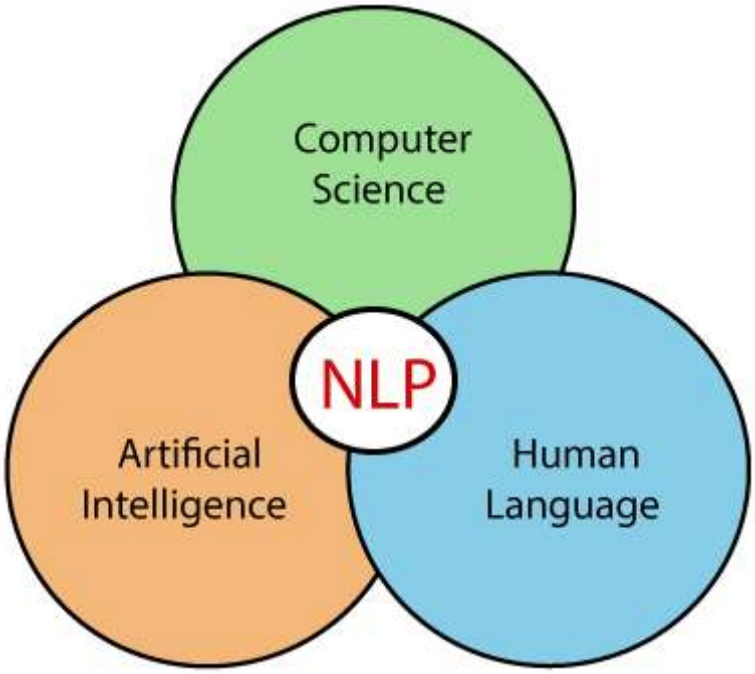
NLP Tutorial - Javatpoint

NLP tutorial provides basic and advanced concepts of the NLP tutorial. Our NLP tutorial is designed for beginners and professionals.

What is NLP?

NLP stands for **Natural Language Processing**, which is a part of **Computer Science, Human language,** and **Artificial Intelligence**. It is the technology that is used by machines to understand, analyse, manipulate, and interpret human's languages. It helps developers to organize knowledge for performing tasks such as **translation, automatic summarization, Named Entity Recognition (NER), speech recognition, relationship extraction,** and **topic segmentation**.



History of NLP

(1940-1960) - Focused on Machine Translation (MT)

The Natural Languages Processing started in the year 1940s.

1948 - In the Year 1948, the first recognisable NLP application was introduced in Birkbeck College, London.

1950s - In the Year 1950s, there was a conflicting view between linguistics and computer science. Now, Chomsky developed his first book syntactic structures and claimed that language is generative in nature.

In 1957, Chomsky also introduced the idea of Generative Grammar, which is rule based descriptions of syntactic structures.

(1960-1980) - Flavored with Artificial Intelligence (AI)

In the year 1960 to 1980, the key developments were:

Augmented Transition Networks (ATN)

Augmented Transition Networks is a finite state machine that is capable of recognizing regular languages.

Case Grammar

Case Grammar was developed by **Linguist Charles J. Fillmore** in the year 1968. Case Grammar uses languages such as English to express the relationship between nouns and verbs by using the preposition.

In Case Grammar, case roles can be defined to link certain kinds of verbs and objects.

For example: "Neha broke the mirror with the hammer". In this example case grammar identify Neha as an agent, mirror as a theme, and hammer as an instrument.

In the year 1960 to 1980, key systems were:

SHRDLU

SHRDLU is a program written by **Terry Winograd** in 1968-70. It helps users to communicate with the computer and moving objects. It can handle instructions such as "pick up the green boll" and also answer the questions like "What is inside the black box." The main importance of SHRDLU is that it shows those syntax, semantics, and reasoning about the world that can be combined to produce a system that understands a natural language.

LUNAR

LUNAR is the classic example of a Natural Language database interface system that is used ATNs and Woods' Procedural Semantics. It was capable of translating elaborate natural language expressions into database queries and handle 78% of requests without errors.

