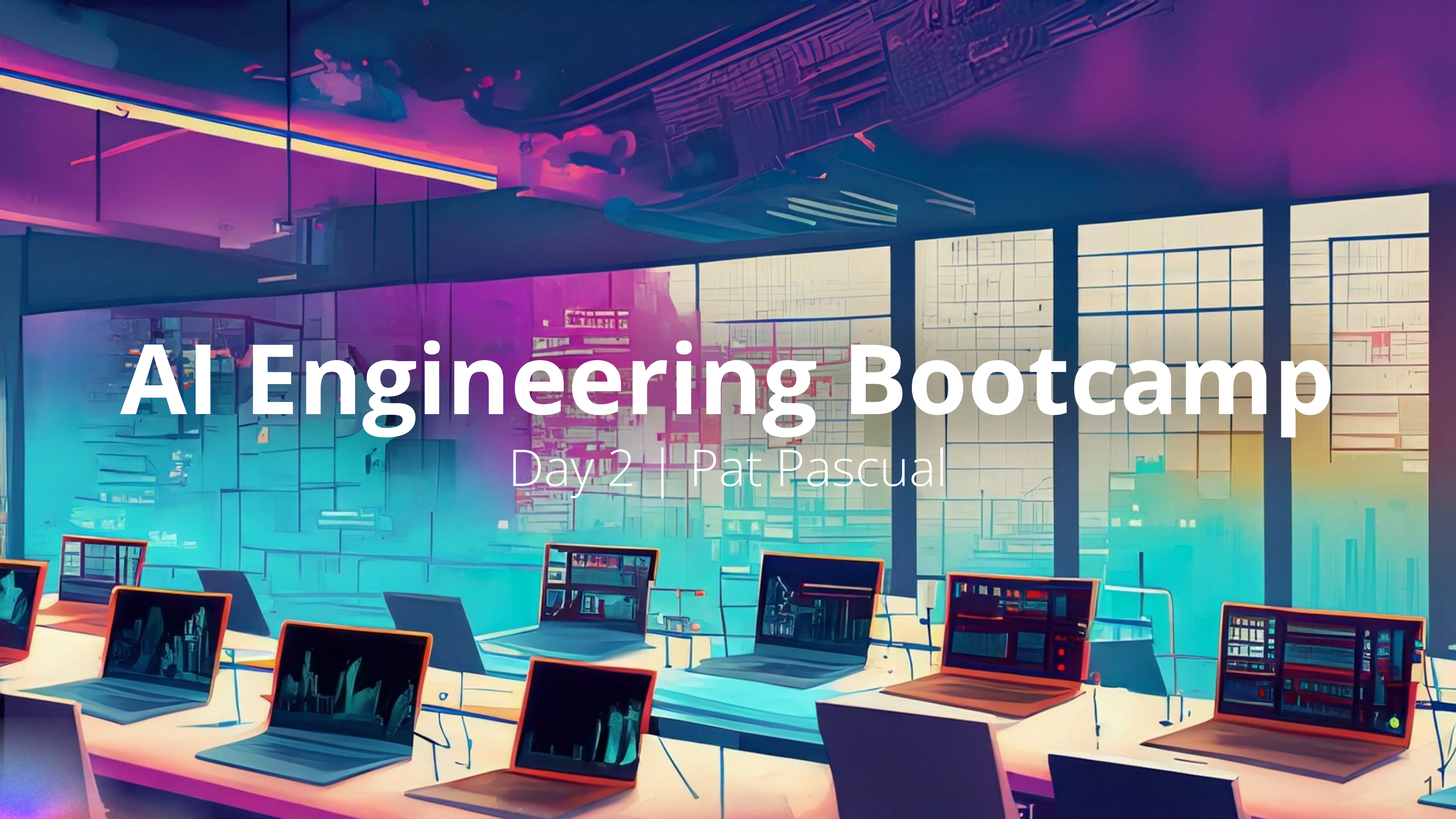


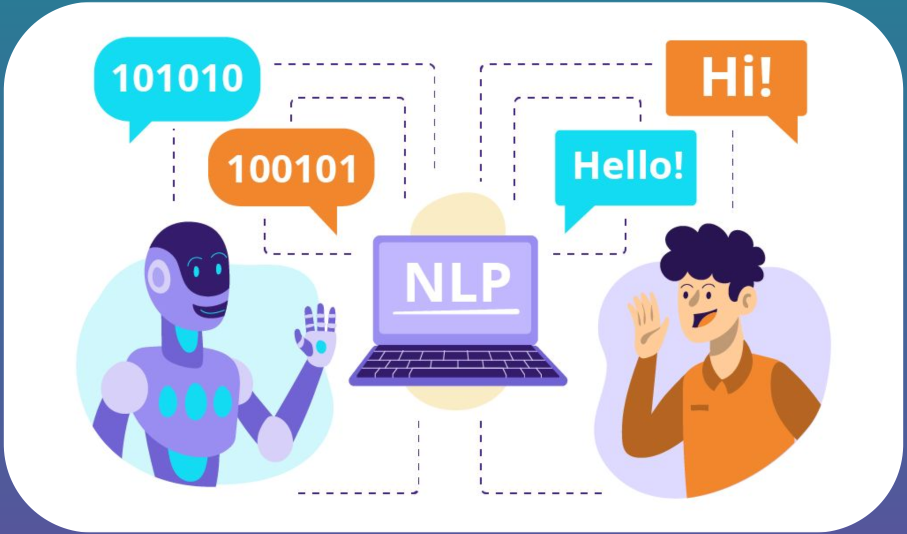
AI Engineering Bootcamp

Day 2 | Pat Pascual

Introduction to NLP

Intro to NLP

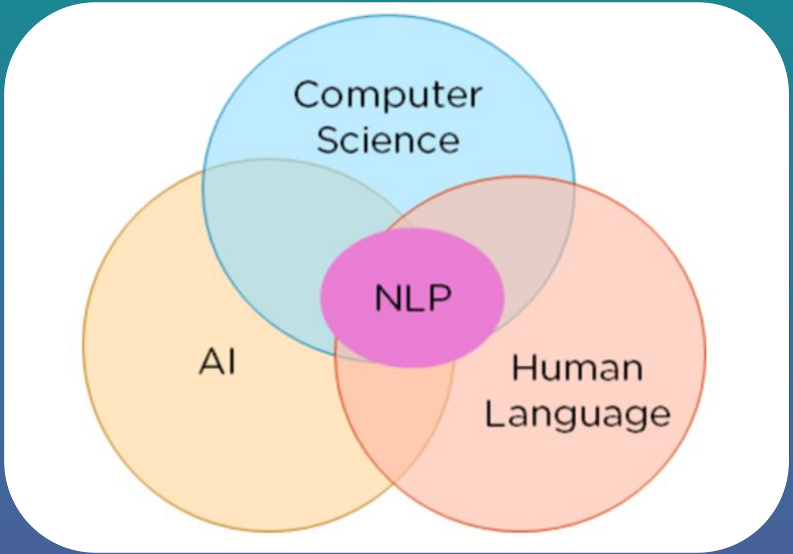
Natural Language Processing is the bridge between language and computation by converting human language into machine-understandable data.



Intro to NLP

The integration of NLP techniques and data science methodologies allows for robust and insightful analysis of textual data.

The synergy of NLP and data science paves the way for extracting meaningful patterns, insights, and information.



