AI SOCIETY

Introduction to the Natural Language Processing Lab

Week #1 - 4 Sept ‘24

NLP Lab

About Me !

Rajat Aayush Jha

Education

2023-25

MS Computer Science @ ASU

Pursuing Thesis, co-advised by Dr. Chitta Baral & Dr. Vivek Gupta

2017-21

Bachelors @ NITK Surathkal, India

Work Experience

Summer’24

Applied Scientist @ Amazon, Seattle, US

Part of International Seller Services;
Worked on bringing explainability using LLMs

2021-23

MTS @ Oracle, India

Part of Oracle Analytics Team

Summer’21

Research Intern @ Siemens, India

Part of Siemens Research & Automation Team;
Worked on Knowledge Fusion

Research Experience

2024-25

Coral Lab & APG @ ASU

Working on problems of visual hallucination in Multimodal LLMs, compositionality issue in Text-to-Image and Text-to-Video models.

2022-23

Xu Lab @ CMU

Worked on Contrastive Unsupervised Representation Learning for Cryo-ET particle detection; Authored the *Feature Enhancement* section of the *Data Mining* chapter for the book titled *Cryogenic Electron Tomography : A Journey from Sample Preparation to Data Mining*

2021-22

MIDAS Lab @ IIIT Delhi

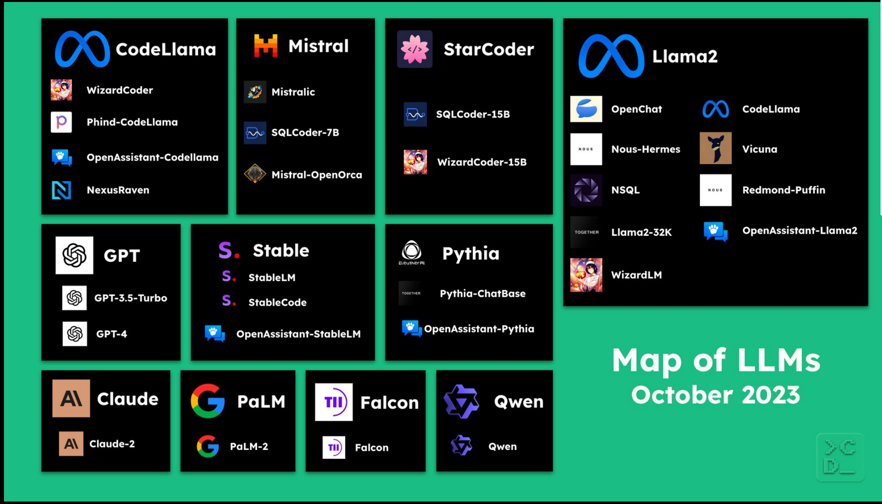
Worked on Persuasion Modelling; Published our work in AAAI 2023

2021-22

PathCheck @ MIT

Worked on evaluating performance of forecasting models; Work published in *Frontiers in Public Health.*

NLP 2.0 : The Era of LLMs



CodeLlama
WizardCoder
Phind-CodeLlama
OpenAssistant-Codellama
NexusRaven

Mistral
Mistralic
SQLCoder-7B
Mistral-OpenOrca

StarCoder
SQLCoder-15B
WizardCoder-15B

Llama2
OpenChat
CodeLlama
Nous-Hermes
Vicuna
NSQL
Redmond-Puffin
Llama2-32K
OpenAssistant-Llama2
WizardLM

