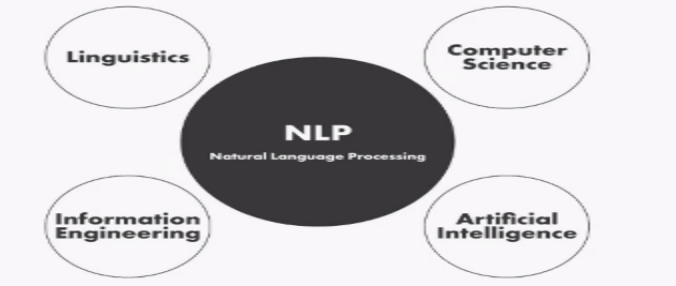
NLP: Day 1

Introduction to NLP:

- **Definitions:**
  - NLP is a field of study focused on teaching computers to understand and process human language
  - NLP combines principles from linguistics, computer science, artificial intelligence, and information engineering to develop algorithms and models that enable computers to understand, process, and generate human language in a way that is useful and meaningful for various applications..
    - **Linguistics:** Linguistics is the scientific study of language. It explores how languages are structured, how they sound, and how meaning is conveyed through words and sentences.
      - For example, linguists might study different grammatical rules in languages or analyze the sounds and patterns in speech.
    - **Computer Science:** Computer science is all about computers and the technologies that power them. It involves creating software programs and building computer systems. Computer scientists develop algorithms and write code that allows computers to perform various tasks, such as playing games, browsing the internet, or processing data.
    - **Artificial Intelligence:** Artificial intelligence is the field that focuses on creating computer systems that can imitate human intelligence. AI enables machines to learn from data, recognize patterns, and make decisions.
      - Examples of AI include voice assistants like Siri or Alexa, self-driving cars, and recommendation systems that suggest movies or products based on your preferences.
    - **Information Engineering:** Information engineering is about effectively managing and using information. It involves designing systems to collect, store, and process data, ensuring that information is organized, accessible, and secure.
      - For instance, information engineers might develop databases, create websites, or implement data analytics tools to extract insights from large datasets.



- **Applications of NLP:**
    - **Text Analysis:** NLP algorithms can analyze and extract information from text data. For example, sentiment analysis can determine the sentiment (positive, negative, or neutral) expressed in customer reviews or social media posts.
    - **Speech Recognition:** NLP enables computers to convert spoken language into written text. Applications like voice assistants (e.g., Siri, Alexa) use NLP to understand and respond to voice commands.
    - **Language Translation:** NLP algorithms can translate text from one language to another. For instance, platforms like Google Translate utilize NLP techniques to provide instant translations between various languages.
    - **Chatbots:** NLP is integral to the development of chatbots. These virtual assistants can understand and respond to user queries, providing automated customer support or information. Chatbots employ NLP to comprehend user input and generate relevant responses.
    - **Named Entity Recognition (NER):** NLP algorithms can identify and extract named entities such as people, organizations, locations, or dates from text. For example, NER can extract names of individuals or organizations from news articles.
    - **Text Summarization:** NLP techniques can condense lengthy documents or articles into shorter summaries while retaining key information. This is useful for quick information retrieval or summarizing news articles.
    - **Question-Answering Systems:** NLP algorithms can be used to build question-answering systems that understand user questions and provide relevant answers. Examples include virtual assistants like IBM Watson or search engines like Google's featured snippets.
    - **Sentiment Analysis:** NLP can determine the sentiment expressed in a piece of text, such as positive, negative, or neutral. This is useful for analyzing customer feedback, social media sentiment, or product reviews.
    - **Language Generation:** NLP can be utilized to generate human-like text. Examples include generating product descriptions, writing news articles, or composing personalized emails.
- **Importance of NLP:**
    - NLP enables computers to understand, analyze, and generate human language, powering applications such as text analysis, speech recognition, language translation, chatbots, and more.

