# NLP Session

# Topics to be covered

1. What is NLP?
2. Preprocessing
3. Bag of Words
4. TF-IDF
5. Word2Vec
6. NLP Process
7. Neural Networks
8. Encoder Decoder Models
9. Transformers
10. ChatGPT

# What is NLP

- NLP is the field of computer science and Artificial Intelligence concerned with enabling computers to understand, interpret, and generate human language.

- Applications of NLP
  - Sentiment Analysis - To determine if sentence is positive, negative or neutral.
  - Machine Translation - Conversion of text in one language into another.
  - Text Summarization - Finding Summary of Text.
  - Chatbots - To deliver response on the basis of user request.
  - Speech Recognition - Process the human speech into written format.

- Terminology in NLP
  - Corpus
  - Documents
  - Vocabulary
  - StopWords

# Preprocessing

- Tokenization - Tokenization is the process of breaking down a piece of text into smaller units called tokens.
  - Eg: “Tom likes to eat pizza.” - [“Tom”, “likes”, “to”, “eat”, “pizza”, “.”]
- Stemming - Stemming is a technique used in natural language processing to reduce words to their base or root form.
  - Eg: “helping” - “help”
- Lemmatization - Lemmatization is similar to stemming, but it goes one step further by reducing words to their dictionary form.
  - Eg: “caring” - “care”
- One Hot Encoding - The element corresponding to the actual category value is set to 1, while all other elements are set to 0.
  - Eg: “Tom likes pizza” -
    - “Tom” - [1 0 0]
    - “likes” - [0 1 0]
    - “pizza” - [0 0 1]

# Bag Of Words

- Main Purpose is to represent the text data in a numerical format that can be processed by machine learning algorithms.

- BOW is a way of counting the frequency of words in a document.

- Example -
  a. "The quick brown fox jumps over the lazy dog."
  b. "The dog chased the cat up the tree."
  c. "The cat meowed loudly and ran away."

  - Unique words in document - ['and', 'away', 'brown', 'cat', 'chased', 'dog', 'fox', 'jumps', 'lazy', 'loudly', 'meowed', 'over', 'quick', 'ran', 'the', 'tree', 'up']

  - BOW vector for Document 1 : [[0 0 1 0 0 1 1 1 1 0 0 0 1 1 2 0 0]]
  - BOW vector for Document 2 : [[0 0 0 1 1 1 0 0 0 0 0 0 0 3 1 1]]
  - BOW vector for Document 3 : [[1 1 0 1 0 0 0 0 0 1 1 0 0 1 1 0 0]]

# TF-IDF

- It is a popular technique used in natural language processing to weight the importance of words in a document or a corpus of documents.

- There are 2 terms -
  - Term Frequency
  - Inverse Document Frequency

- Formula for TF and IDF-
  - TF = (Number of times a words appears in a sentence) / (Total number of words in the sentence)
  - IDF = log(Number of sentences) / (Number of sentences containing the word)

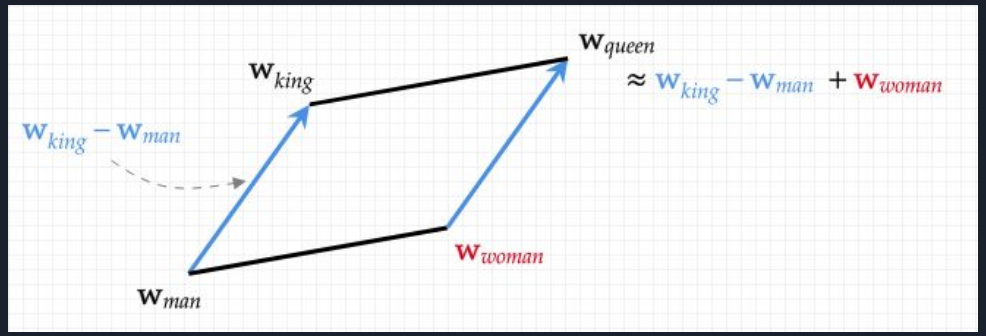
- The TF-IDF score is then calculated by multiplying the TF and IDF scores of each word in a sentence.