Real World Applications of NLP

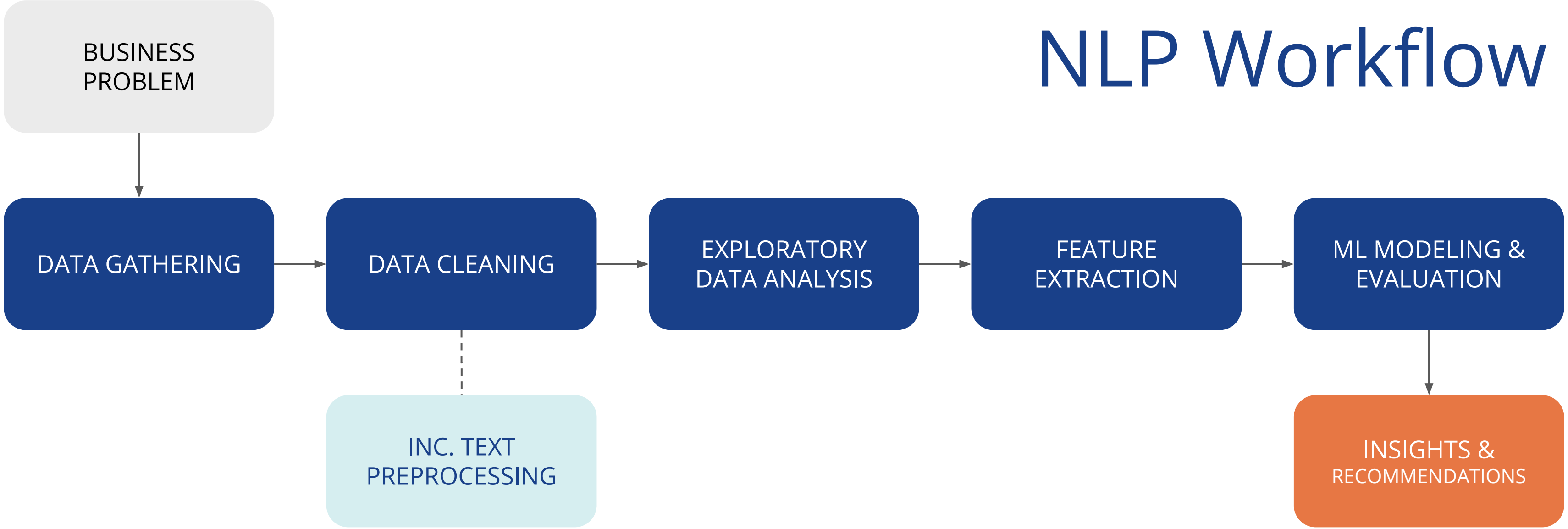
Consumer Products

Research

AI Chatbots

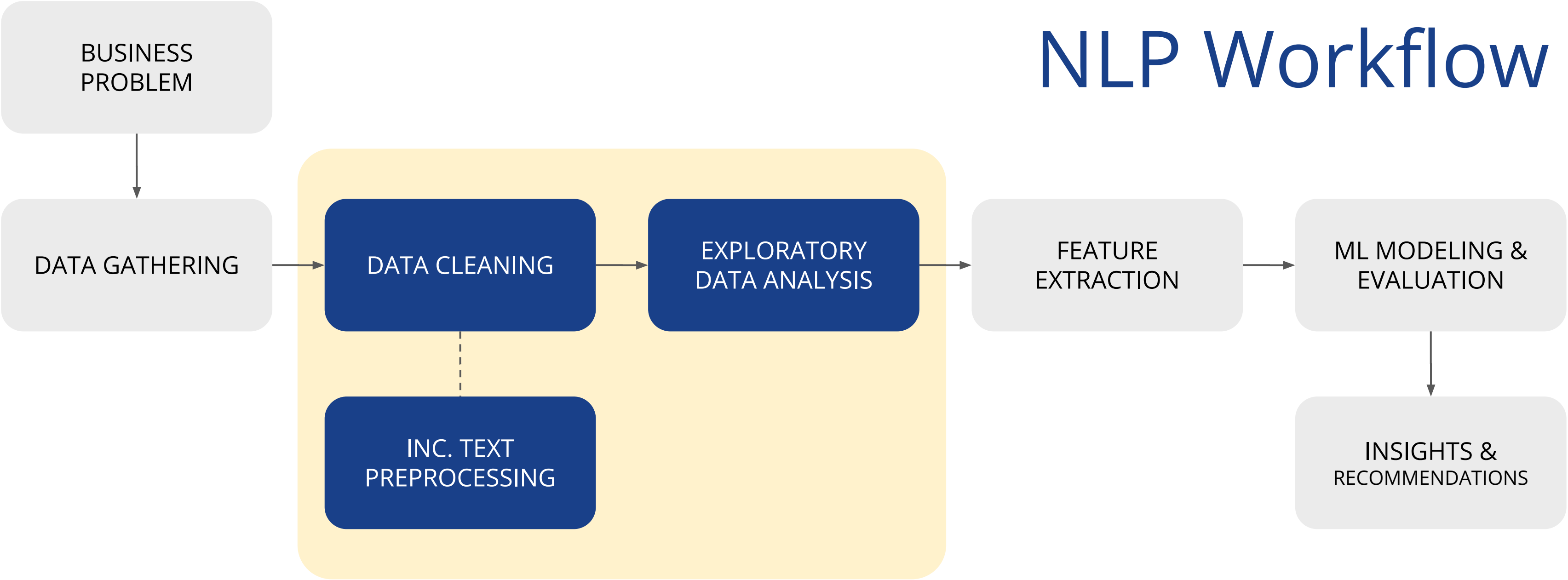
Concept - NLP Workflow

NLP Workflow



Concept - NLP Workflow

NLP Workflow



Text Preprocessing

Concept - Text Preprocessing

What is text preprocessing?

Text preprocessing is where human language is meticulously transformed into a structured format that is easily interpretable by machines.

The process includes, cleaning, normalizing, and organizing raw text.

Concept - Text Preprocessing

NLTK

- Emphasis on teaching NLP concepts
- Wide range of tools for various NLP tasks
- Greater flexibility and customizability
- Ideal for academic and educational purposes

SpaCy

- High-performance, optimized for speed
- Pre-trained models for immediate use
- User-friendly, streamlined API
- Suitable for production and industrial applications

Concept - Text Preprocessing Techniques

Common Text Preprocessing Techniques

- **Tokenization**
- **Noise Removal**
- **Stemming**
- **Lemmatization**

Concept - Tokenization

Tokenization

Process of splitting text into individual words, phrases, or other meaningful elements (tokens).