NLP in the Real World/Core applications:

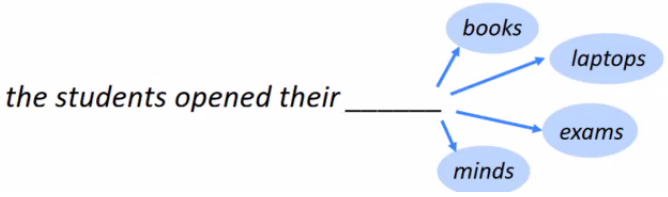
- Email platforms, such as Gmail, Outlook, etc., use NLP extensively to provide a range of product features, such as spam classification, priority inbox, calendar event extraction, auto-complete, etc.
- Voice-based assistants, such as Apple Siri, Google Assistant, Microsoft Cortana, and Amazon Alexa rely or a range of NLP techniques to interact with the user, understand user commands, and respond accordingly.
- Modern search engines, such as Google and Bing, which are the cornerstone of today's internet, use NLP heavily for various subtasks, such as query understanding, query expansion, question answering, information retrieval, and ranking and grouping of the results so on.
- Machine translation services, such as Google Translate, Bing Microsoft Translator, and Amazon Translate are increasingly used in today's world to solve a wide range of scenarios and business use cases.
- Chatbots

Note: *Richard Bandler & Joh Grinder introduced NLP.*

Various NLP Tasks:

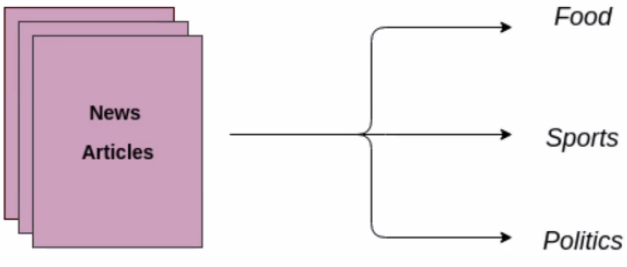
- **Language Modeling:**

  This is the task of predicting what the next word in a sentence will be based on the history of previous words. The goal of this task is to learn the probability of a sequence of words appearing in a given language. Language modeling is useful for building solutions for a wide variety of problems, such as speech recognition, optical character recognition, handwriting recognition, machine translation, and spelling correction.



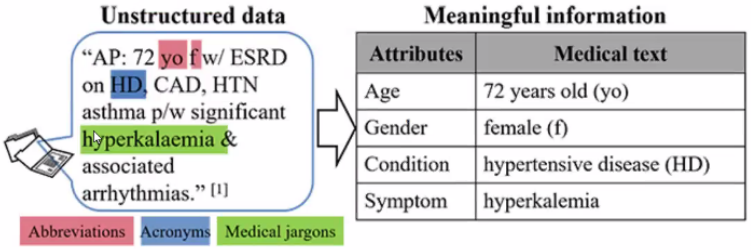
- **Text Classification:**

  This is the task of bucketing the text into a known set of categories based on its content. Text classification is by far the most popular task in NLP and is used in a variety of tools, from email spam identification to sentiment analysis.



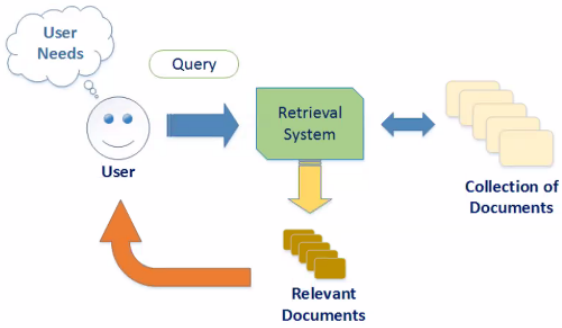
- **Information extraction:**

As the name indicates, this is the task of extracting relevant information from text, such as calendar events from emails or the names of people mentioned in a social media post.



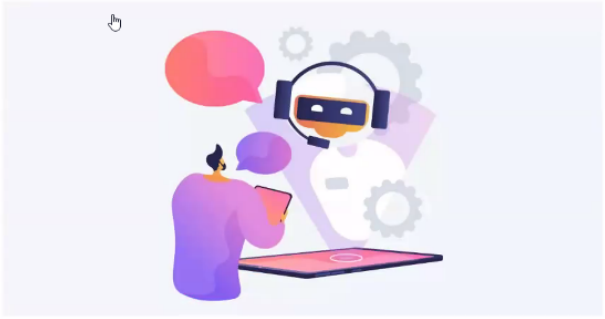
- **Information retrieval:**

This is the task of finding documents relevant to a user query from a large collection. Applications like Google Search are well-known use cases of information retrieval.