# Word2Vec

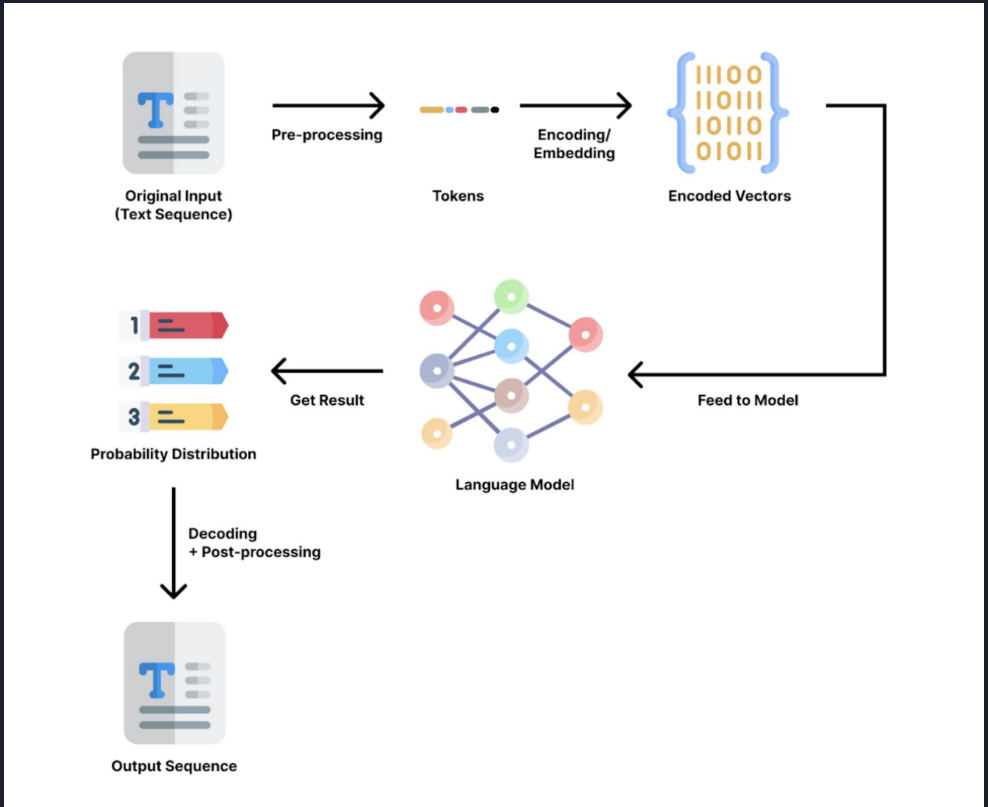
- Main Purpose is to learn vector representations of words. The algorithm is based on the idea that words that appear in similar contexts tend to have similar meanings.

- Two main types of Word2Vec models:
  - Continuous Bag of Words (CBOW)
  - Skip-gram

- Example -
  - King - Man + Woman = Queen



# NLP Process

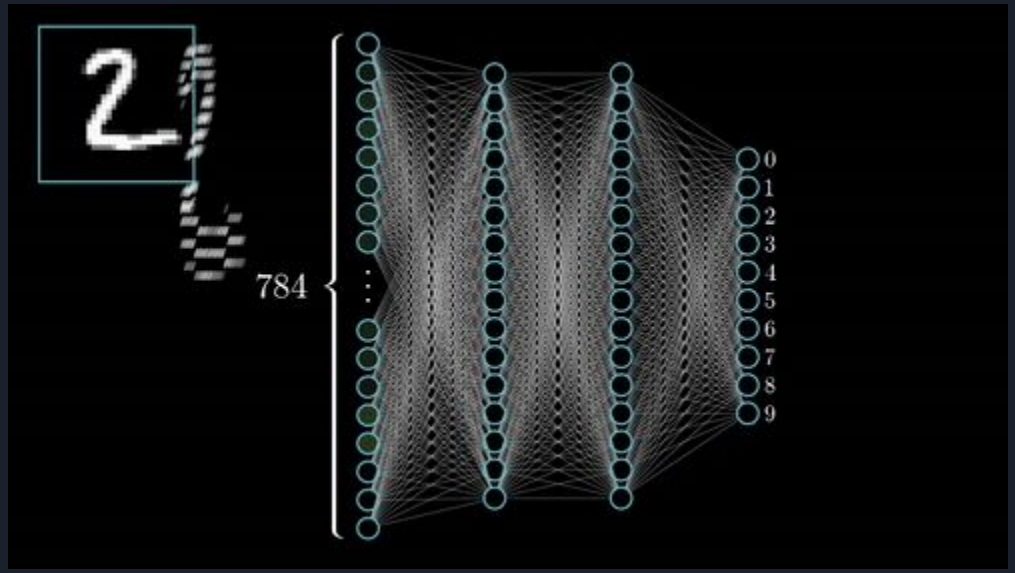

Original Input
(Text Sequence)
Pre-processing
Tokens
Encoding/
Embedding
Encoded Vectors
Feed to Model
Language Model
Get Result
1
2
3
Probability Distribution
Decoding
+ Post-processing
Output Sequence

# NLP Process

- **Preprocessing**: Cleaning the text with techniques like tokenization (breaking down the text into small pieces called tokens), stemming (removing suffixes or prefixes), removing stop words, correcting spelling, etc. For example, “Tom likes to eat pizza.” would be tokenized into [“Tom”, “likes”, “to”, “eat”, “pizza”, “.”]
- **Encoding or embedding**: Turn the cleaned text into a vector of numbers, so that the model can process.
- **Feeding to model**: Pass the encoded input to the model for processing.
- **Getting result**: Get a result of a probability distribution of potential words represented in vectors of numbers from the model.
- **Decoding**: Translate the vector back to human-readable words.
- **Post-processing**: Refine the output with spell checking, grammar checking, punctuation, capitalization, etc.