1980 - Current

Till the year 1980, natural language processing systems were based on complex sets of hand-written rules. After 1980, NLP introduced machine learning algorithms for language processing.

In the beginning of the year 1990s, NLP started growing faster and achieved good process accuracy, especially in English Grammar. In 1990 also, an electronic text introduced, which provided a good resource for training and examining natural language programs. Other factors may include the availability of computers with fast CPUs and more memory. The major factor behind the advancement of natural language processing was the Internet.

Now, modern NLP consists of various applications, like **speech recognition, machine translation,** and **machine text reading**. When we combine all these applications then it allows the artificial intelligence to gain knowledge of the world. Let's consider the example of AMAZON ALEXA, using this robot you can ask the question to Alexa, and it will reply to you.

Advantages of NLP

- NLP helps users to ask questions about any subject and get a direct response within seconds.
- NLP offers exact answers to the question means it does not offer unnecessary and unwanted information.
- NLP helps computers to communicate with humans in their languages.
- It is very time efficient.
- Most of the companies use NLP to improve the efficiency of documentation processes, accuracy of documentation, and identify the information from large databases.

Disadvantages of NLP

A list of disadvantages of NLP is given below:

- NLP may not show context.
- NLP is unpredictable
- NLP may require more keystrokes.
- NLP is unable to adapt to the new domain, and it has a limited function that's why NLP is built for a single and specific task only.

Components of NLP

There are the following two components of NLP -

1. Natural Language Understanding (NLU)

Natural Language Understanding (NLU) helps the machine to understand and analyse human language by extracting the metadata from content such as concepts, entities, keywords, emotion, relations, and semantic roles.

NLU mainly used in Business applications to understand the customer's problem in both spoken and written language.

NLU involves the following tasks -

- It is used to map the given input into useful representation.
- It is used to analyze different aspects of the language.

2. Natural Language Generation (NLG)

Natural Language Generation (NLG) acts as a translator that converts the computerized data into natural language representation. It mainly involves Text planning, Sentence planning, and Text Realization.

Note: The NLU is difficult than NLG.

Difference between NLU and NLG

- NLU: NLU is the process of reading and interpreting language.
  - NLG: NLG is the process of writing or generating language.
- NLU: It produces non-linguistic outputs from natural language inputs.
  - NLG: It produces constructing natural language outputs from non-linguistic inputs.

Applications of NLP

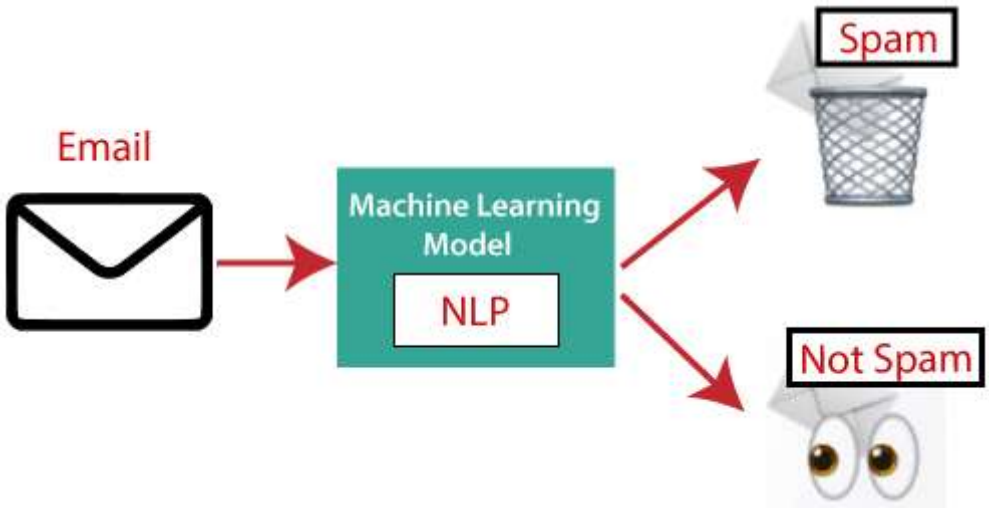
There are the following applications of NLP -

1. Question Answering

Question Answering focuses on building systems that automatically answer the questions asked by humans in a natural language.