GPT
GPT-3.5-Turbo
GPT-4

Stable
StableLM
StableCode
OpenAssistant-StableLM

Pythia
Pythia-ChatBase
OpenAssistant-Pythia

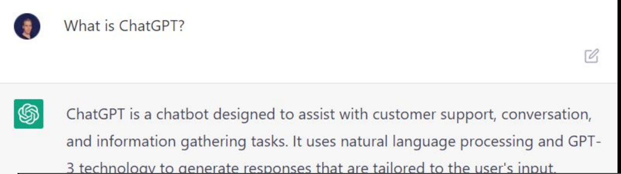
Claude
Claude-2

PaLM
PaLM-2

Falcon
Falcon

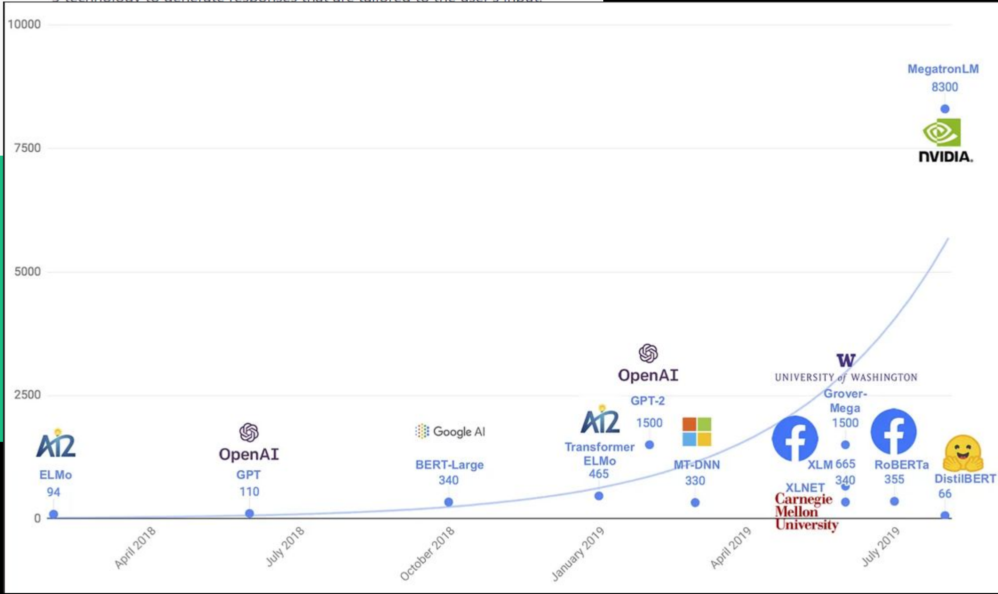
Qwen
Qwen

Map of LLMs
October 2023



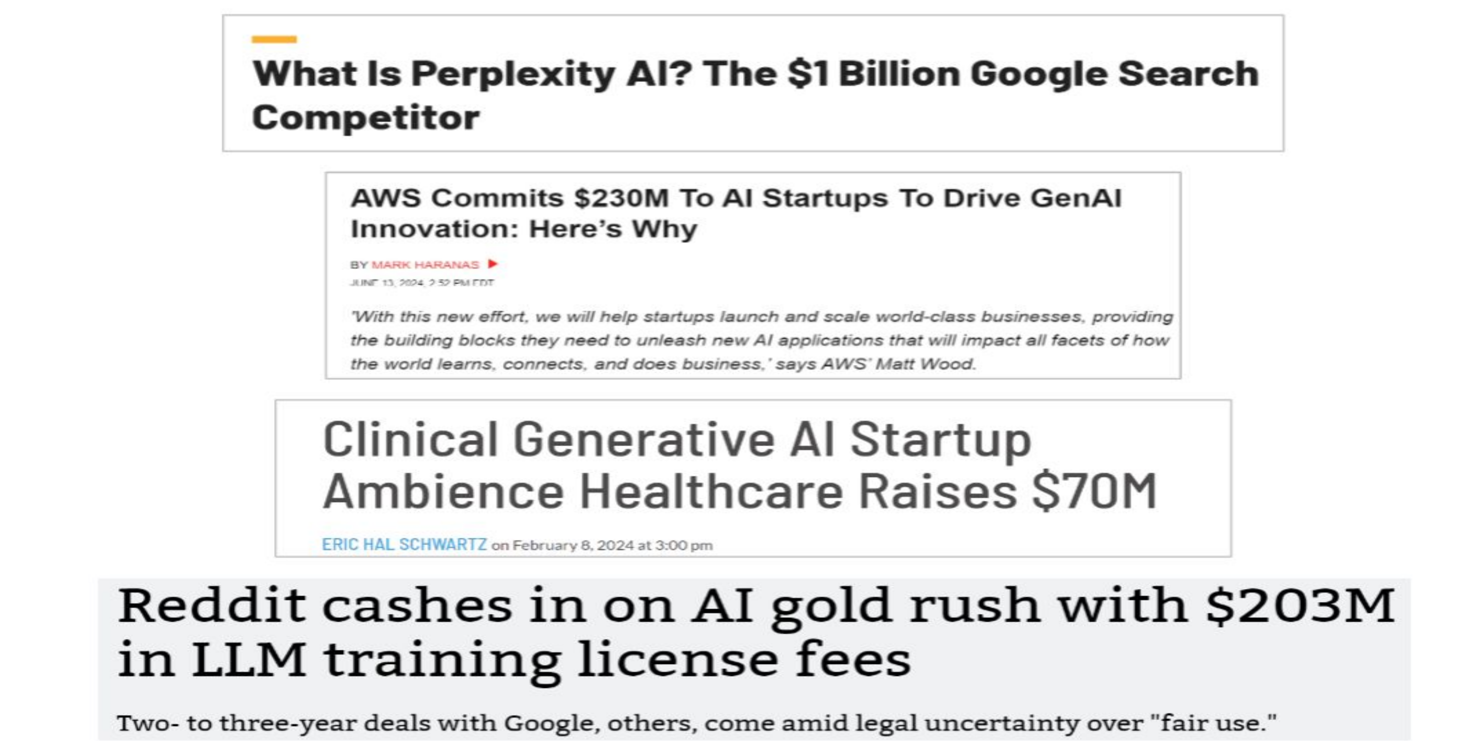
What is ChatGPT?

ChatGPT is a chatbot designed to assist with customer support, conversation, and information gathering tasks. It uses natural language processing and GPT-3 technology to generate responses that are tailored to the user's input.



Source : https://blog.continue.dev/what-llm-to-use/
https://moscow25.medium.com/the-best-deep-natural-language-papers-you-should-read-bert-gpt-2-and-looking-forward-1647f4438797

AI Gold Rush



What Is Perplexity AI? The $1 Billion Google Search Competitor

AWS Commits $230M To AI Startups To Drive GenAI Innovation: Here's Why

BY MARK HARANAS

JUNE 13, 2024, 2:52 PM EDT

'With this new effort, we will help startups launch and scale world-class businesses, providing the building blocks they need to unleash new AI applications that will impact all facets of how the world learns, connects, and does business,' says AWS' Matt Wood.