Done using libraries like NLTK or spaCy, which provide functions to easily tokenize text

Input: **"Hello, world!"**
Output: **["Hello", ",", "world", "!"]**

Concept - Noise Removal

Removing Noise

Involves filtering out irrelevant or extraneous data, such as special characters, numbers, or stop words

Done by defining a list of noise elements and using string manipulation or regular expressions to remove them from the text

Input: **"Hello! Are you there?? #excited"**
Output: **"Hello Are you there excited"**

Concept - Stemming

Stemming

Process of reducing words to their base or root form, often leading to a rough approximation

Often performed using the NLTK library's PorterStemmer or SnowballStemmer

Input: **"running, flies, denying"**
Output: **["run", "fli", "deni"]**

Concept - Lemmatization

Lemmatization

Process of converting words to their dictionary form,
considering the context and part of speech

Done using the spaCy library, which considers the word's part of speech
to find the correct lemma

Input: **"running, flies, denying"**
Output: **["run", "fly", "deny"]**

Exploratory Data Analysis with NLP

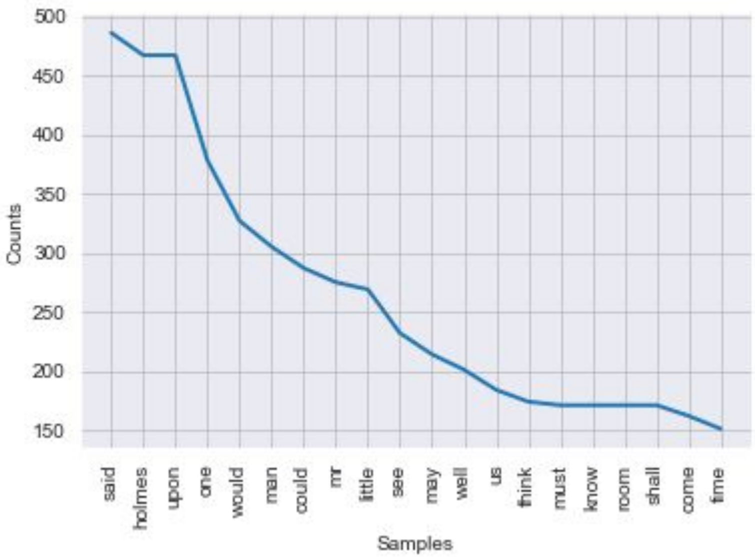
EDA with NLP

What are common EDA techniques that use NLP?