2. Spam Detection

Spam detection is used to detect unwanted e-mails getting to a user's inbox.

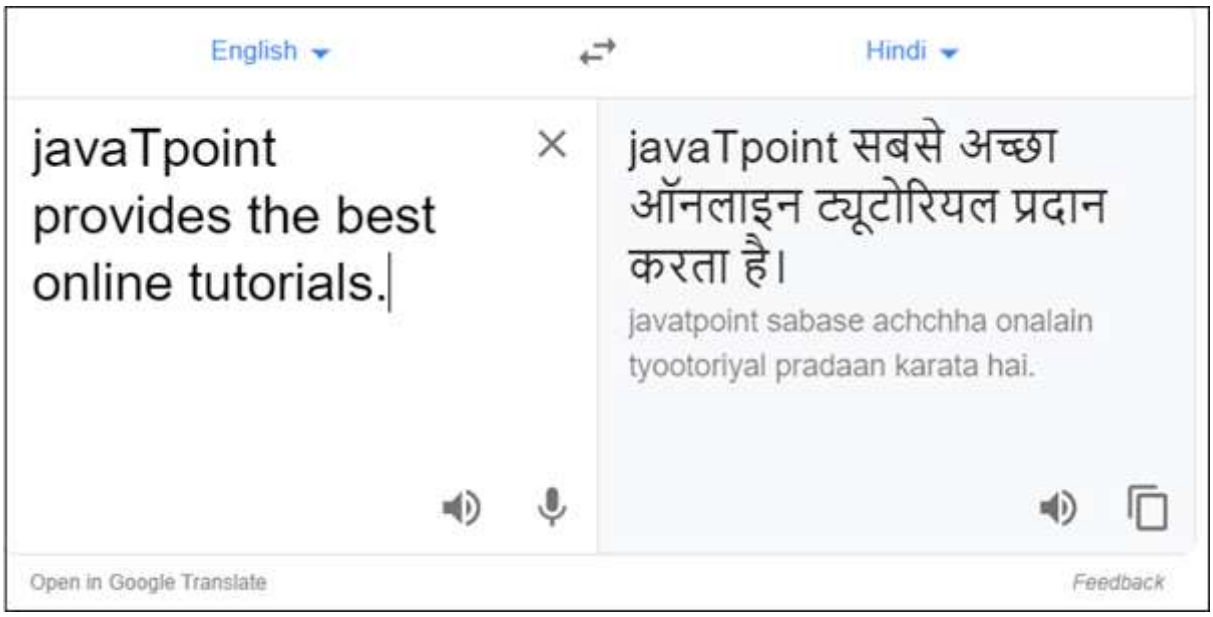


3. Sentiment Analysis

Sentiment Analysis is also known as **opinion mining**. It is used on the web to analyse the attitude, behaviour, and emotional state of the sender. This application is implemented through a combination of NLP (Natural Language Processing) and statistics by assigning the values to the text (positive, negative, or natural), identify the mood of the context (happy, sad, angry, etc.)

4. Machine Translation

Machine translation is used to translate text or speech from one natural language to another natural language.



