Natural Language Processing (NLP)

(CTMTAICS SII P3)

Tuesday : 10:00 a.m. – 12:00 noon
Wednesday : 10:00 a.m. – 12:00 noon
Friday : 10:00a.m. – 11:00 a.m.

Natural Language Processing (NLP)

- Unit 1: Introduction (Regular expression and Finite state Automaton
- Unit 2: Word Level Analysis
- Unit 3: Syntactic Analysis
- Unit 4: Semantic and Pragmatics
- Unit 5: Discourse Analysis and Lexical Resources

Prerequisites :

- ✓ Theory of computation
- ✓ Probability and Statistics
- ✓ Python scripts

TA1 (25)	Mid Sem (50)	End Sem (100)
Unit 1	Unit 1 , 2 and 3	Unit 1,2,3,4 and 5

- ➢ TA2 : Project
- ➢ Lab : 10 -12 Labs

Lab Marks

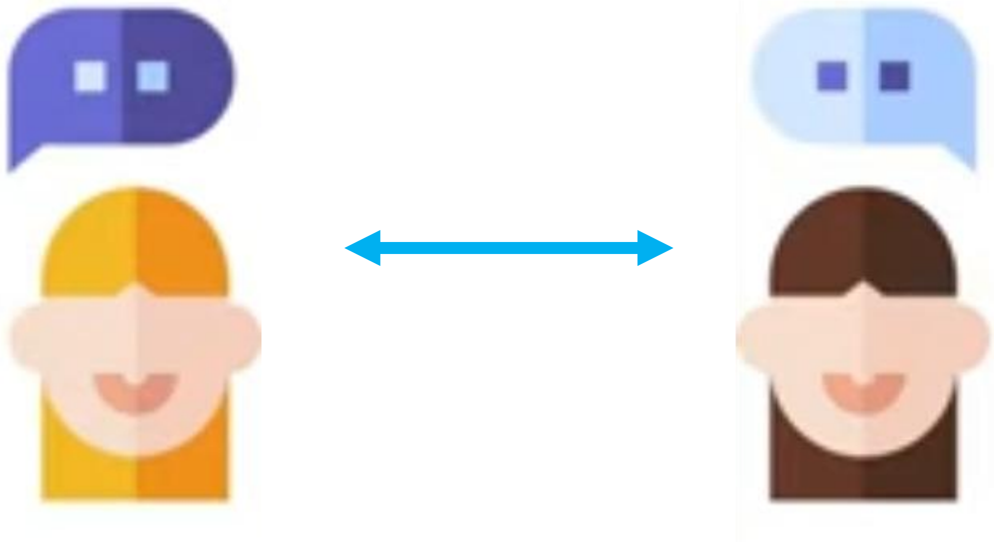
- Continuous Evaluation (100 marks)
- External Examination (100 marks)

Pattern

- ✓ Aim of Assignment
- ✓ Tools used
- ✓ Theory
- ✓ Procedure
- ✓ Result and Discussion
- ✓ Conclusion
- ✓ Code

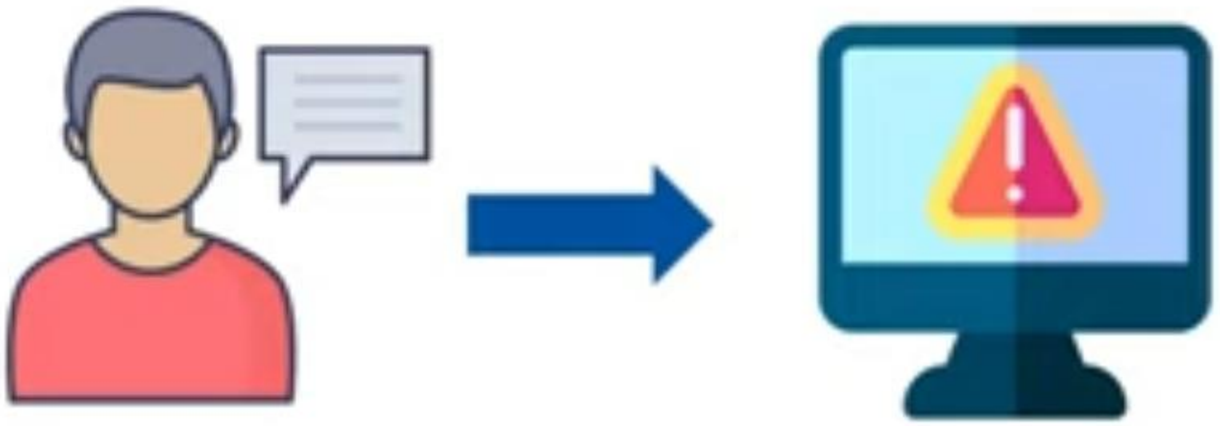
What is Natural Language ?

Natural language refers to the human way of communicating, i.e. through text and speech.

