



SUMMER INTERNSHIP AI/ML DS-2022

NATURAL LANGUAGE PROCESSING


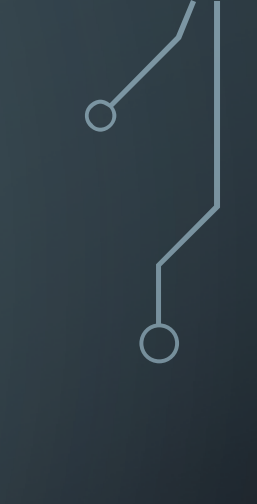

Dr. Uday Pratap Singh

Associate Professor (CE)

PIET



CONTENTS:

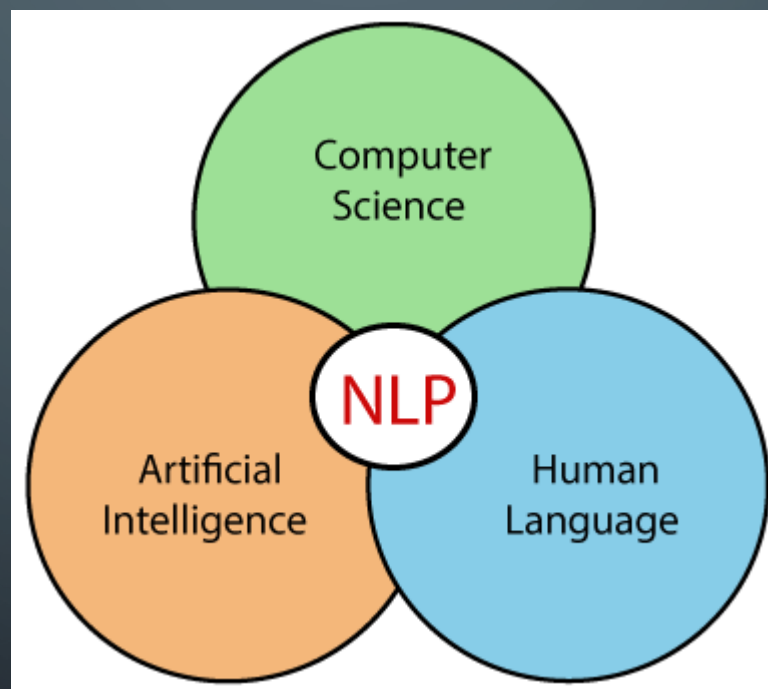
- NLP
 - NLTK
 - NLP Pre-processing
- 
- 
- 

WHAT IS NATURAL LANGUAGE PROCESSING (NLP)?

- Natural Language Processing is an **interdisciplinary field of Artificial Intelligence**.
- It is a technique used to teach a computer to understand Human languages and also interpret just like us.
- It is the art of extracting information, hidden insights from unstructured text.
- It is a sophisticated field that makes computers process text data on a large scale.
- The ultimate goal of NLP is to make computers and computer-controlled bots understand and interpret Human Languages, just as we do.

WHAT IS NATURAL LANGUAGE PROCESSING (NLP)?

- Natural Language Processing is an **interdisciplinary field of Artificial Intelligence**.
- It is a technique used to teach a computer to understand Human languages and also interpret just like us.
- It is the art of extracting information, hidden insights from unstructured text.
- It is a sophisticated field that makes computers process text data on a large scale.
- The ultimate goal of NLP is to make computers and computer-controlled bots understand and interpret Human Languages, just as we do.