Clinical Generative AI Startup Ambience Healthcare Raises $70M

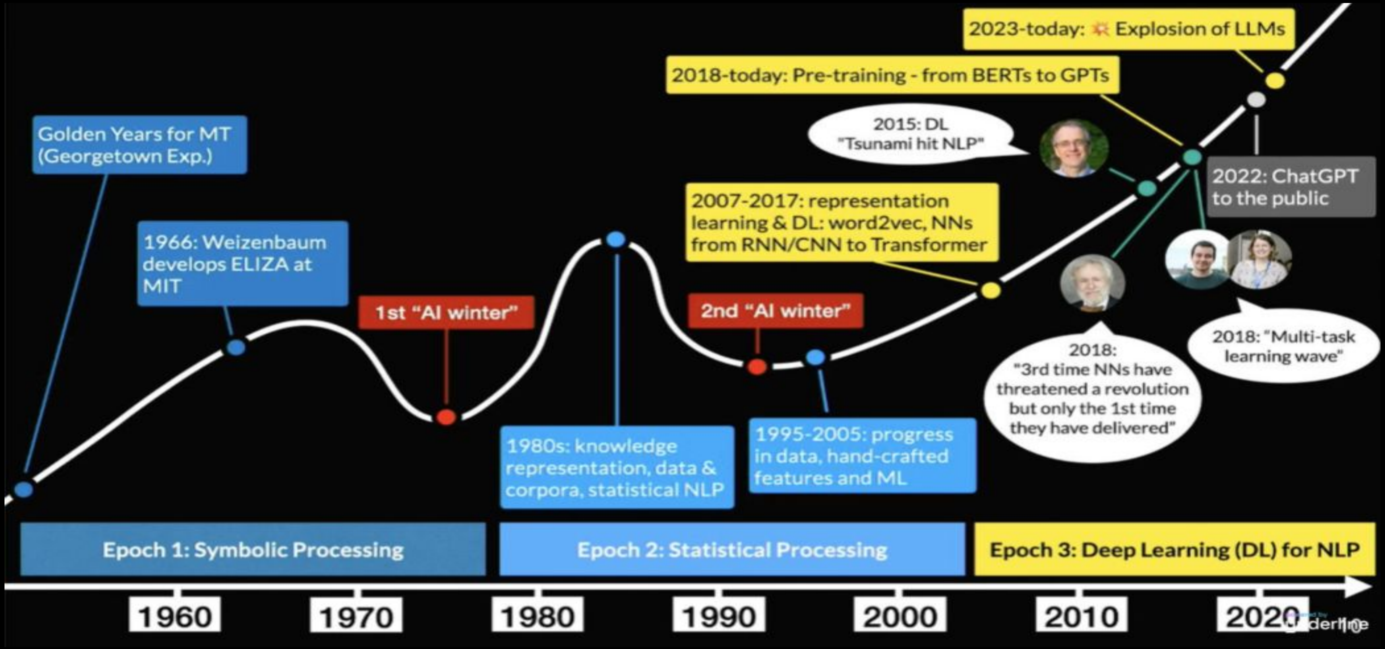
ERIC HAL SCHWARTZ on February 8, 2024 at 3:00 pm

Reddit cashes in on AI gold rush with $203M in LLM training license fees

Two- to three-year deals with Google, others, come amid legal uncertainty over "fair use."

Source : CSE-576 NLP Lecture Slide

Summary of NLP History



Golden Years for MT (Georgetown Exp.)

1966: Weizenbaum develops ELIZA at MIT

1st "AI winter"

1980s: knowledge representation, data & corpora, statistical NLP

2nd "AI winter"

1995-2005: progress in data, hand-crafted features and ML

2007-2017: representation learning & DL: word2vec, NNs from RNN/CNN to Transformer

2015: DL "Tsunami hit NLP"

2018: "3rd time NNs have threatened a revolution but only the 1st time they have delivered"

2018: "Multi-task learning wave"