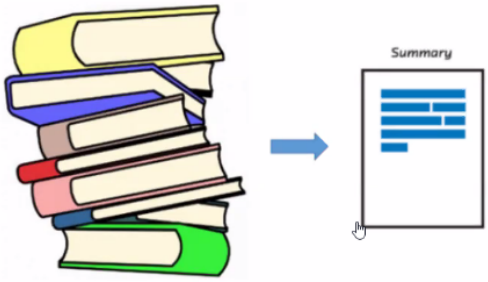
- **Conversational Agent:**

This is the task of building dialogue systems that can converse in human languages. Alexa, Siri, etc., are some common applications of this task.



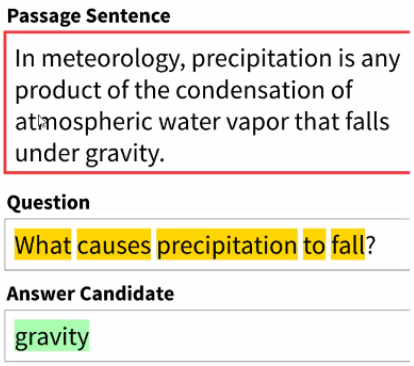
- **Text Summarization:**

  This task aims to create short summaries of longer documents while retaining the core content and preserving the overall meaning of the text.



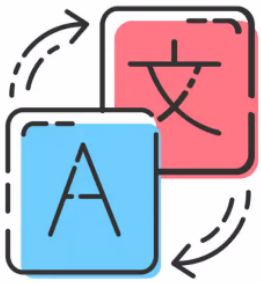
- **Question & Answering:**

  This is the task of building a system that can automatically answer questions posed in natural language.



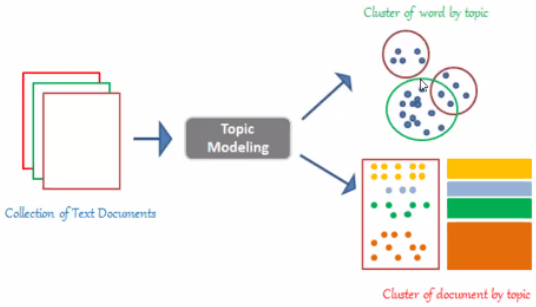
- **Machine Translation:**

  This is the task of converting a piece of text from one language to another. Tools like Google Translate are common applications of this task.

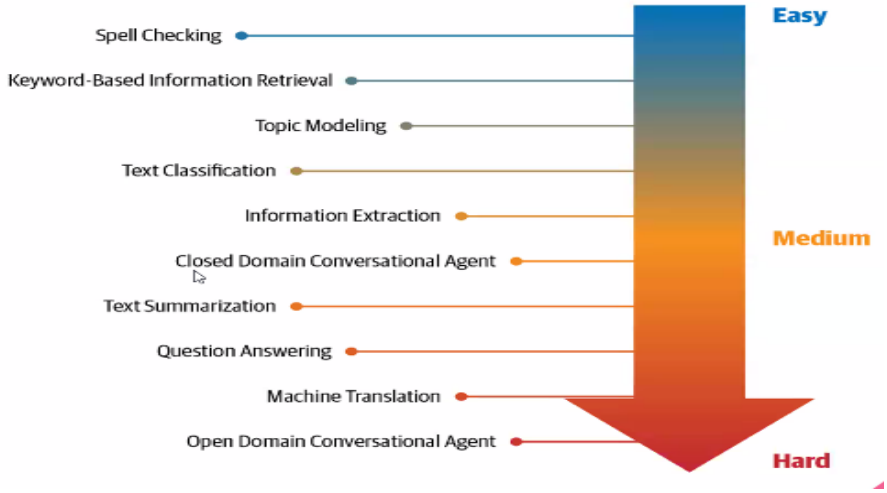


- **Topic Modeling:**

  This is the task of uncovering the topical structure of a large collection of documents. Topic modeling is a common text-mining tool and is used in a wide range of domains, from literature to bioinformatics.



Difficulty in terms of developing comprehensive solutions:



What is Language?

Language is a structured system of communication that involves complex combinations of its constituent components, such as characters, words, sentences, etc.

We can think of human language as composed of four major building blocks: **phonemes, morphemes and lexemes, syntax, and context**.