COMPONENTS OF NLP

- **Natural Language Understanding**
- **Natural Language Generation**

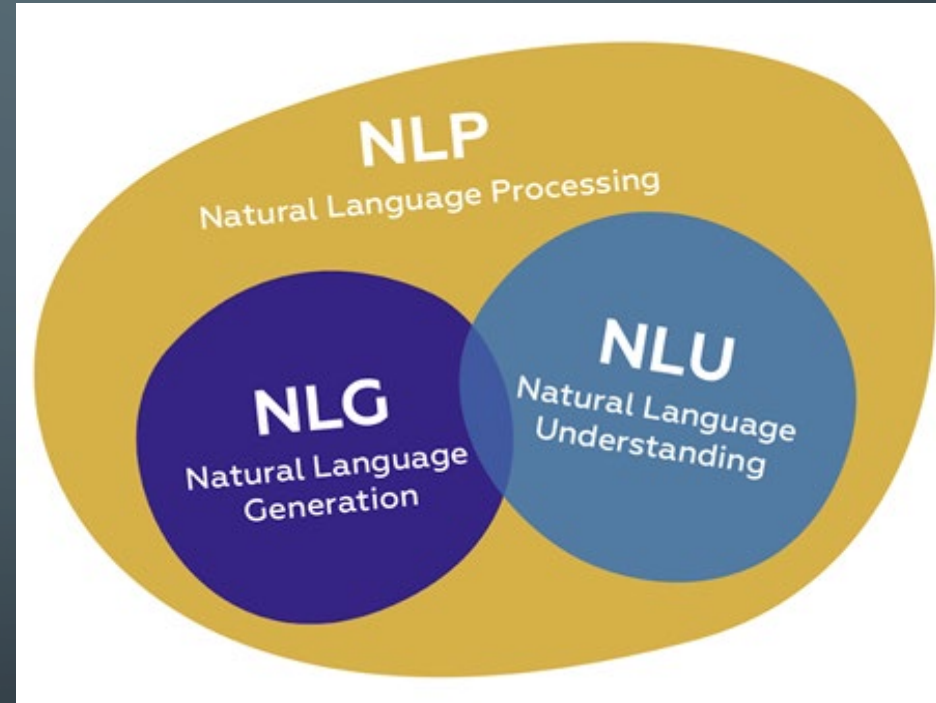


Figure: Components of NLP

NATURAL LANGUAGE UNDERSTANDING:-

- NLU helps the machine to understand and analyze human language by extracting the text from large data such as keywords, emotions, relations, and semantics, etc.

Let's see what challenges are faced by a machine-

He is looking for a match.

- What do you understand by the 'match' keyword?
- This is **Lexical Ambiguity**. It happens when a word has different meanings. Lexical ambiguity can be resolved by using parts-of-speech (POS) tagging techniques.

The Fish is ready to eat.

- What do you understand by the above example?
- This is **Syntactical Ambiguity** which means when we see more meanings in a sequence of words and also Called Grammatical Ambiguity.

NATURAL LANGUAGE GENERATION:-

- It is the process of extracting meaningful insights as phrases and sentences in the form of natural language.
- It consists –
 - **Text planning** – It includes retrieving the relevant data from the domain.
 - **Sentence planning** – It is nothing but a selection of important words, meaningful phrases, or sentences.

APPLICATIONS OF NLP

- Sentimental Analysis
- Chatbots
- Virtual Assistants
- Speech Recognition
- Machine Translation
- Advertise Matching
- Information Extraction
- Grammatical error detection
- Fake news detection
- Text Summarize

LIBRARIES FOR NLP

Here are some of the libraries for leveraging the power of Natural Language Processing.

- Natural Language Toolkit (NLTK)
- spaCY
- Gensim
- Stanford CoreNLP
- TextBlob

WHAT IS NATURAL TOOLKIT?

- **NLTK**, or [Natural Language Toolkit](#), is a Python package that you can use for NLP.
- NLTK is a leading platform for building Python programs to work with human language data.

Installing NLTK

```
pip install nltk
```

DATA PREPROCESSING USING NLTK

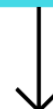
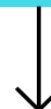
- The process of cleaning unstructured text data, so that it can be used to predict, analyze, and extract information. Real-world text data is unstructured, inconsistent. So, Data preprocessing becomes a necessary step.
- **The various Data Preprocessing methods are:**
 - Tokenization
 - Frequency Distribution of Words
 - Filtering Stop Words
 - Stemming
 - Lemmatization
 - Parts of Speech(POS) Tagging
 - Name Entity Recognition
 - WordNet
- These are some of the methods to process the text data in NLP. The list is not so exhaustive but serves as a great starting point for anyone who wants to get started with NLP.