- Text Statistics
- Word Cloud
- N-gram Visualizations

Concept - Textual Statistics

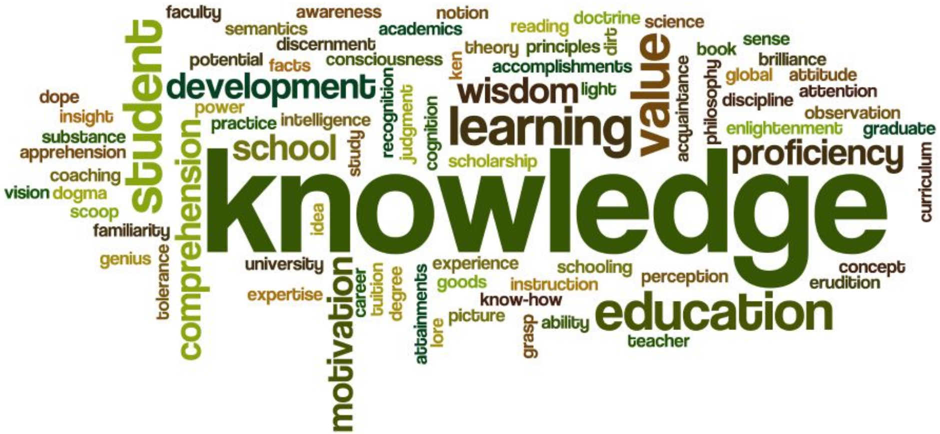
Text statistics involve quantitative analysis of text data, focusing on metrics like word frequency, document length, and lexical diversity



Concept - Word Cloud

NLP visualizations, like **word clouds**, provide graphical representations of text data to facilitate better understanding and insights into its underlying structure.

Utilized in exploring key themes in large textual datasets like books.



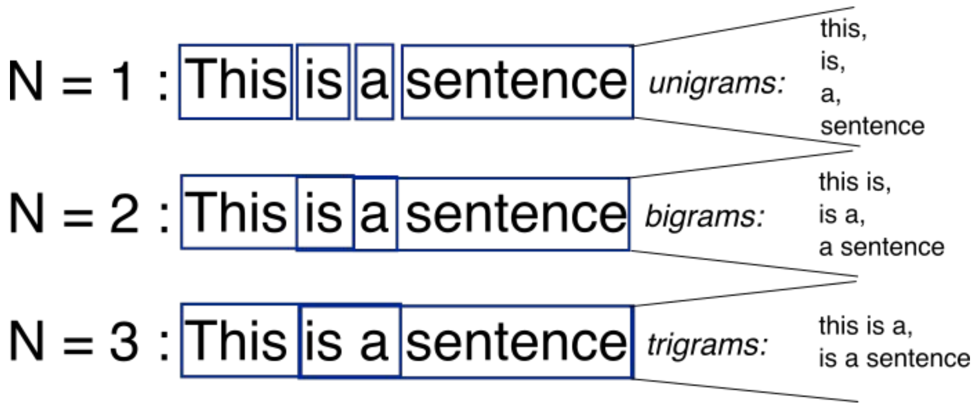
N-Grams

Concept - N-Grams

What are n-grams?

N-grams are contiguous sequences of *n items* from a given sample of text or speech, used to predict the next item in such sequences.

Employed in AI chatbots for better context recognition and response generation



N = 1 : This is a sentence unigrams: this, is, a, sentence

N = 2 : This is a sentence bigrams: this is, is a, a sentence

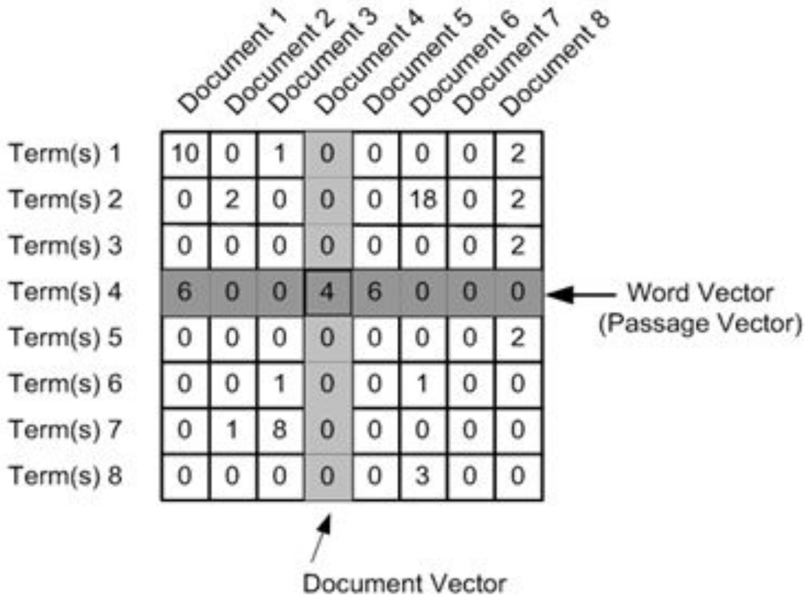
N = 3 : This is a sentence trigrams: this is a, is a sentence

Text Vectorization

Concept - Text Vectorization

Text vectorization

- Converts text into numerical values for algorithms to process
- Higher word weight = More descriptive of the document
- Common vectorization methods
  - **Count Vectorizer**: Counts word occurrences as weights
  - **TF-IDF Vectorizer**: Weighs words based on frequency & uniqueness across documents



	Document 1	Document 2	Document 3	Document 4	Document 5	Document 6	Document 7	Document 8
Term(s) 1	10	0	1	0	0	0	0	2
Term(s) 2	0	2	0	0	0	18	0	2
Term(s) 3	0	0	0	0	0	0	0	2
Term(s) 4	6	0	0	4	6	0	0	0
Term(s) 5	0	0	0	0	0	0	0	2
Term(s) 6	0	0	1	0	0	1	0	0
Term(s) 7	0	1	8	0	0	0	0	0
Term(s) 8	0	0	0	0	0	3	0	0