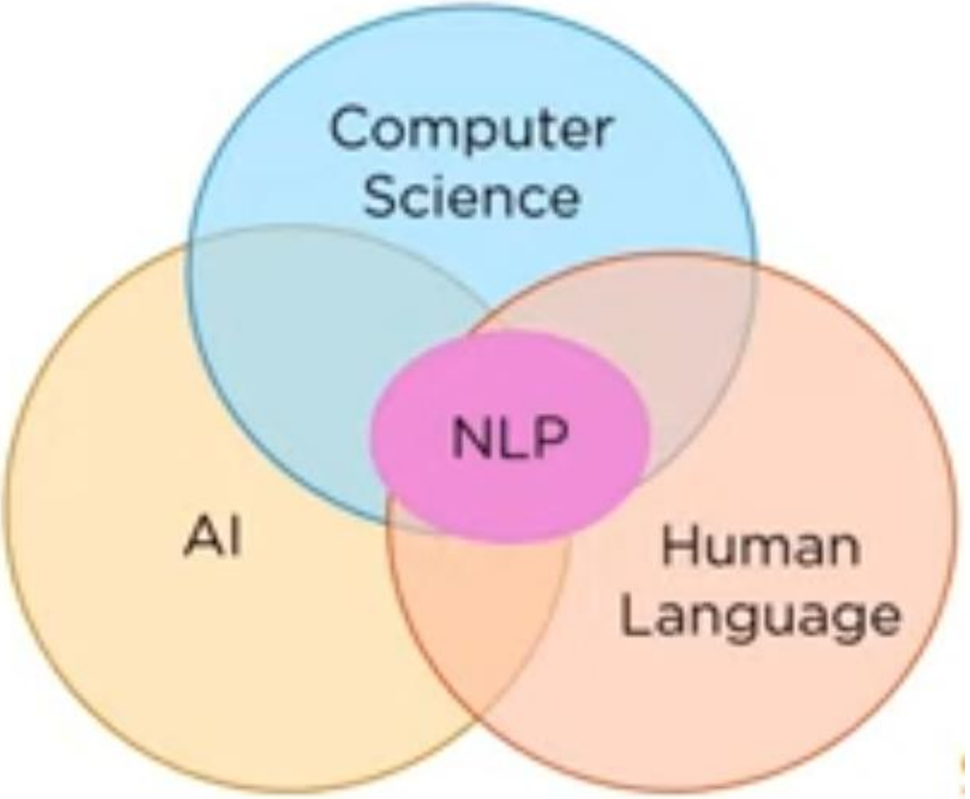


What is Natural Language Processing (NLP)

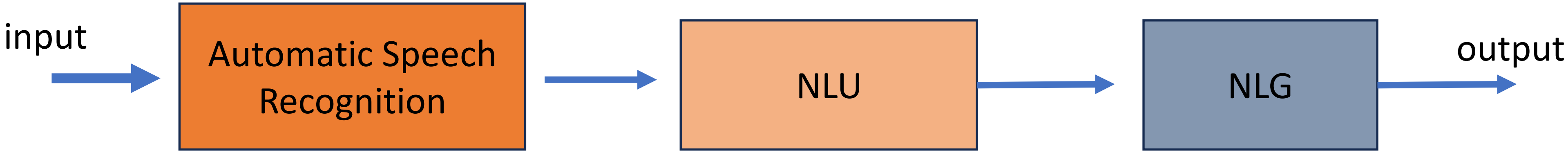
NLP refers to the branch of artificial intelligence that allows the machines to understand the natural language.



What is Natural Language Processing (NLP)