Building Block of Language:

- **Phonemes:**

  Phonemes are the smallest units of sound in a language. They may not have any meaning by themselves but can induce meanings when uttered in combination with other phonemes.

Consonant phonemes, with sample words		Vowel phonemes, with sample words	
1. /b/ - bat	13. /s/ - sun	1. /a/ - ant	13. /oi/ - coin
2. /k/ - cat	14. /t/ - tap	2. /e/ - egg	14. /ar/ - farm
3. /d/ - dog	15. /v/ - van	3. /i/ - in	15. /or/ - for
4. /f/ - fan	16. /w/ - wig	4. /o/ - on	16. /ur/ - hurt
5. /g/ - go	17. /y/ - yes	5. /u/ - up	17. /air/ - fair
6. /h/ - hen	18. /z/ - zip	6. /ai/ - rain	18. /ear/ - dear
7. /j/ - jet	19. /sh/ - shop	7. /ee/ - feet	19. /ure/[4] - sure
8. /l/ - leg	20. /ch/ - chip	8. /igh/ - night	20. /ə/ - corner (the 'schwa' - an unstressed vowel sound which is close to /u/)
9. /m/ - map	21. /th/ - thin	9. /oa/ - boat	
10. /n/ - net	22. /th/ - then	10. /oo/ - boot	
11. /p/ - pen	23. /ng/ - ring	11. /oo/ - look	
12. /r/ - rat	24. /zh/[3] - vision	12. /ow/ - cow	

- **Morphemes and lexemes:**

  A morpheme is the smallest unit of language that has a meaning. It is formed by a combination of phonemes. Not all morphemes are words, but all prefixes and suffixes are morphemes.

unbreakable
un + break + able

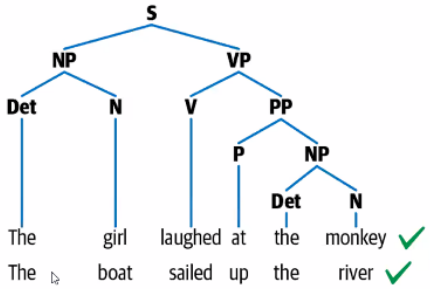
cats
cat + s

tumbling
tumble + ing

unreliability
un + rely + able + ity

- **Syntax:**

  Syntax is a set of rules to construct grammatically correct sentences out of words and phrases in a language. Syntactic structure in linguistics is represented in many different ways. A common approach to representing sentences is a parse tree. In this representation, N stands for noun, V for verb, and P for preposition. Noun phrase is denoted by NP and verb phrase by VP.

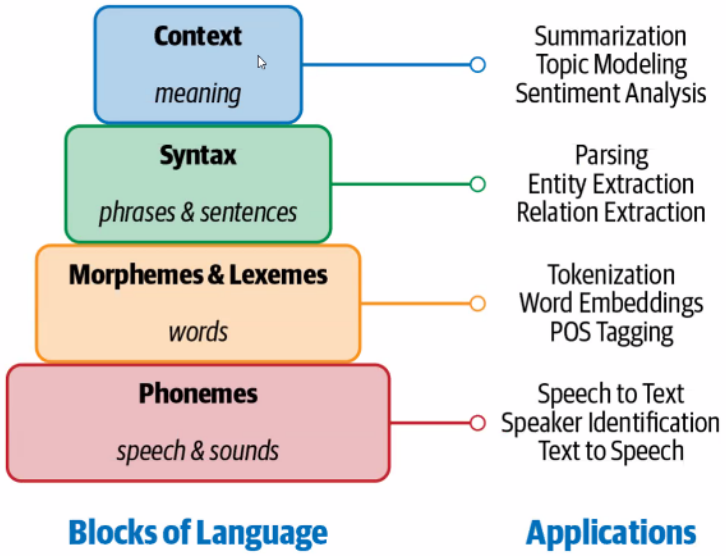


- **Context:**

  Context is how various parts in a language come together to convey a particular meaning. Context includes long-term references, world knowledge, and common sense along with the literal meaning of words and phrases.