TOKENIZING

- The process of breaking down the text data into individual tokens(words, sentences, characters) is known as **Tokenization**. It is a foremost step in **Text Analytics**.
- It's your first step in turning unstructured data into structured data, which is easier to analyze.
- Tokenizing can be done by two ways
 - **Tokenizing by word**
 - **Tokenizing by sentence**

Tokenization

Natural Language Processing



['Natural', 'Language', 'Processing']

- Importing the tokenizer from NLTK

```
from nltk.tokenize import sent_tokenize, word_tokenize
```

STOPWORDS

- Stop words are used to filter some words which are repetitive and don't hold any information. For example, words like – {**that these, below, is, are, etc.**} don't provide any information, so they need to be removed from the text. Stop Words are considered as **Noise**. NLTK provides a huge list of stop words
- Very common words like 'in', 'is', and 'an' are often used as stop words since they don't add a lot of meaning to a text in and of themselves.
- **Content words** give information about the topics covered in the text or the sentiment that the author has about those topics.
- **Context words** give information about writing style. You can observe patterns in how authors use context words in order to quantify their writing style. Once you've quantified their writing style, you can analyze a text written by an unknown author to see how closely it follows a particular writing style so you can try to identify who the author is.

STEMMING

- **Stemming** is a text processing task in which you reduce words to their root, which is the core part of a word. For example, the words “helping” and “helper” share the root “help.”
- Stemming allows you to zero in on the basic meaning of a word rather than all the details of how it’s being used.
- NLTK has
 - Porter Stemmer
 - Snowball Stemmer

- Understemming and overstemming are two ways stemming can go wrong:
- **Understemming** happens when two related words should be reduced to the same stem but aren't. This is a false negative.
- **Overstemming** happens when two unrelated words are reduced to the same stem even though they shouldn't be. This is a false positive.

POS TAGGING

SUMMARY THAT YOU CAN USE TO GET STARTED WITH NLTK'S POS TAGS:

Tags that start with	Deals With
JJ	Adjectives
NN	Nouns
RB	Adverbs
PRP	Pronouns
VB	Verbs

LEMMATIZING

- Like stemming, lemmatization is also used to reduce the word to their root word. Lemmatizing gives the complete meaning of the word which makes sense. It uses vocabulary and morphological analysis to transform a word into a root word.
- For example:
- “engineers” is lemmatized to “engineer”

STEMMING V/S LEMMATIZATION

Stemming	Lemmatization
Stemming is a process that stems or removes last few characters from a word, often leading to incorrect meanings and spelling.	Lemmatization considers the context and converts the word to its meaningful base form, which is called Lemma.
For instance , stemming the word 'Caring' would return 'Car'.	For instance, lemmatizing the word 'Caring' would return 'Care'.
Stemming is used in case of large dataset where performance is an issue	Lemmatization is computationally expensive since it involves look-up tables and what not.

CHUNKING

- While tokenizing allows you to identify words and sentences, **chunking** allows you to identify **phrases**.
- **Note:** A **phrase** is a word or group of words that works as a single unit to perform a grammatical function. **Noun phrases** are built around a noun.
- Here are some examples:
 - “A planet”
 - “A tilting planet”
 - “A swiftly tilting planet”

CHINKING

- Chinking is used together with chunking, but while chunking is used to include a pattern, **chinking** is used to exclude a pattern.

USING NAMED ENTITY RECOGNITION (NER)

- **Named entities** are noun phrases that refer to specific locations, people, organizations, and so on. With **named entity recognition**, you can find the named entities in your texts and also determine what kind of named entity they are.

FURTHER READING

- Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc.

THANK
you