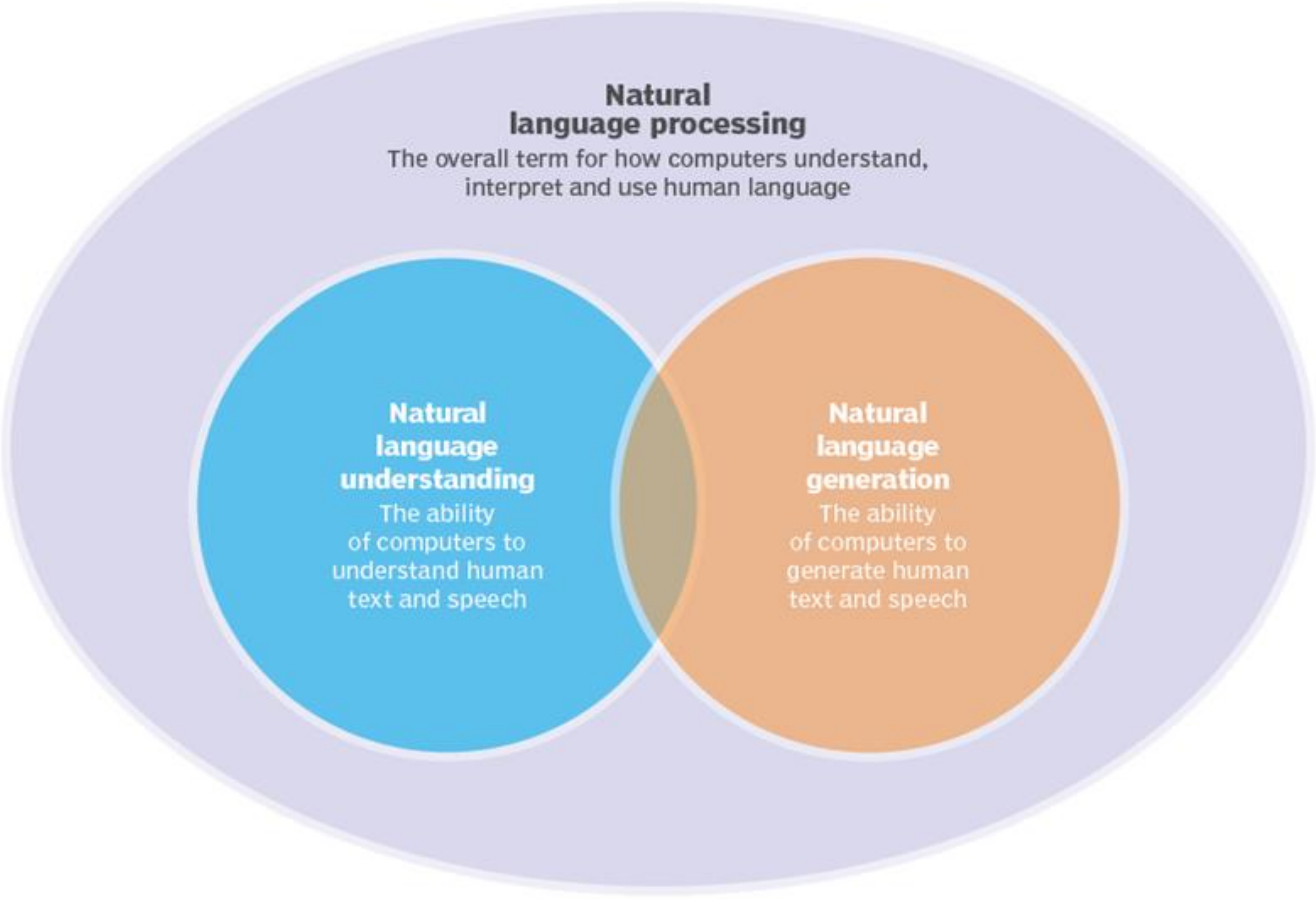
Natural Language Processing



NLU : What do the users say? Their intent and Meaning.

NLG : What should say to user?
It should be Intelligent and conversational.
Deal with structured data.
Text/Sentences Planning (Corpus).

How NLP, NLU and NLG are related



Natural language processing
The overall term for how computers understand, interpret and use human language

Natural language understanding
The ability of computers to understand human text and speech

Natural language generation
The ability of computers to generate human text and speech

What is Natural Language Processing ?

1. NLP is a subfield of linguistics, computer science and artificial intelligence that deals with the interaction between computers and human languages.
2. It enables computers to understand, interpret, and generate natural language, the way humans do.
3. It involves a variety of techniques, including computational linguistics, machine learning, and statistical modeling.
4. It is used in a wide range of industries, including finance, healthcare, education, and entertainment.
5. Some of the main applications of NLP include language translation, speech recognition, sentiment analysis, text classification, and information retrieval.
6. NLP is a rapidly evolving field that is driving new advances in computer science and artificial intelligence.
7. NLP has the potential to transform the way we interact with technology in our daily lives.

Advantages

- **Improves human-computer interaction**: Enables computers to understand and respond to human languages, which improves the overall user experience and makes it easier for people to interact with computers.

- **Automates repetitive tasks:** NLP techniques can be used to automate repetitive tasks, such as text summarization, sentiment analysis, and language translation, which can save time and increase efficiency.

- **Enables new applications:** NLP enables the development of new applications, such as virtual assistants, chatbots, and question answering systems, that can improve customer service, provide information, and more.

- **Improves decision-making:** NLP techniques can be used to extract insights from large amounts of unstructured data, such as social media posts and customer feedback, which can improve decision-making in various industries.

Advantages

- **Improves accessibility:** NLP can be used to make technology more accessible, such as by providing text-to-speech and speech-to-text capabilities for people with disabilities.

- **Facilitates multilingual communication:** NLP techniques can be used to translate and analyze text in different languages, which can facilitate communication between people who speak different languages.

- **Improves information retrieval:** NLP can be used to extract information from large amounts of data, such as search engine results, to improve information retrieval and provide more relevant results.

- **Enables sentiment analysis:** NLP techniques can be used to analyze the sentiment of text, such as social media posts and customer reviews, which can help businesses understand how customers feel about their products and services.

Advantages

- **Improves content creation:** NLP can be used to generate content, such as automated article writing, which can save time and resources for businesses and content creators.

- **Supports data analytics:** NLP can be used to extract insights from text data, which can support data analytics and improve decision-making in various industries.

- **Enhances natural language understanding:** NLP research and development can lead to improved natural language understanding, which can benefit various industries and applications.

Disadvantages

- **Limited understanding of context:** NLP systems have a limited understanding of context, which can lead to misinterpretations or errors in the output.
- **Requires large amounts of data:** NLP systems require large amounts of data to train and improve their performance, which can be expensive and time-consuming to collect.
- **Limited ability to understand idioms and sarcasm:** NLP systems have a limited ability to understand idioms, sarcasm, and other forms of figurative language, which can lead to misinterpretations or errors in the output.
- **Limited ability to understand emotions:** NLP systems have a limited ability to understand emotions and tone of voice, which can lead to misinterpretations or errors in the output.
- **Difficulty with multi-lingual processing:** NLP systems may struggle to accurately process multiple languages, especially if they are vastly different in grammar or structure.