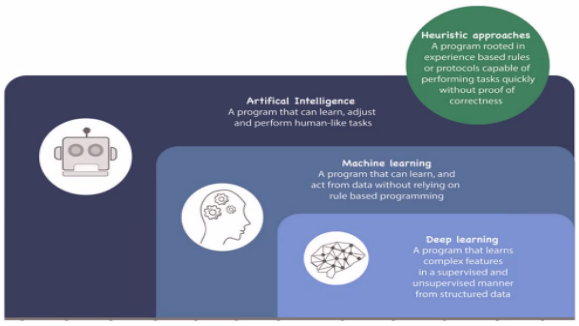
  Complex NLP tasks such as sarcasm detection, summarization, and topic modeling are some of tasks that use context heavily.

The building block of language and its applications:



Approaches to NLP:

The different approaches used to solve NLP problems commonly fall into three categories: **heuristics, machine learning, and deep learning.**



- ***Heuristics means a rule-based approach***

- **Heuristics-Bases NLP:**

  Similar to other early AI systems, early attempts at designing NLP systems were based on building rules for the task at hand.

  Examples:

  - Regular Expression
  - Wordnet
  - Open Mind Common Sense

- **Machine Learning for NLP:**

  Machine learning techniques are applied to textual data just as they're used on other forms of data, such as images, speech, and structured data. Supervised machine learning techniques such as classification and regression methods are heavily used for various NLP tasks.

  - **Naive Bayes**
  - **Support vector machine**
  - **Hidden Markov Model**

- **Deep Learning for NLP:**

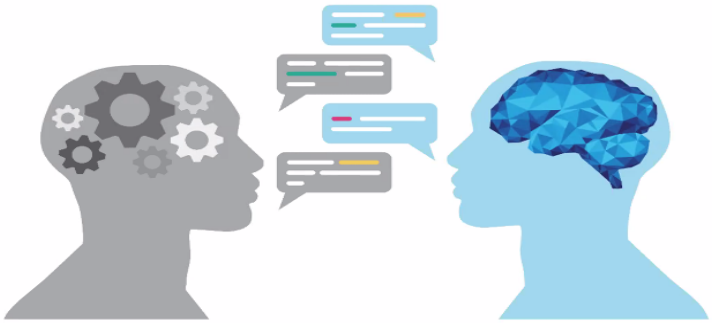
  Huge surge in using neural networks to deal with complex, unstructured data. Language is inherently complex and unstructured. herefore, we need models with better representation and learning capability to understand and solve language tasks. Here are a few popular deep neural network architectures that have become the status quo in NLP.

  - Recurrent neural networks (RNN)
  - Long short-term memory (LSTM)
  - Convolutional neural networks (CNN)
  - Transformers
  - Autoencoders

Why NLP is challenging?

Why Is NLP Challenging?

The **ambiguity and creativity** of human language are just two of the characteristics that make NLP a demanding area to work in.



- **Ambiguity:**

Ambiguity

Ambiguity means uncertainty of meaning!

The man couldn't lift his son because he was so **weak**. — Who was weak?

The man couldn't lift his son because he was so **heavy**. — Who was heavy?

Mary and Sue are **sisters**.
Mary and Sue are **mothers**. — How are Mary and Sue related?

Joan made sure to thank Susan for all the help she had **received**. — Who had received help?

Joan made sure to thank Susan for all the help she had **given**. — Who had given help?

John **promised** Bill to leave, so an hour later he left.
John **ordered** Bill to leave, so an hour later he left. — Who left an hour later?

- **Creativity:**

Language is not just rule driven; there is also a creative aspect to it. Various **styles, dialects, genres, and variations** are used in any language. **Poems** are a great example of creativity in language. Making machines understand creativity is a hard problem not just in NLP, but in AI in general.

& Diversity: There are many languages

- **Common Knowledge:**

  A key aspect of any human language is “common knowledge.” It is the set of all facts that most humans are aware of.

  Example:

  consider two sentences: **“man bit dog” and “dog bit man.”**

What is NLP Pipeline?

Break the problem down into several sub-problems, then try to develop a step-by-step procedure to solve them. Since language processing is involved, we would also list all the forms of text processing needed at each step. This step-by-step processing of text is known as a pipeline.