Word Vector (Passage Vector)

Document Vector

Large Language Models

Concept - LLMs

1

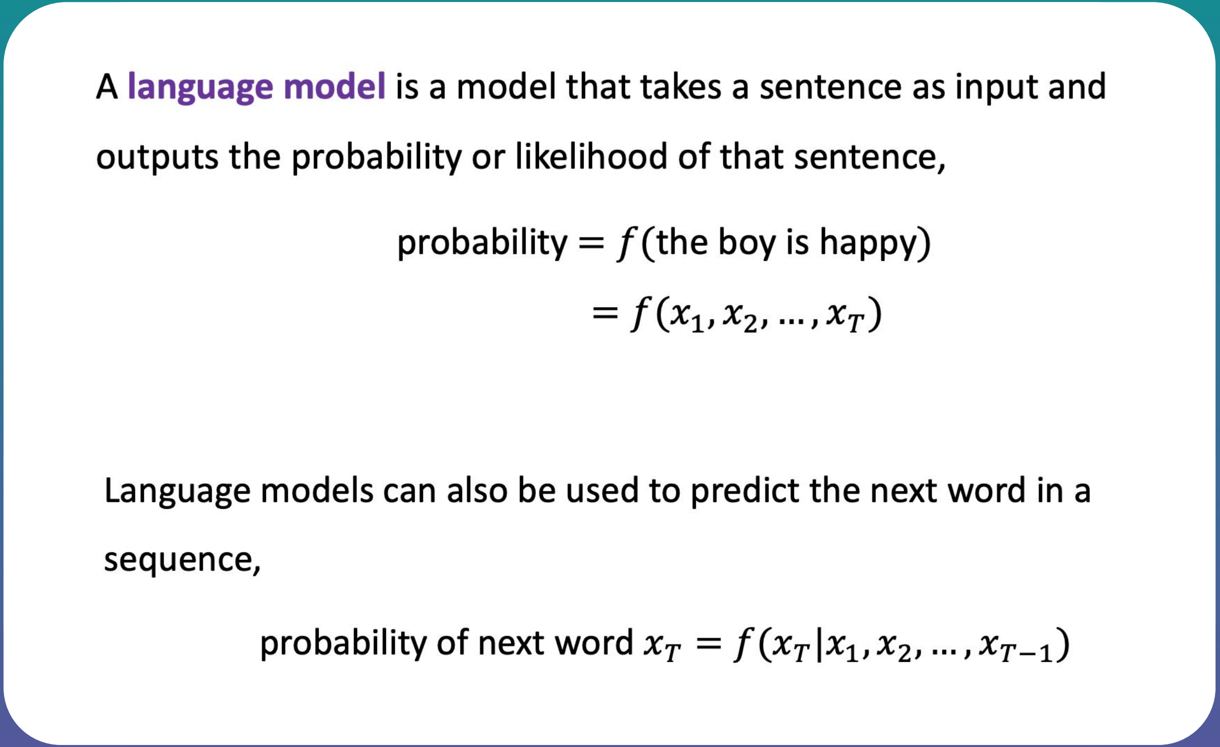
Vast Data

2

Self-Supervised Learning

3

Iteration



A **language model** is a model that takes a sentence as input and outputs the probability or likelihood of that sentence,

$$\text{probability} = f(\text{the boy is happy})$$
$$= f(x_1, x_2, \dots, x_T)$$

Language models can also be used to predict the next word in a sequence,

$$\text{probability of next word } x_T = f(x_T | x_1, x_2, \dots, x_{T-1})$$

LM definition from S. Ibanez (2023), Machine Learning 3 Lecture, Asian Institute of Management.

Concept - LLMs

A **large language model** uses deep neural networks to generate outputs based on patterns learned from training data.

(Current) **TRANSFORMERS**
Uses self-attention, like reading a whole sentence at once

(Old) **RECURRENT NEURAL NETWORKS**
Reading word by word

Sample Transformer-Based LLMs

GPT-4 (*Generative Pretrained Transformer 4*) – developed by OpenAI.

BERT (*Bidirectional Encoder Representations from Transformers*) – developed by Google.

RoBERTa (*Robustly Optimized BERT Approach*) – developed by Facebook AI.

T5 (*Text-to-Text Transfer Transformer*) – developed by Google.

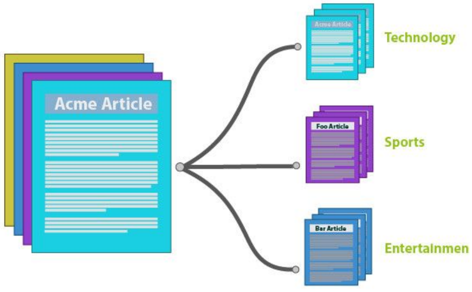
CTRL (*Conditional Transformer Language Model*) – developed by Salesforce Research.

Megatron-Turing – developed by NVIDIA

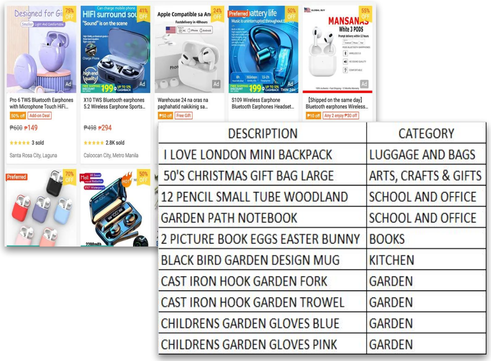
Text Classification with LLMs

Concept - Text Classification

Text Classification is the task of assigning a label or class to a given text.