Disadvantages

- **Dependency on language resources:** NLP systems heavily rely on language resources, such as dictionaries and corpora, which may not always be available or accurate for certain languages or domains.

- **Difficulty with rare or ambiguous words:** NLP systems may struggle to accurately process rare or ambiguous words, which can lead to errors in the output.

- **Lack of creativity:** NLP systems are limited to processing and generating output based on patterns and rules, and may lack the creativity and spontaneity of human language use

- **Ethical considerations:** NLP systems may perpetuate biases and stereotypes, and there are ethical concerns around the use of NLP in areas such as surveillance and automated decision-making.

Application Areas of NLP

- **Speech recognition and transcription:** NLP techniques are used to convert speech to text, which is useful for tasks such as dictation and voice-controlled assistants. (Ex: Google Assistant)
- **Language translation:** NLP techniques are used to translate text from one language to another, which is useful for tasks such as global communication and e-commerce. (Ex: Google Translator)
- **Text summarization:** NLP techniques are used to summarize long text documents into shorter versions, which is useful for tasks such as news summarization and document indexing.
- **Sentiment analysis**: NLP techniques are used to determine the sentiment or emotion expressed in text, which is useful for tasks such as customer feedback analysis and social media monitoring. (Ex: Twitter Information's)
- **Question answering:** NLP techniques are used to answer questions asked in natural language, which is useful for tasks such as chatbots and virtual assistants. (Ex: Chatbots)

Application Areas of NLP

- Contextual Advertisements

- Email Clients (Spam filtering / Smart Reply)

- Social Media (Removing adult content and opinion mining)

Tasks on NLP

- Text / Document Classification
- Sentiment Analysis
- Information Retrieval
- Parts of Speech tagging
- Language Detection and Machine Translation
- Conversational Agents Design
- Knowledge Graph and QA Systems
- Text Summarization
- Topic Modelling
- Spell Checking and Grammar correction
- Speech to Text conversation
- Text Parsing

Approaches in NLP

- Heuristic Approaches (Regular Expression, DFA, etc.)
- Machine Learning Approaches
- Deep Learning Approaches (ANN, RNN, CNN, etc)

Challenges of NLP

1. Ambiguity

Example :

I saw the boy on the beach with my binoculars.

I have never tasted a cake quite like that one before.

Types of Ambiguity

- Lexical Ambiguity (The tank was full of water)
- Syntactic Ambiguity (Old man and women were taken to a safe places)
- Semantic Ambiguity (The car hit the pole while it was moving)
- Pragmatic Ambiguity (The police are comming)

Challenges of NLP

2. Contextual words

Example: I ran to the store because we ran out of milk.

3. Colloquialisms and slang

- Piece of cake, Pulling your leg

4. Synonyms

5. Irony, Sarcasm and tonal difference

Example : That’s just what I needed today!.