- **Data acquisition**
- **Text Preparation**
  - Text Cleanup
  - Basic Preprocessing
  - Advance Preprocessing
- **Feature engineering**
- **Modeling**
- **Evaluation**
- **Deployment**
- **Monitoring and model updating**

- **Point to Remember:**

- **Data Acquisition:**
  - **Available data:** It might be in CSV, in the database etc
  - **Other data:** Using API, websites (web scraping)
  - **No data:** create your own data. Do a survey to get the data
  - **Add more data to Fewer data:**
    - Do data augmentation to get more data. Replace with synonyms
    - Do bigram + lip: Eg. I am subhash or Am I Subhash or Subhash I am

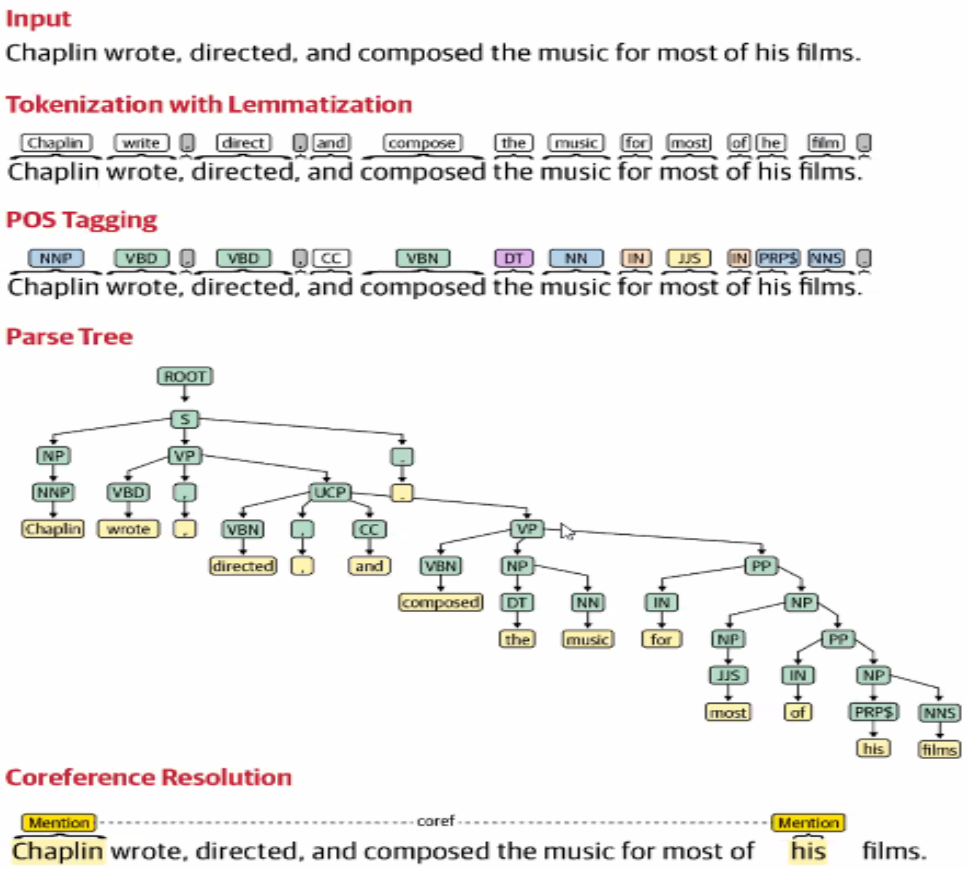
        - Back translate
        - Add noise

- **Text preparation:**
    - **Cleanup:**
        - **Remove HTML tags:** Remove HTML tags from text data. Example: Hello, world! becomes "Hello, world!"
        - **Remove emoji:** Removing emoji symbols from text data. Example: "I love pizza! 🍕" becomes "I love pizza!"
        - **Spelling Correction:** Correcting spelling mistakes in text data. Example: "Helo, wrld!" becomes "Hello, world!"

    - **Basic Preprocessing:**
        - **Tokenization:** Breaking text into individual tokens (words, punctuation marks, etc.) for further analysis. Example: "I love pizza!" becomes ["I", "love", "pizza", "!"]
            - **Word Tokenization:** Breaking text into individual words.
                - Example: "I love pizza!" becomes ["I", "love", "pizza", "!"]
            - **Sentence Tokenization:** Breaking text into individual sentences.
                - Example: "I love pizza! It's my favorite food." becomes ["I love pizza!", "It's my favorite food."]
            - **Character Tokenization:** Breaking text into individual characters.
                - Example: "Hello" becomes ["H", "e", "l", "l", "o"]
            - **Subword Tokenization:** Breaking text into subword units, such as morphemes or syllables.
                - Example: "Unhappiness" becomes ["Un", "happiness"]
        - **Sentence Type:** Identifying the type of sentence (declarative, interrogative, imperative, or exclamatory). Example: "Are you coming?" is an interrogative sentence.
        - **Word Type:** Determining the part of speech of each word (noun, verb, adjective, etc.). Example: "The cat is sleeping" - "cat" is a noun, "is" is a verb, and "sleeping" is a verb.