# Neural Networks

- Neural networks are set of algorithms inspired by the functioning of human brian.



# Types of Neural Network

1. **Artificial Neural Network**:  ANN consist of input layer, hidden layer and output layer. They are majorly used in classification and data prediction tasks.

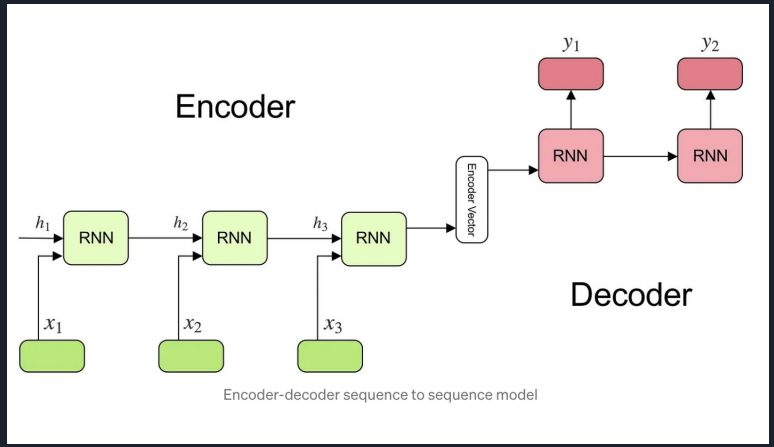
2. **Convolution Neural Network**: CNN uses filters to process image data. CNN can be used for object detection, object classification and recognition.

3. **Recurrent Neural Network**: Recurrent neural networks are networks designed to interpret temporal or sequential information. RNN are commonly used for NLP and audio sequence data.

4. **Long Short Term Memory**: To learn long-term dependencies, our neural network needs memorization power. LSTMs are a special case of RNNs which can do that.

# Encoder Decoder Model

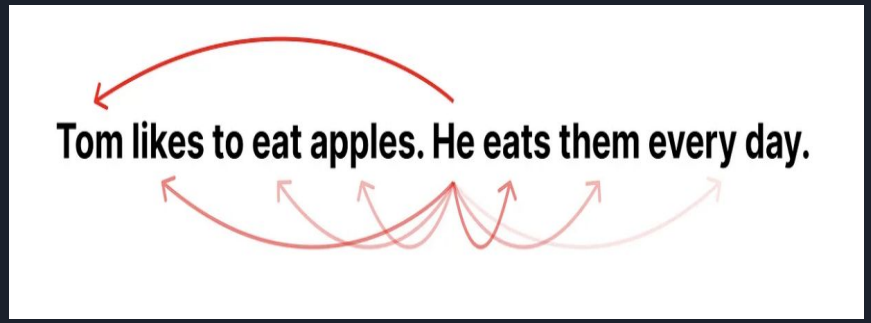
- Also known as Sequence to Sequence model.
- A sequence to sequence model aims to map a fixed-length input with a fixed-length output where the length of the input and output may differ.
- Encoder stores all the important information in a single vector of number and passes that vector to decoder.
- Decoder tries to decode the information from the vector provided by encoder.
- Use Cases for encoder decoder model - machine translation, QA systems, text summarization etc.