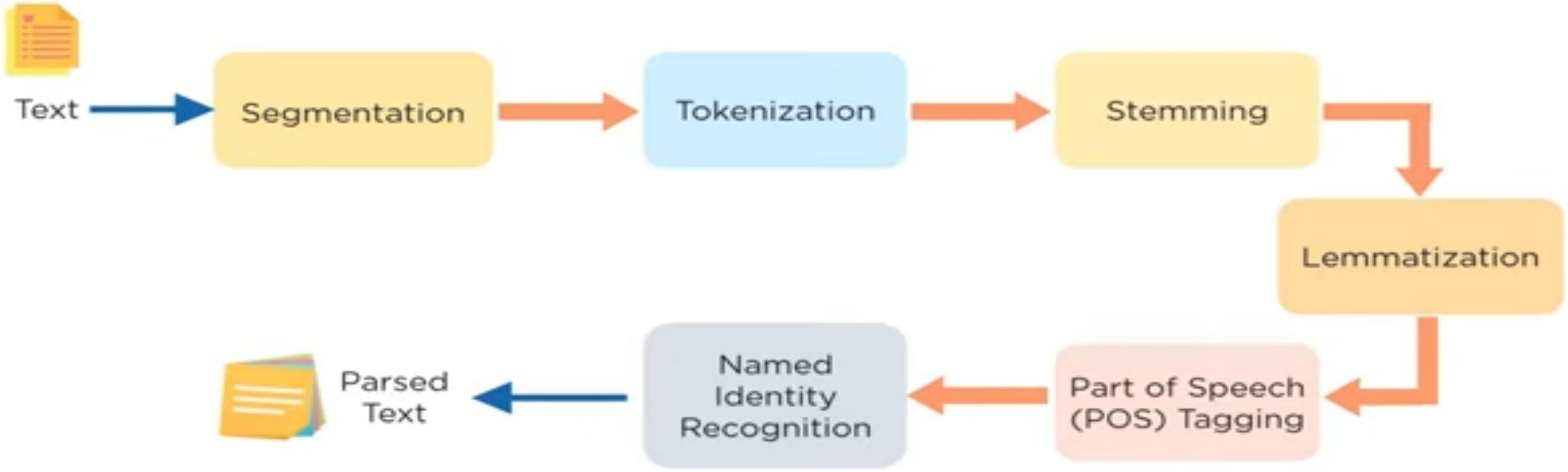
6. Spelling Errors

7.Creativity

Example : Poem, Dialogue, Scripts

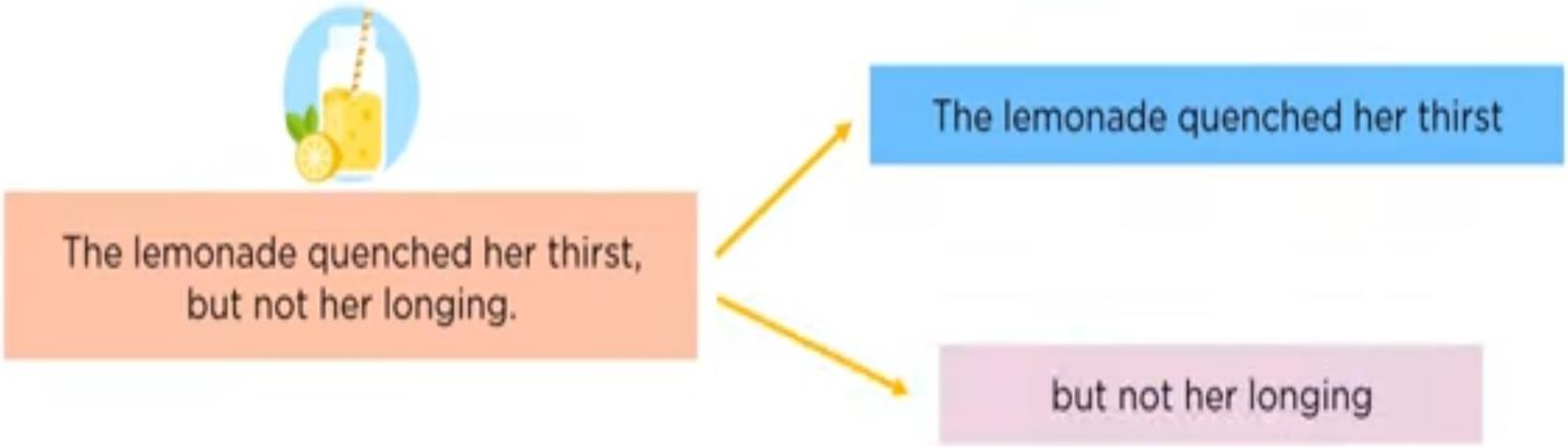
NLP Pipeline

NLP pipeline is a set of steps followed to build an end to end NLP software.



1. Segmentation

- The process of dividing a sentence into its component sentences, Usually along punctuation marks.



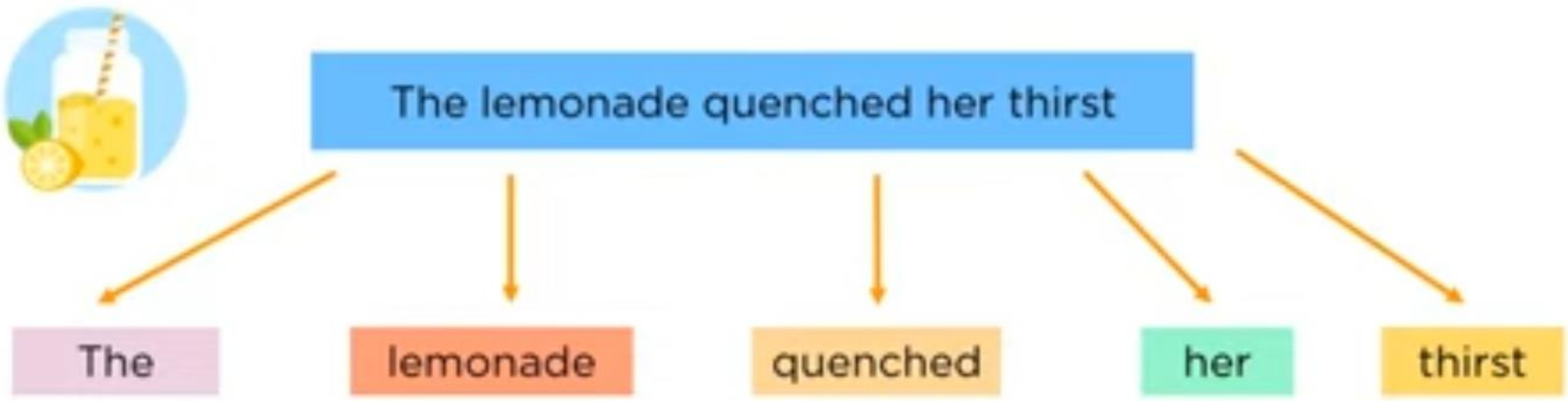
The lemonade quenched her thirst, but not her longing.