2018-today: Pre-training - from BERTs to GPTs

2022: ChatGPT to the public

2023-today: 💥 Explosion of LLMs

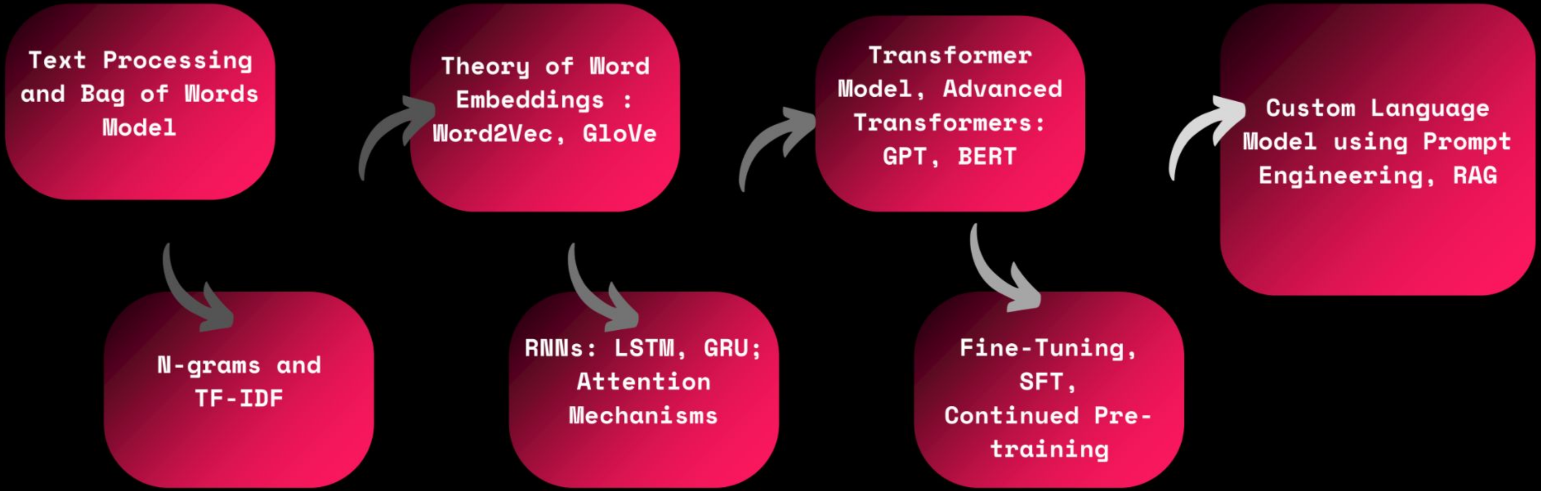
Epoch 1: Symbolic Processing

Epoch 2: Statistical Processing

Epoch 3: Deep Learning (DL) for NLP

Plank, B., Dr (2024, August 14). Are LLMs Widening or Narrowing Our Horizon? Let's Embrace Variation for Trustworthy NLP [Association of Computational Linguistics (ACL) 2024, Keynote 3]. Underline. https://underline.io/events/466/sessions/18198/lecture/104930-keynote-3-barbara-plank

NLP Lab Project : Build your own GPT



Text Processing and Bag of Words Model

N-grams and TF-IDF

Theory of Word Embeddings : Word2Vec, GloVe

RNNs: LSTM, GRU; Attention Mechanisms

Transformer Model, Advanced Transformers: GPT, BERT

Fine-Tuning, SFT, Continued Pre-training

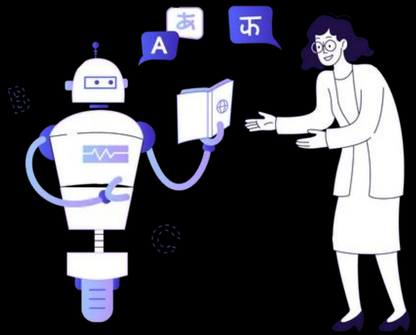
Custom Language Model using Prompt Engineering, RAG

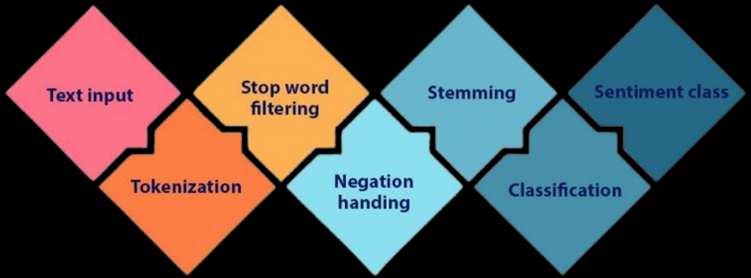
1. Introduction to Natural Language Processing [NLP]

1.1 What is NLP?

Definition: Natural Language Processing is a subfield of artificial intelligence (AI) that focuses on the interaction between computers and humans through natural language.

Applications: Examples include chatbots, sentiment analysis, machine translation, spam detection, and more.