Example: Google Translator

5. Spelling correction

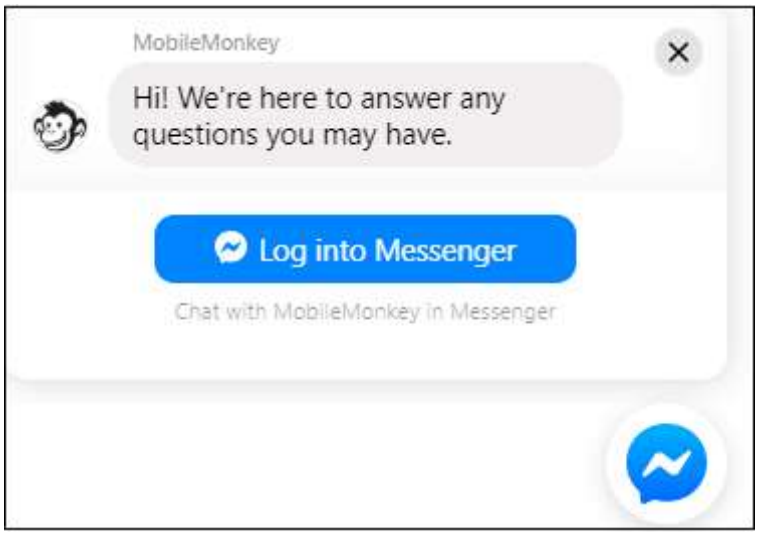
Microsoft Corporation provides word processor software like MS-word, PowerPoint for the spelling correction.

6. Speech Recognition

Speech recognition is used for converting spoken words into text. It is used in applications, such as mobile, home automation, video recovery, dictating to Microsoft Word, voice biometrics, voice user interface, and so on.

7. Chatbot

Implementing the Chatbot is one of the important applications of NLP. It is used by many companies to provide the customer's chat services.



8. Information extraction

Information extraction is one of the most important applications of NLP. It is used for extracting structured information from unstructured or semi-structured machine-readable documents.

9. Natural Language Understanding (NLU)

It converts a large set of text into more formal representations such as first-order logic structures that are easier for the computer programs to manipulate notations of the natural language processing.

How to build an NLP pipeline

There are the following steps to build an NLP pipeline -

Step1: Sentence Segmentation

Sentence Segment is the first step for building the NLP pipeline. It breaks the paragraph into separate sentences.

Example: Consider the following paragraph -

Independence Day is one of the important festivals for every Indian citizen. It is celebrated on the 15th of August each year ever since India got independence from the British rule. The day celebrates independence in the true sense.

Sentence Segment produces the following result:

1. "Independence Day is one of the important festivals for every Indian citizen."
2. "It is celebrated on the 15th of August each year ever since India got independence from the British rule."

3. "This day celebrates independence in the true sense."

Step2: Word Tokenization

Word Tokenizer is used to break the sentence into separate words or tokens.

Example:

JavaTpoint offers Corporate Training, Summer Training, Online Training, and Winter Training.

Word Tokenizer generates the following result:

"JavaTpoint", "offers", "Corporate", "Training", "Summer", "Training", "Online", "Training", "and", "Winter", "Training", "."

Step3: Stemming

Stemming is used to normalize words into its base form or root form. For example, celebrates, celebrated and celebrating, all these words are originated with a single root word "celebrate." The big problem with stemming is that sometimes it produces the root word which may not have any meaning.

For Example, intelligence, intelligent, and intelligently, all these words are originated with a single root word "intelligen." In English, the word "intelligen" do not have any meaning.

Step 4: Lemmatization

Lemmatization is quite similar to the Stamming. It is used to group different inflected forms of the word, called Lemma. The main difference between Stemming and lemmatization is that it produces the root word, which has a meaning.

For example: In lemmatization, the words intelligence, intelligent, and intelligently has a root word intelligent, which has a meaning.

Step 5: Identifying Stop Words

In English, there are a lot of words that appear very frequently like "is", "and", "the", and "a". NLP pipelines will flag these words as stop words. **Stop words** might be filtered out before doing any statistical analysis.

Example: He **is a** good boy.

Note: When you are building a rock band search engine, then you do not ignore the word "The."

Step 6: Dependency Parsing

Dependency Parsing is used to find that how all the words in the sentence are related to each other.

Step 7: POS tags

POS stands for parts of speech, which includes Noun, verb, adverb, and Adjective. It indicates that how a word functions with its meaning as well as grammatically within the sentences. A word has one or more parts of speech based on the context in which it is used.

Example: "Google" something on the Internet.