The lemonade quenched her thirst

but not her longing

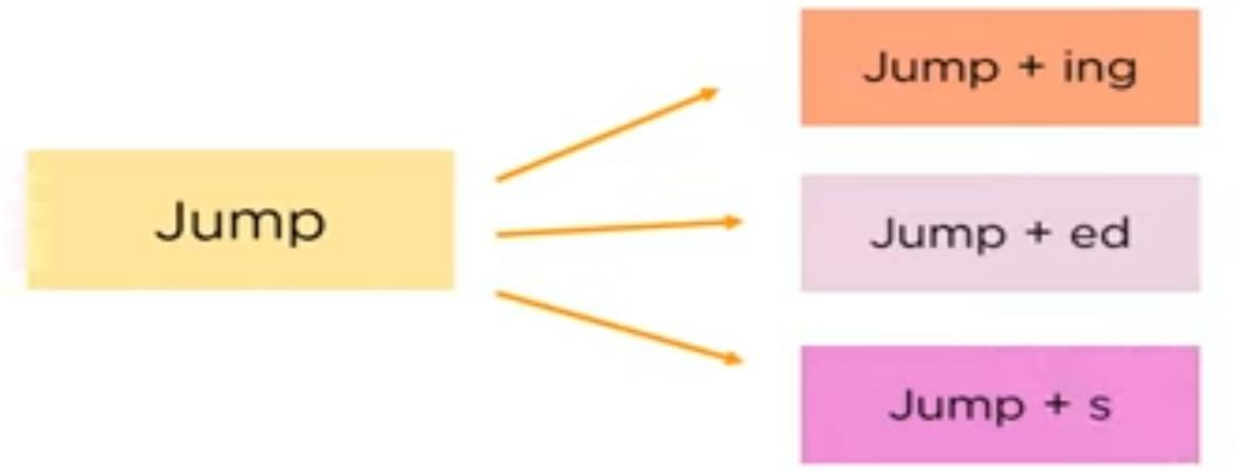
2. Tokenization

- The process of splitting sentences into their constituent words is called Tokenization



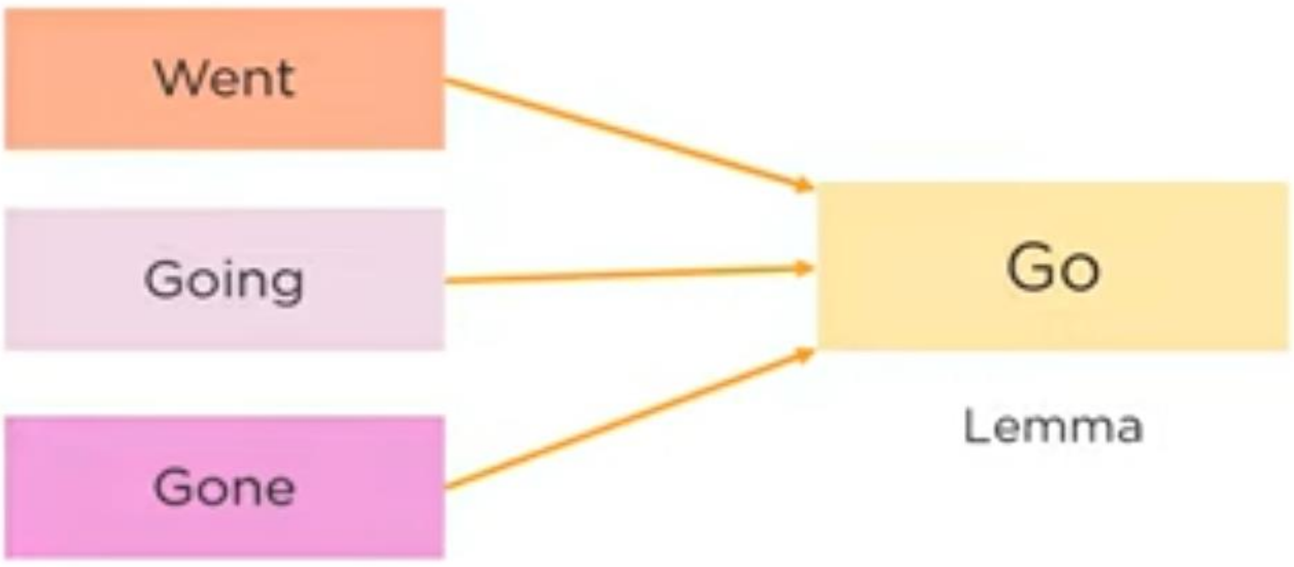
3. Stemming

- The process of obtaining the word stem of a new word.
- Word Stem give new words upon adding affixes to them.



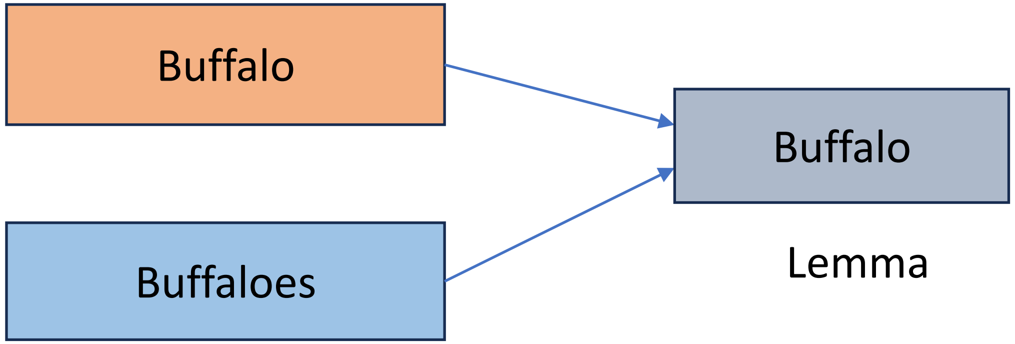
4. Lemmatization

- The process of obtaining the Root Stem of a word.
- Root Stem give new base form of a word.



