Word Embeddings (LLM NOTES)

[Mapping notes — P111
Positional Embeddings — P112
Vector DB's — P95]

- in reality tokenization frequently divides words, not whole words every time but we will focus on it by thinking tokens are clean whole words

truth: $\boxed{\text{This}}\boxed{\text{pro}}\boxed{\text{cess}}\boxed{\text{(}}\boxed{\text{known}}\boxed{\text{fanci}}\boxed{\text{fully}}$ Lie: $\boxed{\text{this}}\ \boxed{\text{process}}\ \boxed{\text{(}}\ \boxed{\text{known}}\ \boxed{\text{fancifully}}$

token ← ↳ and same for output token it can be any token not always a full word

- The model has a pre defined Vocabulary, some list of all possible tokens (words, symbols, spaces, number etc) → Not always whole
  and the first matrix we will see is the embedding matrix has a single column for each of these words these colums determine what vector each word get in that first step we label W_E = Embedding Matrix.

All words ≈ 50K = 50K tokens

$$\begin{array}{cccccc} \text{1} & \text{aback} & \text{blue} & \text{boy} & \cdots \text{zing} & ; \\ \end{array}$$
$$\begin{bmatrix} 2.1 & 2.1 & 5 & 7 & \cdots & 2 & 5 \\ 3.2 & 2 & 6 & 4 & \cdots & 7 & 7 \\ 5.1 & 3 & 7 & 9 & \cdots & 6 & 8 \\ 6 & 4 & 8 & 1 & \cdots & 8 & 9 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 9 & 9.1 & 2 & 5 & \cdots & 7 & 9 \end{bmatrix}$$

Embedding Matrix W_E

the sky was blue

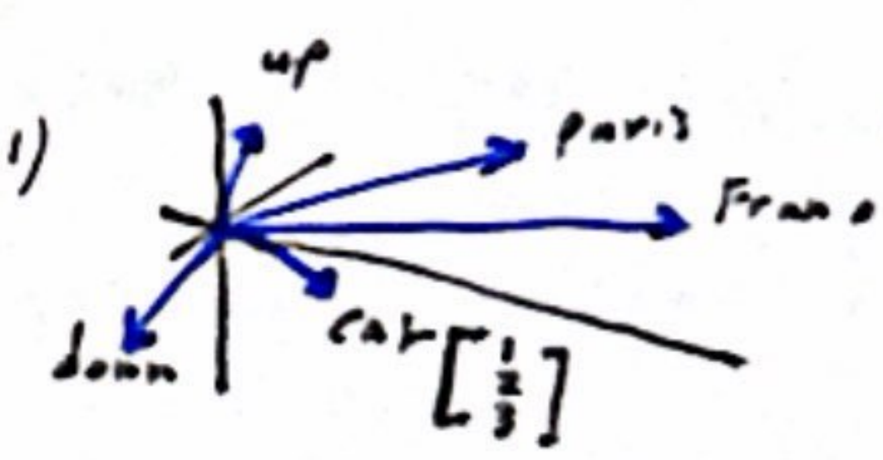
$$\begin{bmatrix} \vdots \end{bmatrix} \begin{bmatrix} \vdots \end{bmatrix} \begin{bmatrix} \vdots \end{bmatrix} \begin{bmatrix} 5 \\ 6 \\ 7 \\ 8 \end{bmatrix}$$

} EM picks the vector for blue

★ Values begin random and learned through data ★

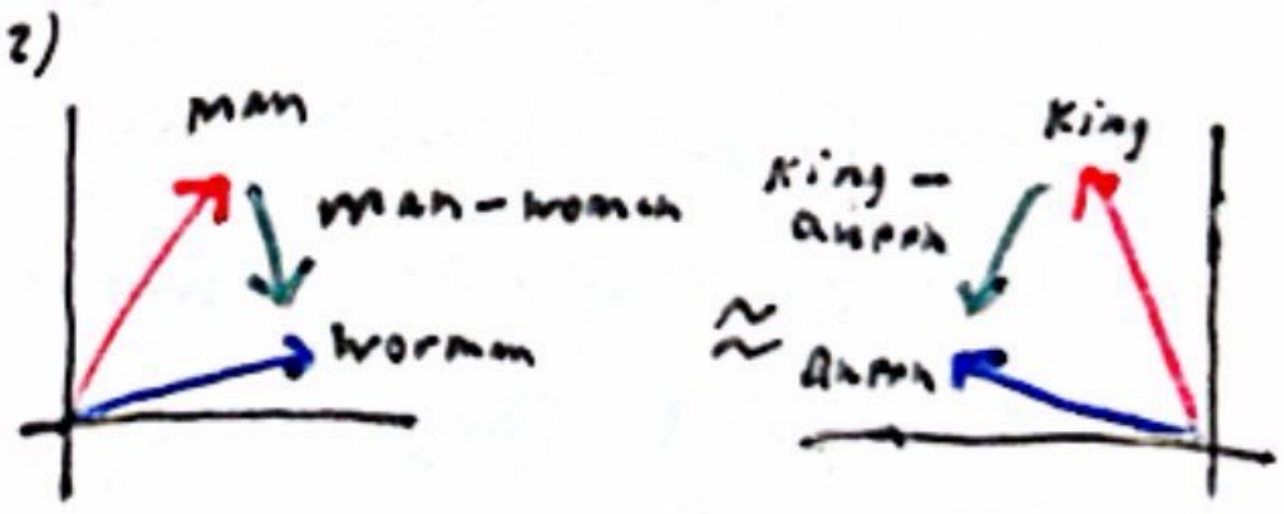
- We embed these words as tokens which means its a vector in some very high dimention space like 12,288 dimentions in GPT-3. So we visulize in 3D. as it learns these embedding in training it settles on embedding's whos direction have meaning for Ex paris and France's embedding are close as there related but up and down for ex are opposite in meaning and space. in the Ex 2 we see how Embedding are related and why using vectors is good say we did not know what a Female leader is we could do King + (women − man) and finding the vector (embedding) closest to that. $E(\text{queen}) \approx E(\text{King}) + E(\text{women}) - E(\text{man})$

Ex 1)



⇒ embedding space 3D

Ex 2)



$E(\text{man}) - E(\text{women}) \approx E(\text{King}) - E(\text{women})$

- this is a perfect Ex in reality the might be further apart and are impossible to visulize. they are leaned in training.
- further more "Queen" in traning data would not always be a female leader Ex "Queen the band" hence the embedding for Queen wont be the given formula but close still.