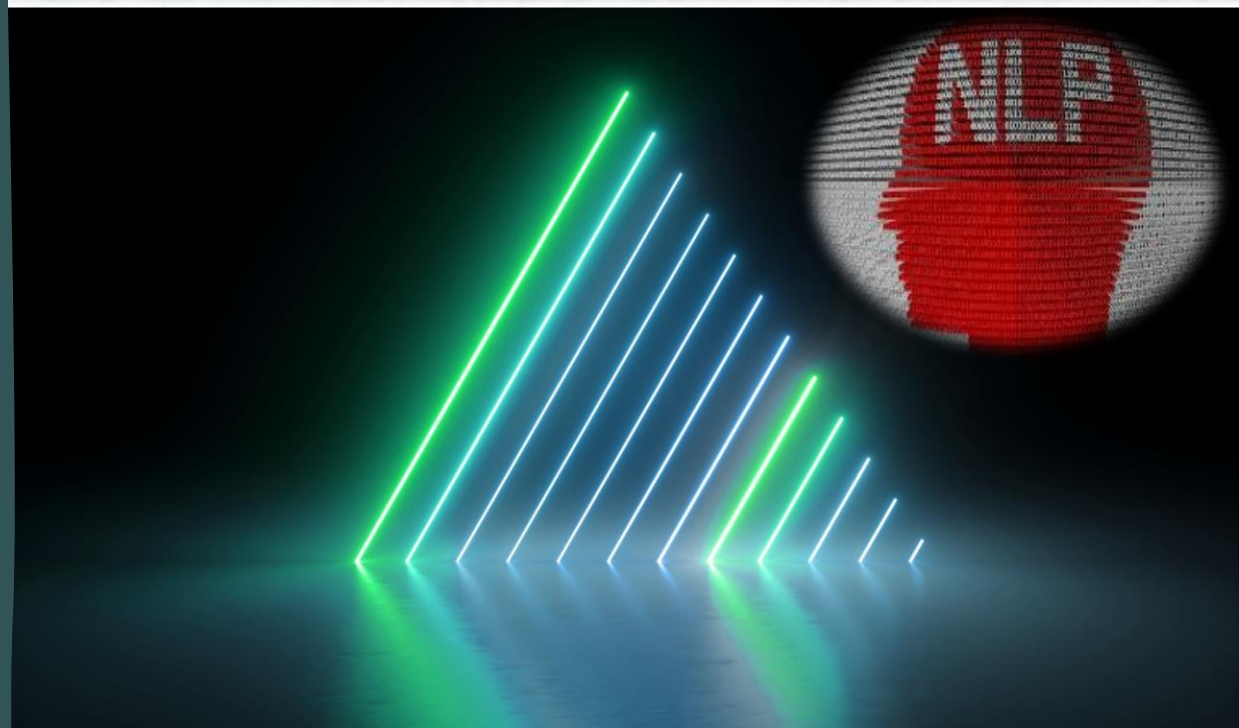
SESSION 2022-23 CLASS- 10 AI-417

NATURAL LANGUAGE PROCESSING

NLP



AI Domains

- Data Science
- Computer Vision
- Natural Language Processing

WHAT IS NLP?

Natural Language Processing (NLP) is a branch of Artificial Intelligence that enables computers to process human language in the form of text or voice data, understand its full meaning and mimic human conversation.

"Okay Google" with Android phones, or interacting with Siri on iPhones or iPads, or interacting with Amazon Alexa all these examples of NLP.

Tasks of NLP

- ✓ **Speech recognition (speech-to-text)** - is the task of taking input from audio or voice data and converting it into text data.

  e.g. when we say something to Siri or Google Assistant, it shows the text of what we said.