Article Classification



DESCRIPTION	CATEGORY
I LOVE LONDON MINI BACKPACK	LUGGAGE AND BAGS
50'S CHRISTMAS GIFT BAG LARGE	ARTS, CRAFTS & GIFTS
12 PENCIL SMALL TUBE WOODLAND	SCHOOL AND OFFICE
GARDEN PATH NOTEBOOK	SCHOOL AND OFFICE
2 PICTURE BOOK EGGS EASTER BUNNY	BOOKS
BLACK BIRD GARDEN DESIGN MUG	KITCHEN
CAST IRON HOOK GARDEN FORK	GARDEN
CAST IRON HOOK GARDEN TROWEL	GARDEN
CHILDRENS GARDEN GLOVES BLUE	GARDEN
CHILDRENS GARDEN GLOVES PINK	GARDEN

Product Categorization

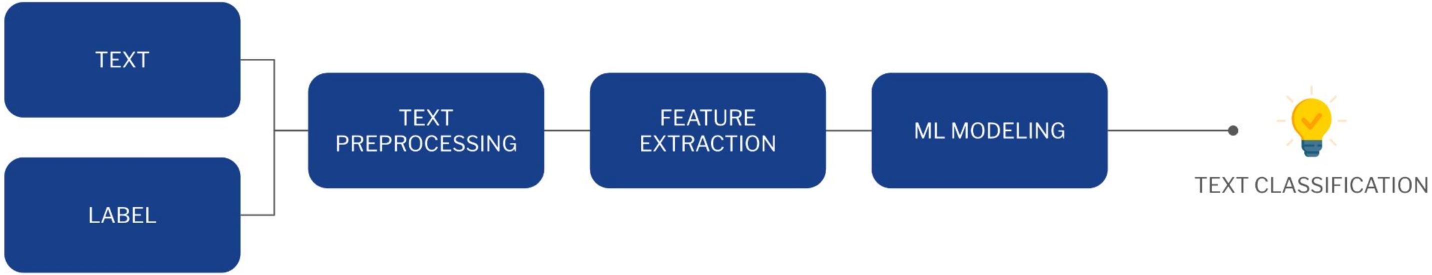
SPAM

NOT SPAM

Email Classification

