DSN

Large Language model

Natural Language Processing

NLP is a branch of artificial intelligence (AI) that focuses on the interaction between computers and human languages.

Its goal is to enable machines to understand, interpret, and generate human language in a way that is both valuable and meaningful.

NLP combines linguistics (the study of language structure) with computer science (the development of algorithms and computational models).

Rule Based Systems (1950s - 1970s):

Early NLP systems were heavily reliant on handcrafted rules, grammars, and dictionaries.

Researchers focused on developing algorithms that could parse sentences and perform basic tasks like translation (e.g., the Georgetown-IBM experiment in1954, one of the first machine translation experiments).

Statistical Modelling (1980s - Early 2000s)

Statistical methods in NLP refer to techniques that rely on statistical models and data-driven approaches to process and analyze language.

Unlike rule-based methods, which are manually designed, statistical methods typically involve learning patterns and structures from large corpora of text. These methods are often used in combination with machine learning techniques to perform various NLP tasks.

Deep Learning Revolution (2010s - Present)

Neural Networks:

There was a major shift in with the creation of RNN's, improving the ability to process sequential data like text and speech

Word embeddings:

Techniques like Word2Vec and GloVe allowed words to be represented as dense vectors (embedding space), capturing semantic meanings and relationships between words.

Attention mechanisms and Transformers:

In 2017, the Transformer model was introduced in the paper "Attention is All You Need". This architecture revolutionized NLP by enabling more efficient parallel processing and allowing the model to focus on relevant parts of the input data (i.e., attention). The Transformer model formed the basis of models like BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), and others.

Fundamental Concepts

Corpus:

A corpus is a large and structured collection of data, which can consist of text, audio, images, or other forms of data, typically used for training, testing, or analyzing models.

Tokens:

Tokens are the smallest meaningful units of text, generated during tokenization.

Tokenization splits text into words, punctuation, symbols, or subword segments, depending on the tokenization method used. Tokens are the fundamental units processed by NLP models.

Vocabulary:

often referred to as a "vocab" is the set of unique tokens extracted from a corpus.

Each token within the vocabulary typically maps to a unique numerical identifier, which models use internally for computation.

Probability Basics

Joint Probability
probability of two events occurring together.
example: rolling two dice and getting 4 on the first die and 3 on the second.

Conditional Probability
probability of an event given that another event has occurred.
example: Probability of drawing Ace of Spades given the card drawn is a spade.

Bayes' Rule
updates probability based on new information (reverses conditional probability).
example: Probability of having a disease after a positive medical test.

N Gram Models

Is a probabilistic model that predicts the next word in a sequence based on the previous N - 1 words.

It is one of the earliest and most intuitive statistical language models.

Traditional Statistical NLP Techniques

- Word frequency: Bag-of-words, TF-IDF
- POS tagging
- Markov Chains & Hidden Markov Models (HMM)
- Lab: Build a simple POS tagger

Word Frequency

Word frequency analysis simply counts how many times each word appears in a given text or collection of texts (corpus).

- Bag of Words (BoW):

counts word frequency, ignores grammar and order.

- Term Frequency–Inverse Document Frequency (TF-IDF):

adjusts for common words across documents.

Part of Speech Tagging

POS tagging is an NLP technique that involves assigning specific grammatical categories or labels (such as nouns, verbs, adjectives, adverbs, pronouns, etc.) to individual words within a sentence.

Markov Chains & Hidden Markov Models

Markov chain is a mathematical system where the probability of transitioning to the next state depends only on the current state, not on the history of how the system arrived at that state.

A Hidden Markov Model (HMM) is a Markov process where the true states are hidden, but you observe outputs (emissions) that are probabilistically linked to those hidden states.

Thank you