Example:

There’s a Buffalo grazing in the field.
There are Buffaloes grazing in the field.


Buffalo

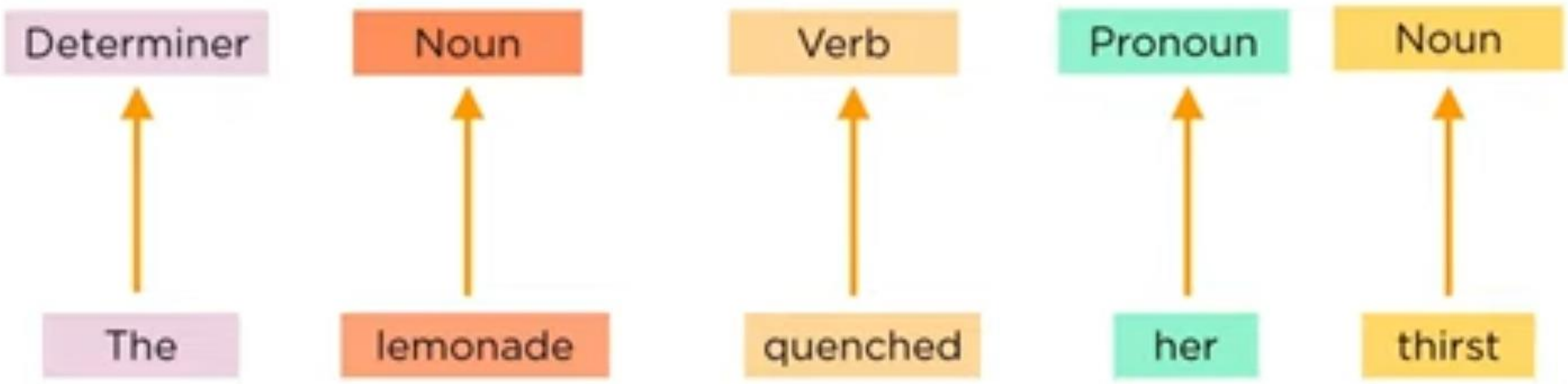
Buffaloes

Buffalo

Lemma

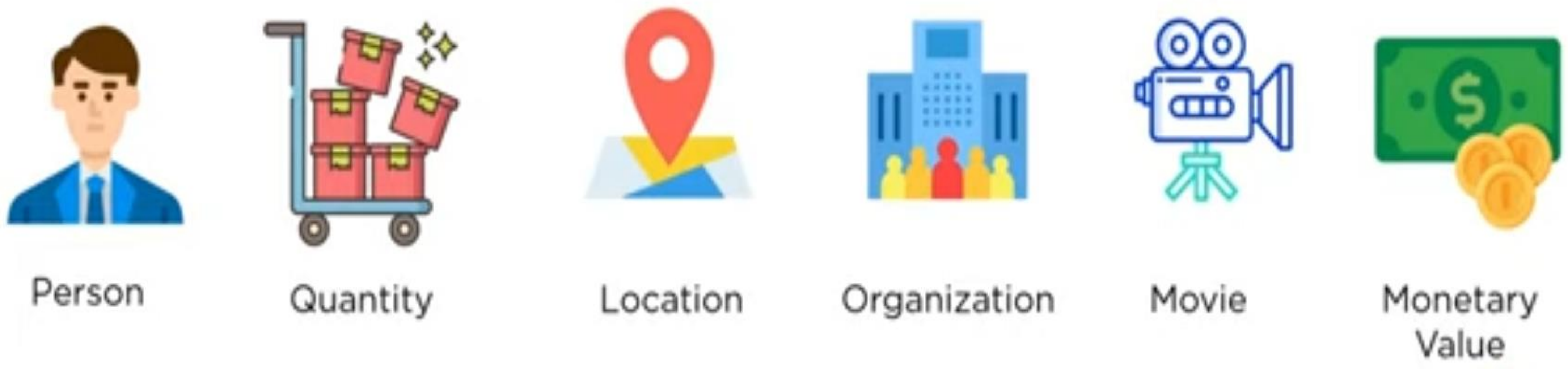
5. Part of Speech Tagging

- Identifies which part of speech a word belongs to.
- It tags a word as a verb, noun, pronoun etc.



6. Named Entity Recognition

- Classifying the words into subcatagories.
- The subcatagories are person, Quantity, Location, Organization, Movie, Monetary value etc.



Step #1: Sentence Segmentation Breaking the piece of text in various sentences.