In the above example, Google is used as a verb, although it is a proper noun.

Step 8: Named Entity Recognition (NER)

Named Entity Recognition (NER) is the process of detecting the named entity such as person name, movie name, organization name, or location.

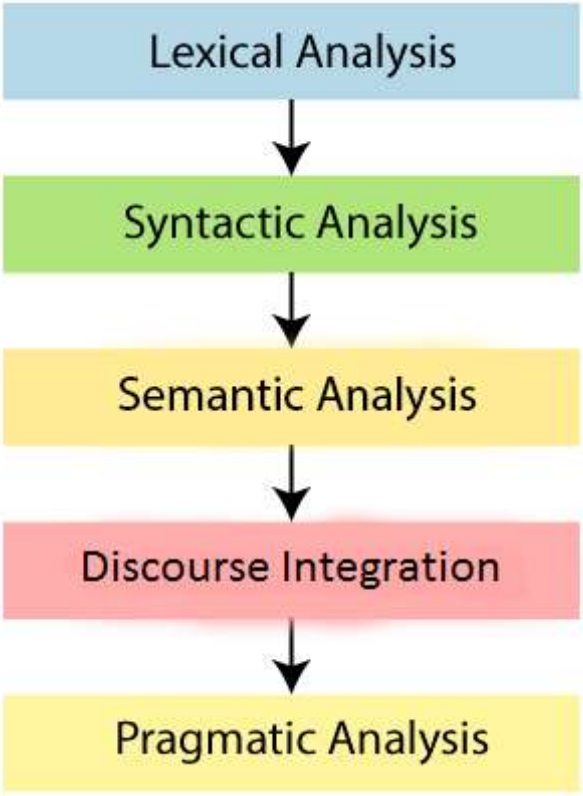
Example: Steve Jobs introduced iPhone at the Macworld Conference in San Francisco, California.

Step 9: Chunking

Chunking is used to collect the individual piece of information and grouping them into bigger pieces of sentences.

Phases of NLP

There are the following five phases of NLP:



1. Lexical Analysis and Morphological

The first phase of NLP is the Lexical Analysis. This phase scans the source code as a stream of characters and converts it into meaningful lexemes. It divides the whole text into paragraphs, sentences, and words.

2. Syntactic Analysis (Parsing)

Syntactic Analysis is used to check grammar, word arrangements, and shows the relationship among the words.

Example: Agra goes to the Poonam

In the real world, Agra goes to the Poonam, does not make any sense, so this sentence is rejected by the Syntactic analyzer.

3. Semantic Analysis

Semantic analysis is concerned with the meaning representation. It mainly focuses on the literal meaning of words, phrases, and sentences.

4. Discourse Integration

Discourse Integration depends upon the sentences that proceeds it and also invokes the meaning of the sentences that follow it.

5. Pragmatic Analysis

Pragmatic is the fifth and last phase of NLP. It helps you to discover the intended effect by applying a set of rules that characterize cooperative dialogues.

For Example: "Open the door" is interpreted as a request instead of an order.

Why NLP is difficult?

NLP is difficult because Ambiguity and Uncertainty exist in the language.

Ambiguity

There are the following three ambiguity -

- **Lexical Ambiguity**

Lexical Ambiguity exists in the presence of two or more possible meanings of the sentence within a single word.

Example:

Manya is looking for a **match**.

In the above example, the word match refers to that either Manya is looking for a partner or Manya is looking for a match. (Cricket or other match)

- **Syntactic Ambiguity**

Syntactic Ambiguity exists in the presence of two or more possible meanings within the sentence.

Example:

I saw the girl with the binocular.

In the above example, did I have the binoculars? Or did the girl have the binoculars?

- **Referential Ambiguity**

Referential Ambiguity exists when you are referring to something using the pronoun.

Example: Kiran went to Sunita. She said, "I am hungry."

In the above sentence, you do not know that who is hungry, either Kiran or Sunita.

NLP APIs

Natural Language Processing APIs allow developers to integrate human-to-machine communications and complete several useful tasks such as speech recognition, chatbots, spelling correction, sentiment analysis, etc.