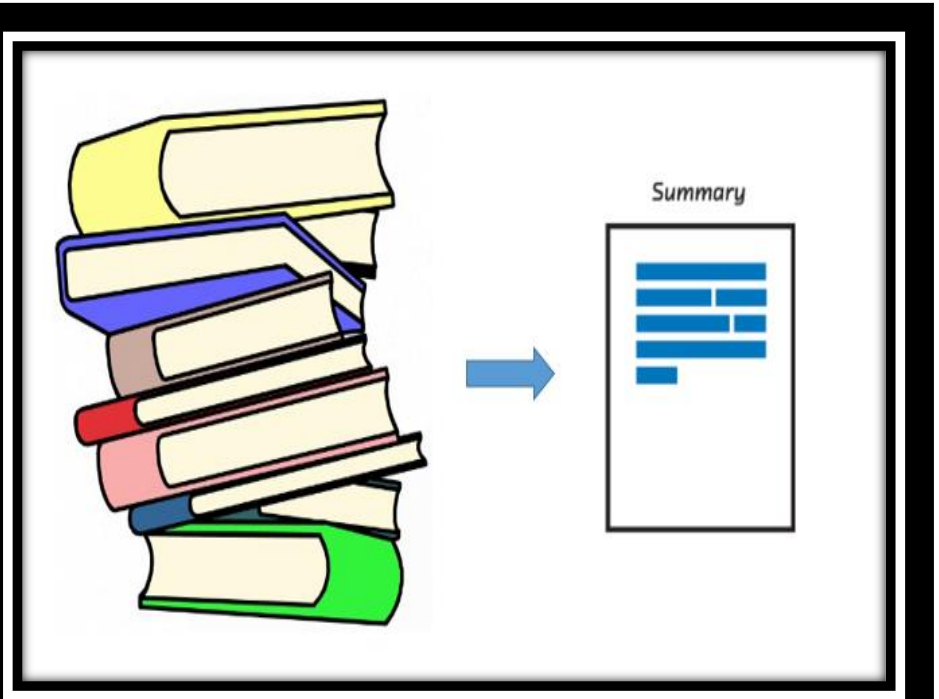
- ✓ **Word sense disambiguation** - The selection of the meaning of a word with multiple meanings.

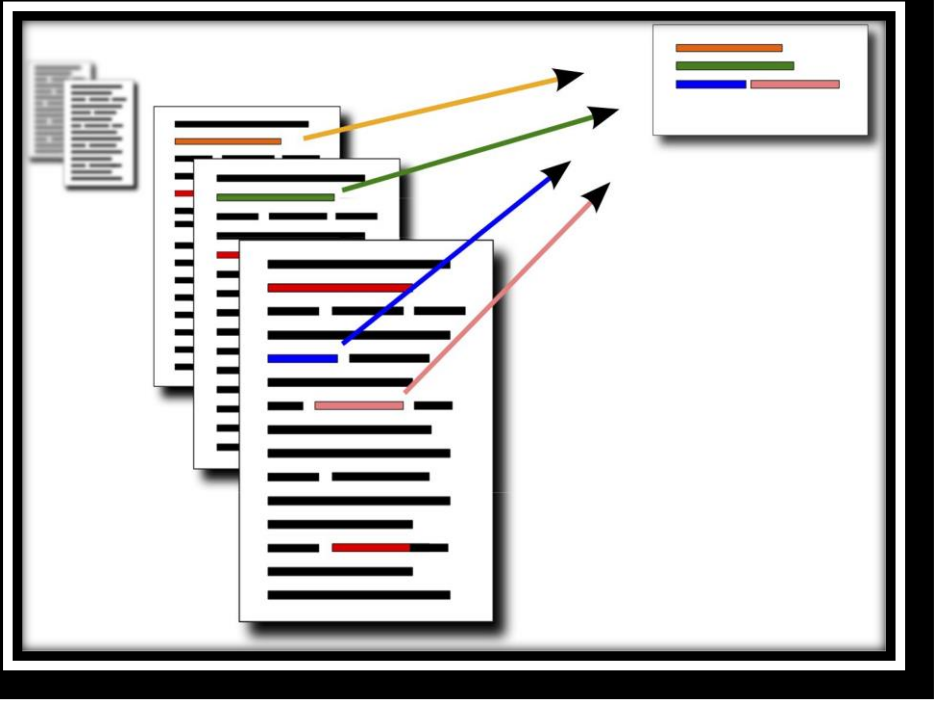
- ✓ **Sentiment analysis-** attempts to extract subjective qualities – attitudes, emotions, sarcasm, confusion, and suspicion – from the text.