    - **Optional Preprocessing:**
        - **Remove Stopwords:** Removing common words (e.g., "the", "is", "are") that do not add much meaning to the text. Example: "I am going to the park" becomes "going park".
        - **Stemming**: Reducing words to their base or root form. Example: "running" becomes "run", "jumps" becomes "jump".
        - Remove Punctuation: Remove punctuation marks from the text. Example: "Hello, world!" becomes "Hello world"
        - **Lowercasing:** Converting all words in the text to lowercase. Example: "Hello World" becomes "hello world".

    - **Language Detection:** Identifying the language of the text. Example: "Bonjour, comment ça va?" is detected as French.
    - **Lemmatization:** Reducing words to their base form (lemma) using vocabulary and morphological analysis. Example: "running" becomes "run", "better" becomes "good".
- **Advanced Preprocessing:**
    - **Named Entity Recognition (NER):** Identifying and classifying named entities like names, locations, organizations, etc., in the text. Example: "Apple Inc. is headquartered in Cupertino" - "Apple Inc." is recognized as an organization and "Cupertino" as a location.
    - **POS Tagging:** Assigning part-of-speech tags to each word in the text. Example: "The cat is sleeping" - "cat" is tagged as a noun, "is" as a verb, and "sleeping" as an adjective.
    - **Dependency Parsing:** Analyzing the grammatical structure of the sentence by identifying relationships between words. Example: "The cat is sleeping" - "cat" is the subject, "is" is the verb, and "sleeping" is the predicate.
    - **Named Entity Disambiguation:** Resolving the ambiguity of named entities based on context. Example: "I saw a bat" - disambiguating between a flying mammal and sports equipment based on the context.



- **Feature Engineering:**
  - **Text Vectorization:**
    - **Bag of Words:** Representing text as a collection of word frequencies. Example: "I love pizza and pasta" becomes [1, 1, 1, 1] for ["I", "love", "pizza", "and", "pasta"].
    - **Tf-Idf (Term Frequency-Inverse Document Frequency):** Adjusting word frequencies based on their importance in the corpus. Example: "I love pizza" has higher weight than "I love cats" since "pizza" is less common than "cats".
  - **Encoding:** Converting categorical or textual data into numerical form for machine learning algorithms. Example: Converting labels like "cat," "dog," and "bird" into 0, 1, and 2, respectively.
  - **Modelling:**
    - Cloud API: AWS, GCP, AZURE: Cloud-based services provided by major providers (Amazon Web Services, Google Cloud Platform, Microsoft Azure) for hosting, deploying, and scaling machine learning models or NLP applications in the cloud.
  - **Evaluation:**
    - **Intrinsic Evaluation:**
      - False Positive (FP): Incorrectly predicted positive instances.
      - False Negative (FN): Incorrectly predicted negative instances.
      - Accuracy Score: Measure of the model's overall accuracy in classification tasks, calculated as the ratio of correct predictions to total predictions.
    - **Extrinsic Evaluation:**
      - After Deployment, Feedback from Users: Collecting feedback from real users after deploying an NLP system or application to assess its performance and user satisfaction. This evaluation helps improve the system based on practical usage.
  - **Common Terms in NLP:**
    - **Corpus:** Refers to a collection of text documents used for analysis or model training. Example: A collection of news articles, a set of customer reviews, or a compilation of scientific papers.
    - **Vocabulary:** The set of all unique words or terms present in a corpus. Example: In a news corpus, the vocabulary might include words like "politics," "economy," "sports," etc.
    - **Documents:** Refers to individual units of text in a corpus, often represented as rows in a dataset. Example: Each news article or customer review can be considered a document.
    - **Word:** A single unit of language with its own meaning, typically separated by spaces or punctuation marks. Example: In the sentence "I love pizza," "love" and "pizza" are individual words.