(Ex) Text Classification Pipeline

Supervised Text Classification


TEXT

LABEL

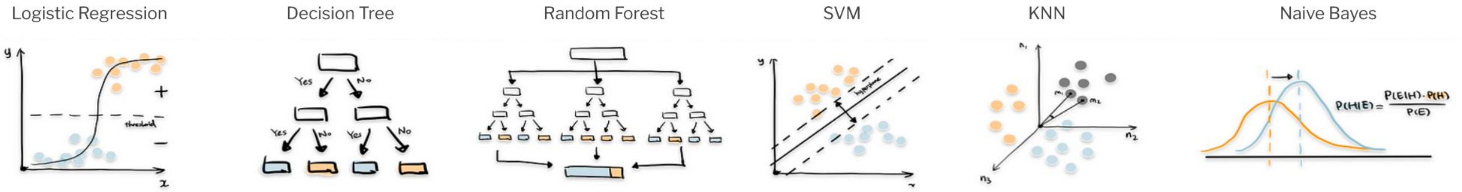
TEXT PREPROCESSING

FEATURE EXTRACTION

ML MODELING

TEXT CLASSIFICATION

CLASSIFICATION ALGORITHMS


Logistic Regression

Decision Tree

Random Forest

SVM

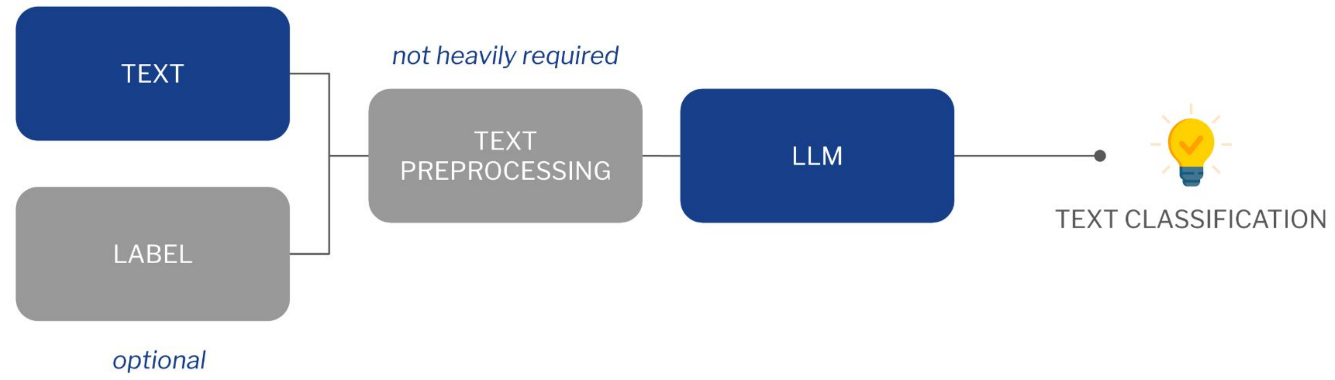
KNN

Naive Bayes

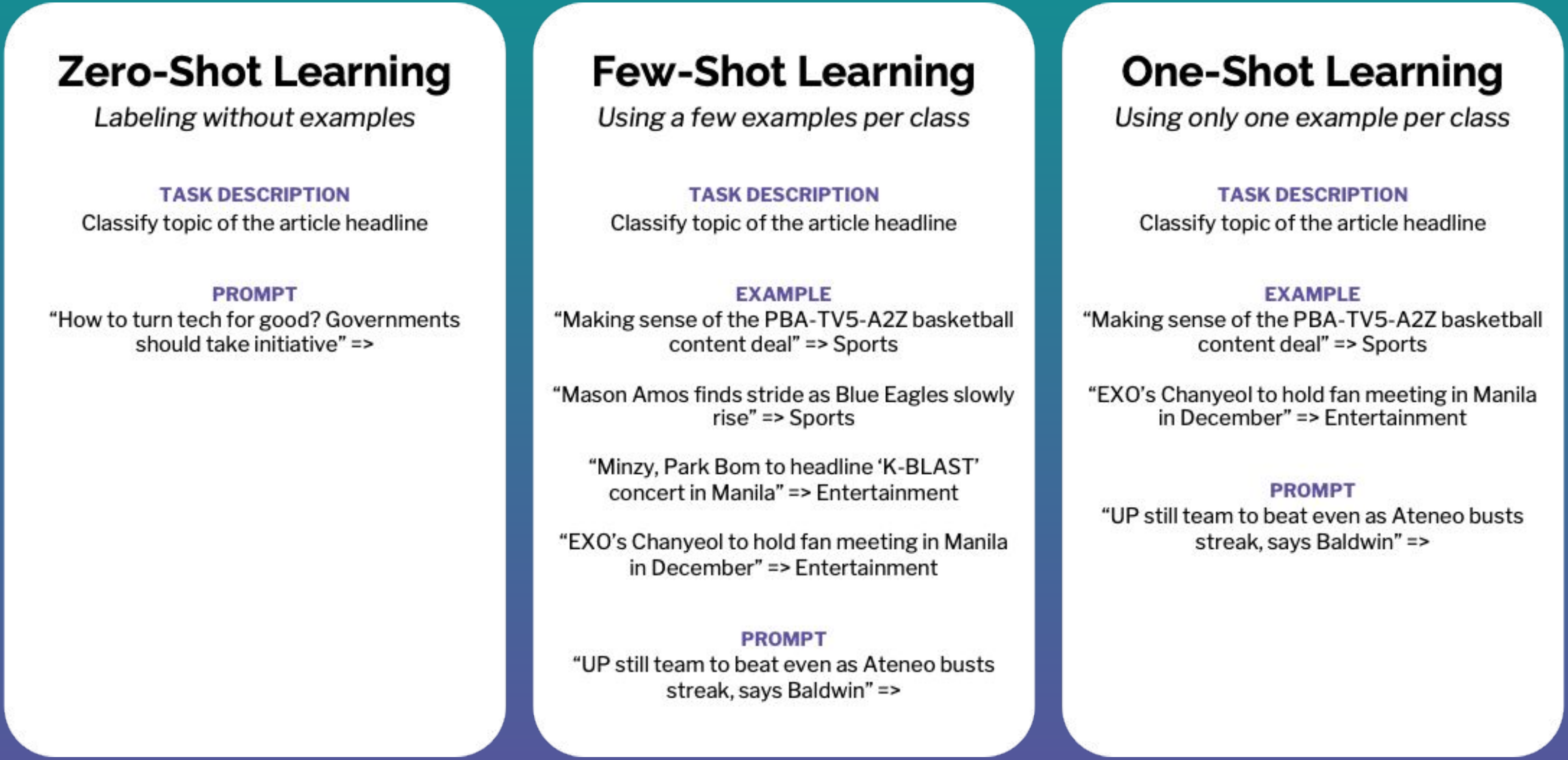
(Ex) Text Classification Pipeline

Text Classification with LLMs

Models like GPT-4 are capable of **zero-shot learning** due to their pre-trained language understanding.