Encoder-decoder sequence to sequence model

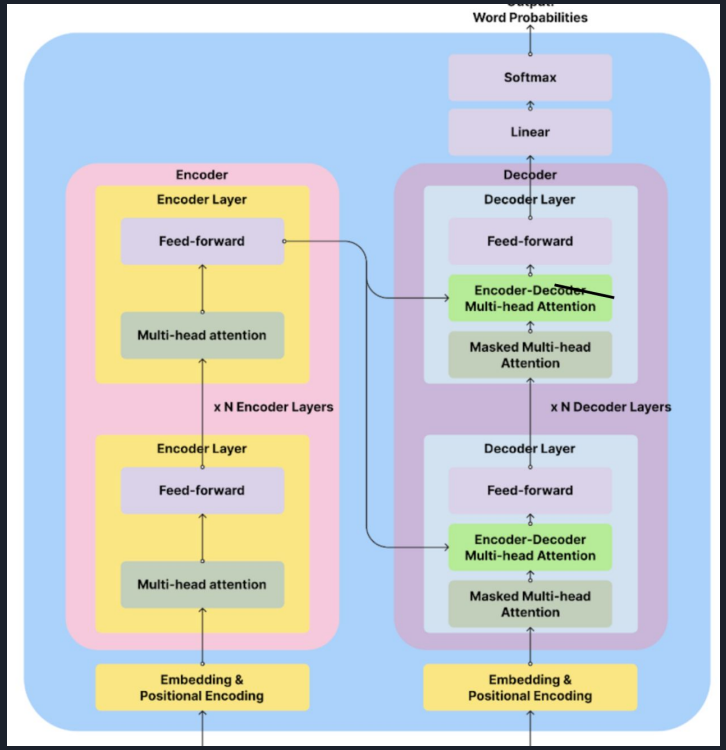
# Transformers

- Transformer is an architecture for transforming one sequence into another one with the help of two parts (Encoder and Decoder).
- The transformer can understand contexts in sequential data like text, speech, or music better with mechanisms called attention and self-attention.
- Attention allows the model to focus on the most relevant parts of the input and output by learning the relevance or similarity between the elements.



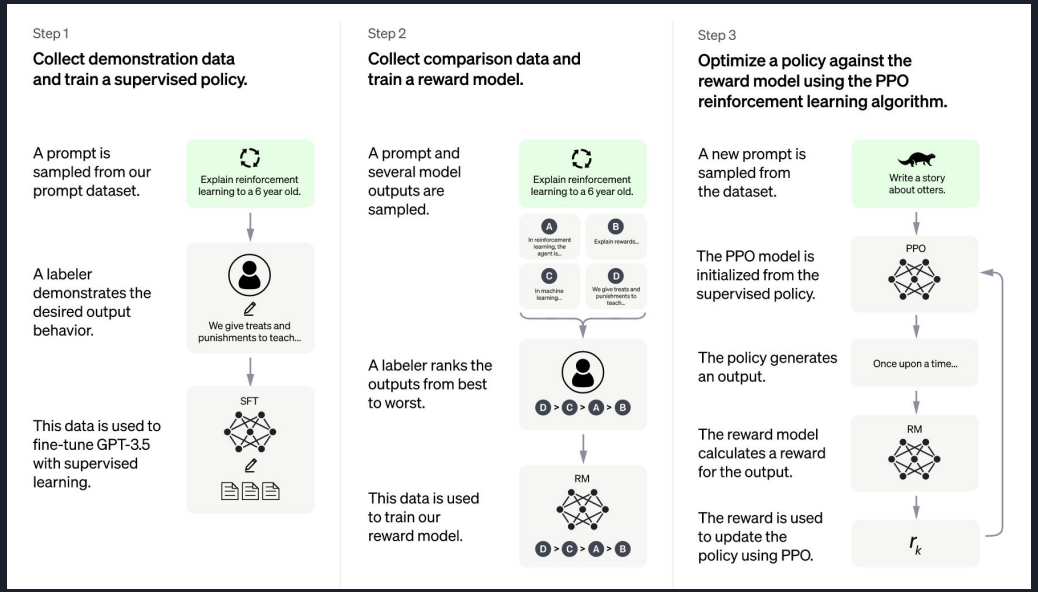
# Transformers

- **Embedding & Positional Encoding**: turning words into vectors of numbers.
- **Encoder**: extract features from the input sequence and analyze the meaning and context of it.
- **Decoder**: generate the output sequence based on the output from the encoder and the previous output tokens
- **Linear & Softmax Layer**: turning the vector into a probability distribution of output words.



# ChatGPT

- It is a variant of the popular GPT-3 (Generative Pertained Transformer 3) model, which has been trained on a massive amount of text data to generate human-like responses to a given input.



# ChatGPT

- The first step is **supervised learning** from human examples. Researchers first provided the pre-trained GPT with a curated, labeled dataset of prompt and response pairs written by human labelers.From this step, they get a supervised fine-tuned (SFT) model.
- The second step is **training a reward model (RM)** to rate the responses from the generative model. Researchers used the SFT model to generate multiple responses from each prompt and asked human labelers to rank the responses from best to worse by quality, engagement, informativeness, safety, coherence, and relevance. The prompts, responses, and rankings are fed to a reward model to learn human preferences of the responses through supervised learning. The reward model can predict a scalar reward value based on how well the response matches human preferences.
- In the third step, researchers used the reward model to optimize the SFT model's policy through **reinforcement learning**. The SFT model will generate a response from a new prompt; the reward model will assess the response and give it a reward value that approximates human preference; the reward is then used to optimize the generative model by updating its parameters. For example, if the generative model generates a response that the reward model thinks humans may like, it will get a positive reward to continue generating similar responses in the future; and vice versa.

Thank You