Text input

Tokenization

Stop word filtering

Negation handing

Stemming

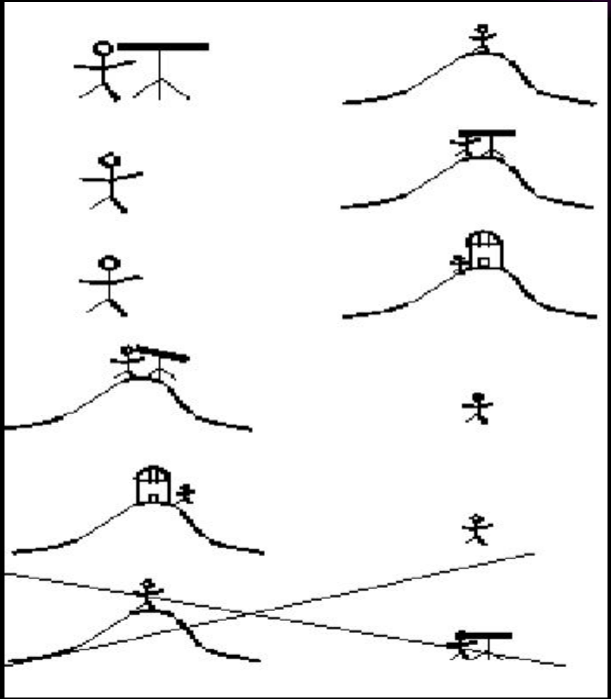
Classification

Sentiment class

1.2 Importance of NLP

- Understanding human language allows machines to better interpret, analyze, and generate text.
- NLP transforms vast amounts of unstructured textual data into actionable insights.
- NLP enables more natural and intuitive communication between humans and machines.
- Processing is crucial in NLP to accurately interpret and disambiguate the complexities of human language.

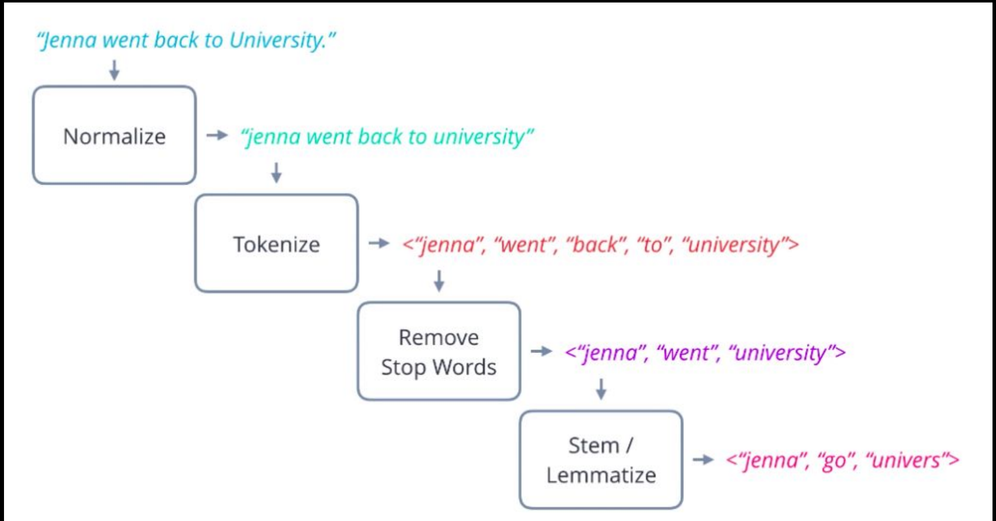
Example: *“I saw a man on a hill with a telescope”*



2. Basic Text Processing

2.1 Text Processing

Definition: Preparing raw text data for analysis by cleaning and transforming it into a format suitable for machine learning models.


"Jenna went back to University."
Normalize
"jenna went back to university"
Tokenize
<"jenna", "went", "back", "to", "university">
Remove Stop Words
<"jenna", "went", "university">
Stem / Lemmatize
<"jenna", "go", "univers">

2.2 Steps in Text Processing

STEP 1

- **Tokenization**
  - **Definition:** Splitting text into individual words (tokens) or phrases.
  - **Types:** Word tokenization, Sub-word tokenization, Sentence tokenization.
  - **Example:**
    - **Input:** "NLP is fun!"
    - **Output:** ["NLP", "is", "fun", "!"]
  - Popular Tokenization Techniques : Byte Pair Encoding, WordPiece, Unigram.