Concept - Few-Shot Learning



Zero-Shot Learning

Labeling without examples

TASK DESCRIPTION

Classify topic of the article headline

PROMPT

“How to turn tech for good? Governments should take initiative” =>

Few-Shot Learning

Using a few examples per class

TASK DESCRIPTION

Classify topic of the article headline

EXAMPLE

“Making sense of the PBA-TV5-A2Z basketball content deal” => Sports

“Mason Amos finds stride as Blue Eagles slowly rise” => Sports

“Minzy, Park Bom to headline ‘K-BLAST’ concert in Manila” => Entertainment

“EXO’s Chanyeol to hold fan meeting in Manila in December” => Entertainment

PROMPT

“UP still team to beat even as Ateneo busts streak, says Baldwin” =>

One-Shot Learning

Using only one example per class

TASK DESCRIPTION

Classify topic of the article headline

EXAMPLE

“Making sense of the PBA-TV5-A2Z basketball content deal” => Sports

“EXO’s Chanyeol to hold fan meeting in Manila in December” => Entertainment

PROMPT

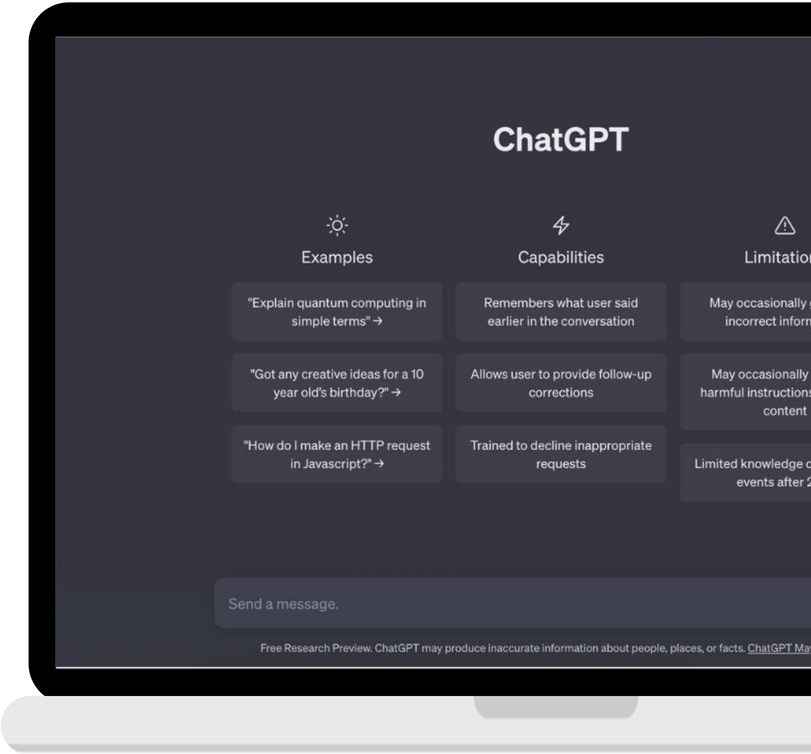
“UP still team to beat even as Ateneo busts streak, says Baldwin” =>

Prompting Best Practices

Custom Prompting

Why do we need more flexibility in structuring our prompts?

- Instead of assigning predefined labels, we may want to **extract insights**, **summarize**, or **generate structured responses**.
    - Example: Instead of **classifying sentiment**, we may want the model to **explain why a review is negative** or **suggest improvements**.
- The quality of the output depends on how well we **structure our prompts**.



ChatGPT

Examples

"Explain quantum computing in simple terms" →

"Got any creative ideas for a 10 year old's birthday?" →

"How do I make an HTTP request in Javascript?" →

Capabilities

Remembers what user said earlier in the conversation

Allows user to provide follow-up corrections

Trained to decline inappropriate requests

Limitatio

May occasionally incorrect inforn

May occasionally harmful instruction content

Limited knowledge o events after 2

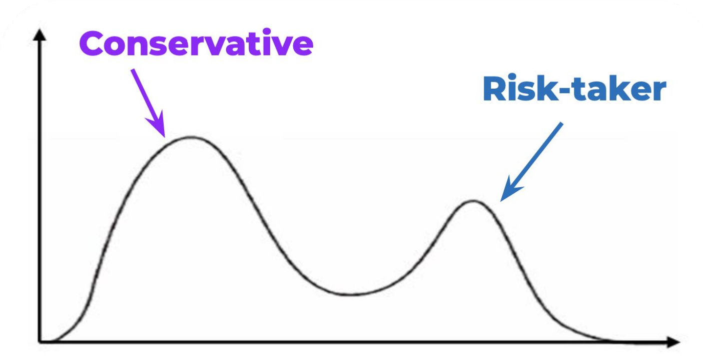
Send a message.

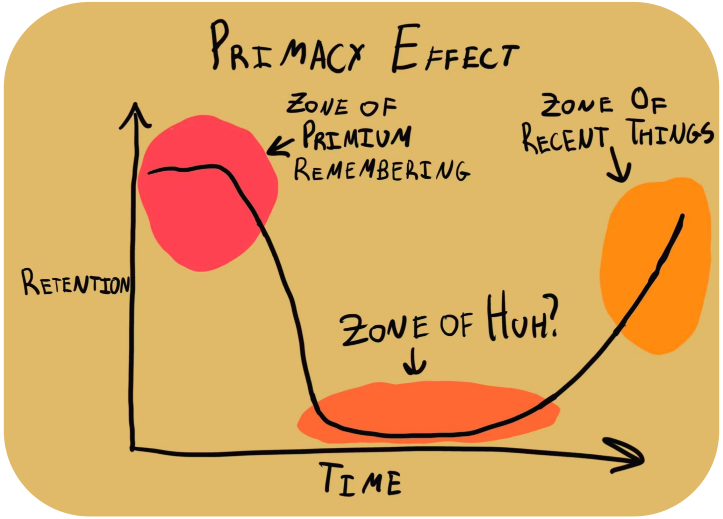
Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT Ma

Concept - Prompting Best Practices

How should we be constructing our prompts?

- Instructing
  - Inspire from multimodal distribution
  - Recency bias





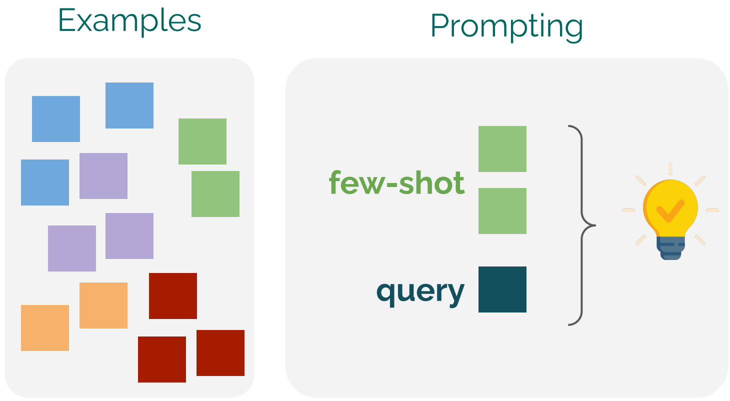
Concept - Prompting Best Practices

How should we be constructing our prompts?

- Few-shot
  - Cover the output dimensions
  - Make the few-shot as relevant as possible

Persona

Length	Short/ Average Joe	Short/ Expert
	Long/ Average Joe	Long/ Expert



Concept - Prompting Best Practices

How should we be constructing our prompts?

- Reasoning
    - Chain of thoughts

“Let’s think step by step.”