pdf / image ...

Script of a movie

My cat ate the cake.
It is bad.

→ I am sitting on a bank.
River

→ I am sitting in a bank.
finance

Foundational ML models

→ LLMs
- translate language
- analyse sentiments
- chatbots conversations

Capable of understanding
- → Complex data.
- → Identify the relationship b/w words
- → generate new text → coherent → grammatically correct

→ Autocomplete a LLM?] NO

↳ language model

Earlier

→ { predict the prob of Once upon a Big word. time → 50%, dream → 15%, reaction → 4% }

Human Language → bigger

LLM → predict the next sentence / paragraph / doc

My → name

My name → is

My name is →

NLP

Break the text → tokenization

is, an, are, the

Stop words X

Understand → Stemming & Lemmatization

Embedding

I ~~like~~ Taj Mahal ~~is~~ beautiful

Tokenization

Hello, How are you ?

"Hello", ",", "How", "are", "you" } Word tokenization

Don't waste food

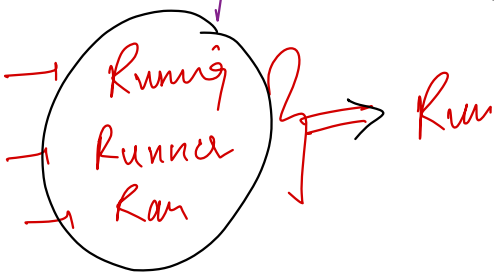
English

Chinese or Japanese

["Do", "n't", "waste", "food"] } Sub words

["D", "o", "n", "'", "t" } Character tokenization

Stemming & Lemmatization



→ NLTK

→ Spacy

more Acc ↑ more comp expensive

Chop-it off Approach

↓

Stemming

Lemmatization

Semantic Connections

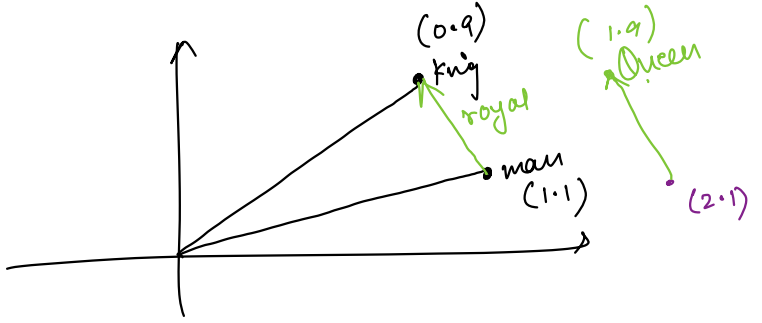
Change
Changing
Changes
Changed
Changer

Chang

change

Embeddings

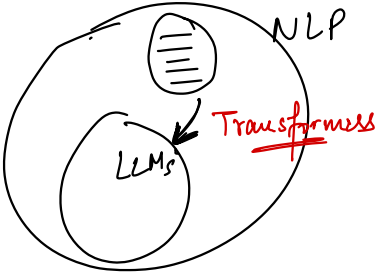
Semantic relation b/w the words



✓ King − man = ✓

King − man + woman = ✓

"Word2vec"



Sequential processing

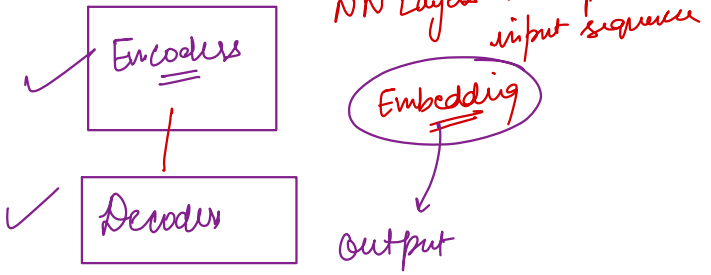
↳ connections b/w words

More Context

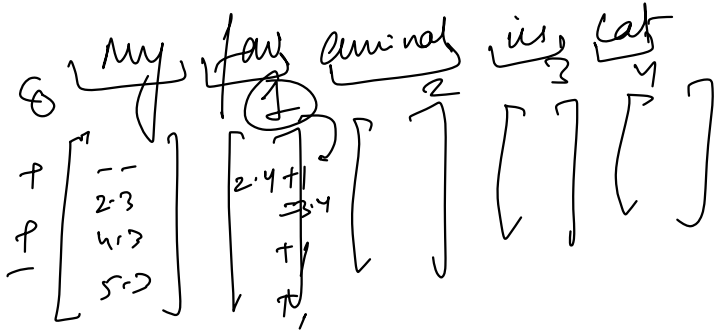
Sequence ✓

Encoder & decoder architecture

Auto-encoders



Token embedding + Positional embedding



Learning in a transformer

(1) Encoder learning] (*) Compresses the input info into compact representation

(*) Identifies imp features

(*) Disregards the irrelevant & redundant info.

(*) Create a condensed summary

(2) Decoder learning

(AE) reconstruct from encoder's compact rep.

Traning

→ Under compact rep — loss f...

→ Learns patterns / rel within data

→ generate the output (similar to input)

Scoring process

(*) Encoder → rep state

(*) Decoder — generate output

(*) output compared expected output using the loss func.

Autoencoders → Remove noise

Attention

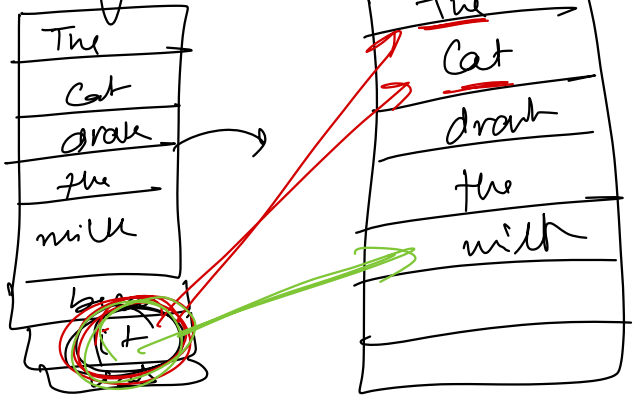
I played with the bat

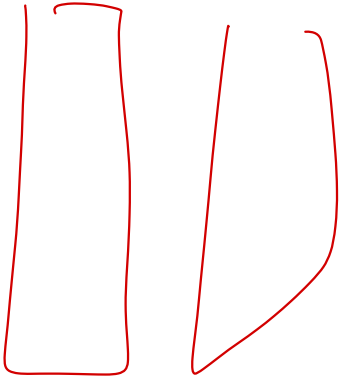
The bat flew in the night

The cat drank the milk because it was hungry.

it — Sweet.

Self - Attention





Queries = question you are seeking

Key = potential matches in your query found in each book.

values = actual piece of info → when key match your query.

Library

Step 1 for each word in the sentence, model with create a query : Key, value

"Cat" → query ⇒ looking for actions association

Step 2 Swing

↳ how related the words are.

Step 3 weighted Summary