2.2 Steps in Text Processing

STEP 1

- **Byte Pair Encoding - A Sub-word Tokenization technique**
  - Used by a lot of Transformer models, including GPT, GPT-2, RoBERTa, BART, and DeBERTa.
  - How it works?
    - Initialize with Character Tokens
    - Count Character Pairs
    - Merge the Most Frequent Pair
    - Repeat the Process
    - Stop at a Desired Vocabulary Size
  - Advantages : Compact Vocabulary, Handling Rare Words, Efficiency.

2.2 Steps in Text Processing

STEP 2

- **Lowercasing**
    - **Definition:** Converting all characters in text to lowercase to ensure uniformity.
    - **Example:**
        - **Input:** "NLP is FUN"
        - **Output:** "nlp is fun"

2.2 Steps in Text Processing

STEP 3

- **Removing Punctuation and Special Characters**
    - **Definition:** Stripping out punctuation marks, special characters, and numbers.
    - **Tools:** NLTK, SpaCy libraries.
    - **Example:**
        - **Input:** "Hello, World!"
        - **Output:** "Hello World"

2.2 Steps in Text Processing

STEP 4

- **Removing Stopwords**
    - **Definition:** Removing common words (e.g., "the", "is", "in") that don’t contribute much to the meaning.
    - **Tools:** Regex or built-in Python string functions.
    - **Example:**
        - **Input:** "NLP is fun and interesting"
        - **Output:** "NLP fun interesting"

2.2 Steps in Text Processing

STEP 5

- **Stemming:** Reducing words to their base or root form.
    - **Example:** "running", "runner" -> "run".
    - **Libraries:** NLTK (PorterStemmer, SnowballStemmer).
- **Lemmatization:** Reducing words to their dictionary form (lemma).
    - **Example:** "running" -> "run", "better" -> "good".
    - **Libraries:** NLTK, SpaCy.

2.2 Steps in Text Processing

STEP 6

- Normalization
    - **Definition:** Converting different forms of a word to a single form.
    - **Example:**
        - Converting numbers into words, handling contractions (e.g., "don't" to "do not").

3. Bag of Words (BoW) Model

3.1 Introduction to Bag of Words (BoW)

Definition: A BoW model represents text as a collection of word frequencies or occurrences without considering grammar and word order.

How it works:

- **Vocabulary Creation:** Create a list of all unique words (vocabulary) from the text corpus.
- **Vectorization:** Convert text documents into vectors of word counts or binary values (presence/absence).

3.1 Introduction to Bag of Words (BoW)

Example:

Documents:

- "NLP is fun"
- "I love NLP"

Vocabulary: ["NLP", "is", "fun", "I", "love"]

BoW representation:

- Doc 1: [1, 1, 1, 0, 0]
- Doc 2: [1, 0, 0, 1, 1]

3.2 Advantages and Limitations of BoW

Advantages

- Simple and easy to implement.
- Works well with smaller datasets.
- Effective for text classification tasks.

Limitations

- Ignores the order and semantics of words.
- Can lead to a sparse matrix for large vocabularies.
- Does not handle synonyms well.
- Sensitive to the presence of stopwords and noisy data.

4. Practical Applications of Text Processing and BoW Model

4. Text Processing & BoW Model

Text Classification: Use BoW to convert text into numerical format for machine learning models (e.g., spam detection).

Sentiment Analysis: Analyze sentiment of reviews or social media posts using BoW and text preprocessing techniques.

Information Retrieval: Create search algorithms that match queries to documents based on word frequencies.

Recommendation Systems: Build simple recommendation systems based on text similarity using BoW vectors.

Kahoot Time!

Group Picture!