- ***Input :*** *San Pedro is a town on the southern part of the island of Ambergris Caye in the Belize District of the nation of Belize in Central America. According to 2015 mid-year estimates the town has a population of about 16444. It is the second largest town in the Belize District and largest in the Belize Rural South constituency.*

- *Output :*

✓*1.San Pedro is a town on the southern part of the island of Ambergris Caye in the Belize District of the nation of Belize, in Central America.*

✓*2. According to 2015 mid-year estimates the town has a population of about 16444.*

✓*3. It is the second-largest town in the Belize District and largest in the Belize Rural South constituency.*

Step #2: Word Tokenization Breaking the sentence into individual words called as tokens. We can tokenize them whenever we encounter a space, we can train a model in that way. Even punctuations are considered as individual tokens as they have some meaning.

- ***Input :*** *San Pedro is a town on the southern part of the island of Ambergris Caye in the Belize District of the nation of Belize in Central America. According to 2015 mid-year estimates the town has a population of about 16444. It is the second-largest town in the Belize District and largest in the Belize Rural South constituency.*

- ***Output :*** *‘San’ , ‘Pedro’, ’ is’, ’a’, ’town’ , ‘on’ , ‘the’ ,’southern’ ,’part’ , ’of’, ‘island, ‘Ambergris’ ,’Caye’ ,’in’ ,‘Belize’ ,’District’ , ‘nation’ ‘Central’ ‘America’, ‘According’ , ‘to’ , ‘2015’ , ‘mid-year’ ,’estimates’ , ‘has’ , ‘population’ , ‘about’ ,’16444’, ‘It’ , ‘second-largest’ , ‘and’ , ‘largest’ , ‘Rural’ , ‘South’ , ‘constituency’.*

Step #3: Lemmatization Feeding the model with the root word.

- Large
- Largest

✓ South
✓ Southern

➢ Estimate
➢ Estimates
➢ Estimation

Step #4: Predicting Parts of Speech for each token

Input :

‘San’, ‘Pedro’, ’is’, ’a’, ’town’, ‘on’, ‘the’,’southern’,’part’,’of’, ‘island, ‘Ambergris’,’Caye’, ’in’,‘Belize’,’District’, ‘nation’ ‘Central’ ‘America’, ‘According’, ‘to’, ‘2015’, ‘mid-year’,’estimates’, ‘has’, ‘population’, ‘about’,’16444’, ‘It’, ‘second-largest’, ‘and’, ‘largest’, ‘Rural’, ‘South’, ‘constituency’.

Output :

Town - common noun

Is - verb

The - determiner

Of- Preposition

In-Preposition

San – noun

And so on.

- **Step #5:** Identifying stop words.

- There are various words in the English language that are used very frequently like ‘a’, ‘and’, ‘the’ etc. These words make a lot of noise while doing statistical analysis.
- We can take these words out.
- Some NLP pipelines will categorize these words as stop words, they will be filtered out while doing some statistical analysis.
- Definitely, they are needed to understand the dependency between various tokens to get the exact sense of the sentence.
- The list of stop words varies and depends on what kind of output are you expecting.

- **Step 6.1:** Dependency Parsing
- This means finding out the relationship between the words in the sentence and how they are related to each other.
- We create a parse tree in dependency parsing, with root as the main verb in the sentence.
- If we talk about the first sentence in our example, then 'is' is the main verb and it will be the root of the parse tree.
- We can construct a parse tree of every sentence with one root word(main verb) associated with it. We can also identify the kind of relationship that exists between the two words.
- In our example, 'San Pedro' is the subject and 'island' is the attribute. Thus, the relationship between 'San Pedro' and 'is', and 'island' and 'is' can be established. Just like we trained a Machine Learning model to identify various parts of speech, we can train a model to identify the dependency between words by feeding many words. It's a complex task though. In 2016, Google released a new dependency parser Parsey McParseface which used a deep learning approach.

- **Step 6.2:** Finding Noun Phrases We can group the words that represent the same idea.
-
- For example – It is the second-largest town in the Belize District and largest in the Belize Rural South constituency. Here, tokens ‘second’, ‘largest’ and ‘town’ can be grouped together as they together represent the same thing ‘Belize’.
- We can use the output of dependency parsing to combine such words.
- Whether to do this step or not completely depends on the end goal, but it’s always quick to do this if we don’t want much information about which words are adjective, rather focus on other important details.

•**Step #7:** Named Entity Recognition(NER)

•San Pedro is a town on the southern part of the island of Ambergris Caye in the Belize District of the nation of Belize, in Central America.

•Here, the NER maps the words with the real world places.

•The places that actually exist in the physical world. We can automatically extract the real world places present in the document using NLP.

•If the above sentence is the input, NER will map it like this way:

San Pedro - Geographic Entity

Ambergris Caye - Geographic Entity

Belize - Geographic Entity

Central America - Geographic Entity

- **Step #8:** Coreference Resolution

- San Pedro is a town on the southern part of the island of Ambergris Caye in the Belize District of the nation of Belize, in Central America. According to 2015 mid-year estimates, the town has a population of about 16, 444. It is the second-largest town in the Belize District and largest in the Belize Rural South constituency.

- Here, we know that ‘it’ in the sentence 6 stands for San Pedro, but for a computer, it isn’t possible to understand that both the tokens are same because it treats both the sentences as two different things while it’s processing them. Pronouns are used with a high frequency in English literature and it becomes difficult for a computer to understand that both things are same.

THANK YOU