Nome: Fernando Buligon Antunes **Data:**02/04/2025

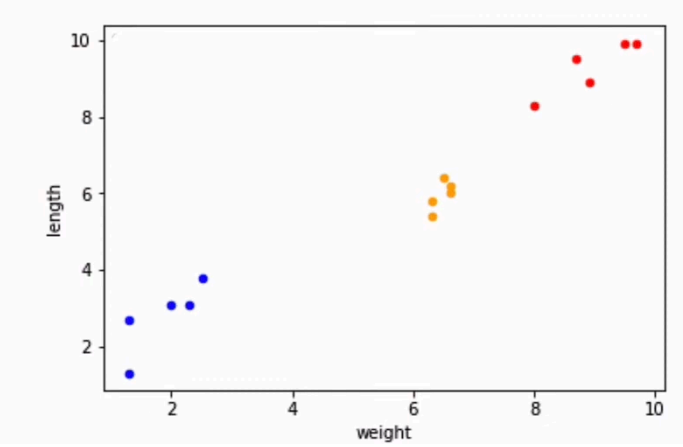
Introdução a Embeddings

No vídeo é falado sobre word embeddings, que são representações de palavras, sendo um dos fatores mais importantes quando o assunto é processamento de linguagem natural. O início é dado com o apresentador definindo linguagem natural como “um sistema complexo de comunicação usado para expressar significados”, e em cima dessa definição, ele fala que o objetivo de word embedding é capturar esses significados por trás dessa linguagem natural de uma forma em que a máquina consiga processar.

Uma maneira antiga de demonstrar a palavra para o computador era o one hot vector, em que era criado um vetor de inteiros com tamanho suficiente para conter n palavras, cada índex representava uma palavra diferente, então o que representasse tal palavra ficava com 1 e o resto com zero, o problema por trás disso era que não era possível representar similaridade, então por exemplo, palavras como hotel, motel e pousada iam ser vistas como total diferentes, apesar de possuir significados semelhantes.

Com base nesse problema surgiu a ideia de “semântica distributiva”, onde “o significado de uma palavra é dado pelas palavras que aparecem junto a ela”, ou seja, pelo contexto. Através dessa maneira o modelo passa a ser capaz de entender que apesar da escrita das palavras serem diferentes, elas podem possuir o mesmo significado ou semelhantes, por exemplo, homem e mulher, são dois casos diferentes mas o modelo consegue encontrar uma relação entre eles.

Essa similaridade é calculada com o cosseno de distância, então quanto mais próxima uma da outra, mais similar é, por exemplo, na imagem abaixo cada pontinho representa uma palavra, os pontinhos vermelhos, como estão próximos, são todos similares, mas os amarelos não são tão parecidos por estarem um pouco longe, e os azuis são completamente diferentes.



O vídeo é finalizado com o autor apresentando o Word2vec, GloVe e BERT, que são alguns frameworks famosos para trabalhar com word embedding.

Criando um LLM do Zero

Foram feitas algumas coisas como:

- Criação de um vocabulário: transformamos um texto em um vocabulário, nele temos todos os caracteres presentes (importante citar que cada caractere só foi adicionado uma única vez)
- Codificador: Com o vocabulário criado, foram adicionados indexes a cada caractere e através do lambda foi montado um codificador.
- Decodificador: Foi feito o processo reverso do codificador, ao invés de adicionar um índex para cada letra, foi adicionado uma letra para cada índex, com isso foi criado um decodificador novamente usando o lambda.
- Tiktoken: Foi feito um codificador e um decodificador usando a biblioteca tiktoken.
- GPT x BERT: Foi feita uma comparação a como os modelos GPT e BERT fazem predições, basicamente o gpt vai da esquerda para a direita, sempre prevendo qual vai ser a próxima palavra, e o bert coloca uma máscara em palavras aleatórias.