e ele latir se eu estou ouvindo, estou vendo então é essa palavra que é que vai fazer mais sentido. Então a gente tem um modelo transformador, o próprio Chat GPT, o Bing, fazem isso. A gente brinca que a máquina, melhor dizendo, a mágica acontecendo.

Então, essa é a estratégia no modelo de atenção. A gente já conseguiu, literalmente, quebrar em fatias. Eu já consigo entender quais são as palavras. Essa probabilidade de palavras, uma que vem depois da outra. Depois, a gente está trabalhando com aquilo que faça sentido, a nossa semântica. E, obviamente, aqui os nossos gatilhos, o modelo de atenção. Aquilo que sim, faz sentido, é a probabilidade dessa palavra vir com base naquilo que a gente treinou, com base também em todos os anteriores. Fazer sentido, probabilidades entre outros.

Então, essa aqui a gente vai elencar como sendo a mais provável. Então, cada uma dessas etapas ao se encaixar, vão trazer para nós os resultados esperados.

Essa aqui é a etapa... Atenção!