A list of NLP APIs is given below:

- **IBM Watson API**
  IBM Watson API combines different sophisticated machine learning techniques to enable developers to classify text into various custom categories. It supports multiple languages, such as English, French, Spanish, German, Chinese, etc. With the help of IBM Watson API, you can extract insights from texts, add automation in workflows, enhance search, and understand the sentiment. The main advantage of this API is that it is very easy to use.
  Pricing: Firstly, it offers a free 30 days trial IBM cloud account. You can also opt for its paid plans.
- **Chatbot API**
  Chatbot API allows you to create intelligent chatbots for any service. It supports Unicode characters, classifies text, multiple languages, etc. It is very easy to use. It helps you to create a chatbot for your web applications.
  Pricing: Chatbot API is free for 150 requests per month. You can also opt for its paid version, which starts from $100 to $5,000 per month.
- **Speech to text API**
  Speech to text API is used to convert speech to text
  Pricing: Speech to text API is free for converting 60 minutes per month. Its paid version starts form $500 to $1,500 per month.
- **Sentiment Analysis API**
  Sentiment Analysis API is also called as **'opinion mining'** which is used to identify the tone of a user (positive, negative, or neutral)
  Pricing: Sentiment Analysis API is free for less than 500 requests per month. Its paid version starts form $19 to $99 per month.
- **Translation API by SYSTRAN**
  The Translation API by SYSTRAN is used to translate the text from the source language to the target language. You can use its NLP APIs for language detection, text segmentation, named entity recognition, tokenization, and many other tasks.
  Pricing: This API is available for free. But for commercial users, you need to use its paid version.
- **Text Analysis API by AYLIEN**
  Text Analysis API by AYLIEN is used to derive meaning and insights from the textual content. It is available for both free as well as paid from$119 per month. It is easy to use.
  Pricing: This API is available free for 1,000 hits per day. You can also use its paid version, which starts from $199 to S1, 399 per month.

- **Cloud NLP API**
  The Cloud NLP API is used to improve the capabilities of the application using natural language processing technology. It allows you to carry various natural language processing functions like sentiment analysis and language detection. It is easy to use.
  Pricing: Cloud NLP API is available for free.
- **Google Cloud Natural Language API**
  Google Cloud Natural Language API allows you to extract beneficial insights from unstructured text. This API allows you to perform entity recognition, sentiment analysis, content classification, and syntax analysis in more the 700 predefined categories. It also allows you to perform text analysis in multiple languages such as English, French, Chinese, and German.
  Pricing: After performing entity analysis for 5,000 to 10,000,000 units, you need to pay $1.00 per 1000 units per month.

NLP Libraries

Scikit-learn: It provides a wide range of algorithms for building machine learning models in Python.

Natural language Toolkit (NLTK): NLTK is a complete toolkit for all NLP techniques.

Pattern: It is a web mining module for NLP and machine learning.

TextBlob: It provides an easy interface to learn basic NLP tasks like sentiment analysis, noun phrase extraction, or pos-tagging.

Quepy: Quepy is used to transform natural language questions into queries in a database query language.

SpaCy: SpaCy is an open-source NLP library which is used for Data Extraction, Data Analysis, Sentiment Analysis, and Text Summarization.

Gensim: Gensim works with large datasets and processes data streams.

Difference between Natural language and Computer Language

- Natural Language: Natural language has a very large vocabulary.
  - Computer Language: Computer language has a very limited vocabulary.
- Natural Language: Natural language is easily understood by humans.
  - Computer Language: Computer language is easily understood by the machines.
- Natural Language: Natural language is ambiguous in nature.

- Computer Language: Computer language is unambiguous.

Prerequisite

Before learning NLP, you must have the basic knowledge of Python.

Audience

Our NLP tutorial is designed to help beginners.

Problem

We assure that you will not find any problem in this NLP tutorial. But if there is any mistake or error, please post the error in the contact form.