- ✓ **Natural Language Generation –** is the opposite of speech recognition (i.e. text-to-speech) where the generated or processed information is reproduced in the form f human language. e.g. when we ask for some information from Siri or Google Assistant or Alexa and it replies in speech form, in a human-like language.

Application of

NLP





1. Automatic Text Summarisation

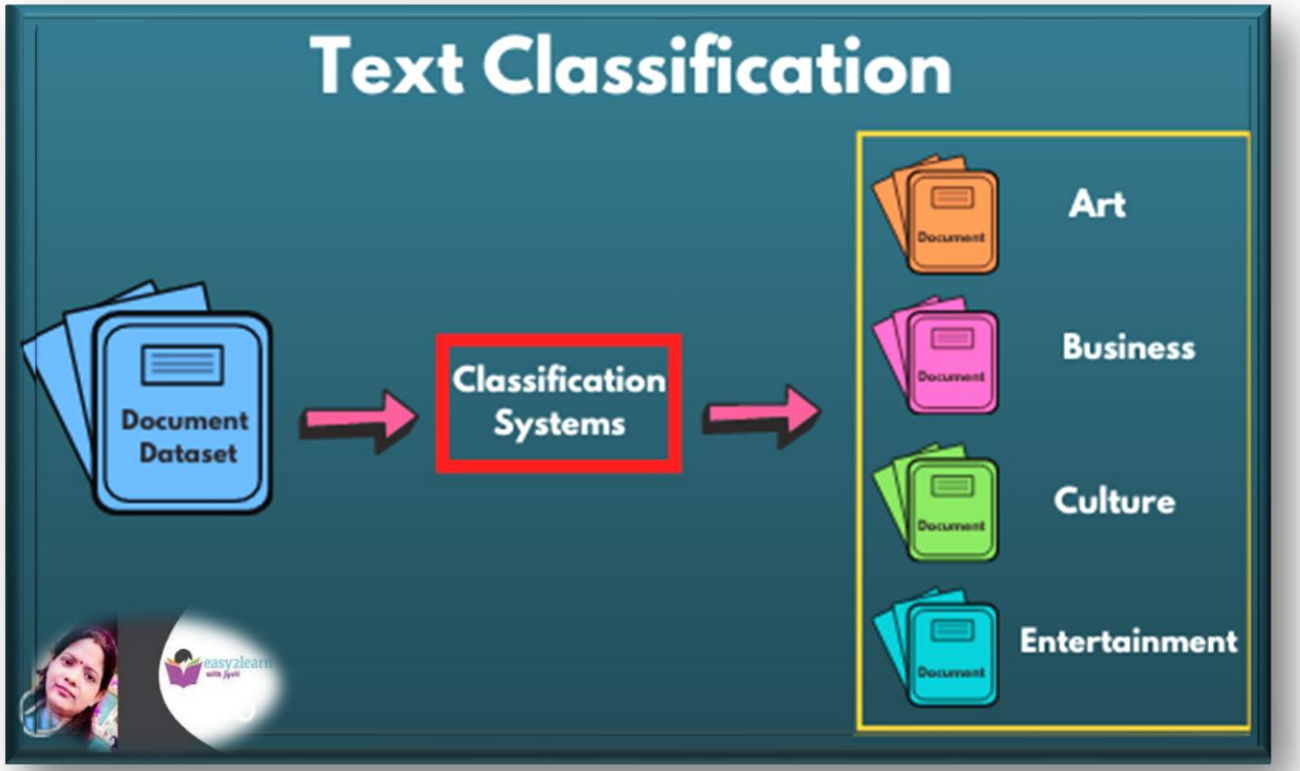
- An NLP technique where a computer program shortens longer texts and generates summaries to pass the intended message as the most important information is retained.
- Text Summarisation is most helpfully applied in Academic, research, or healthcare settings.
- Online and digital Newsletters use this for personalized content.
- Agencies apply this on social media posts to make the persona sketch of the person.

2. Sentiment Analysis

- ✓ The most exciting feature of NLP.
- ✓ It analyses the text or speech-to-text to recognize the sentiment or emotion expressed in it.
- ✓ Through NLP, sentiment analysis categorizes words as positive, negative, or neutral.
- ✓ NLP-enabled sentiment analysis understands the nuances and emotions in human voice and text.

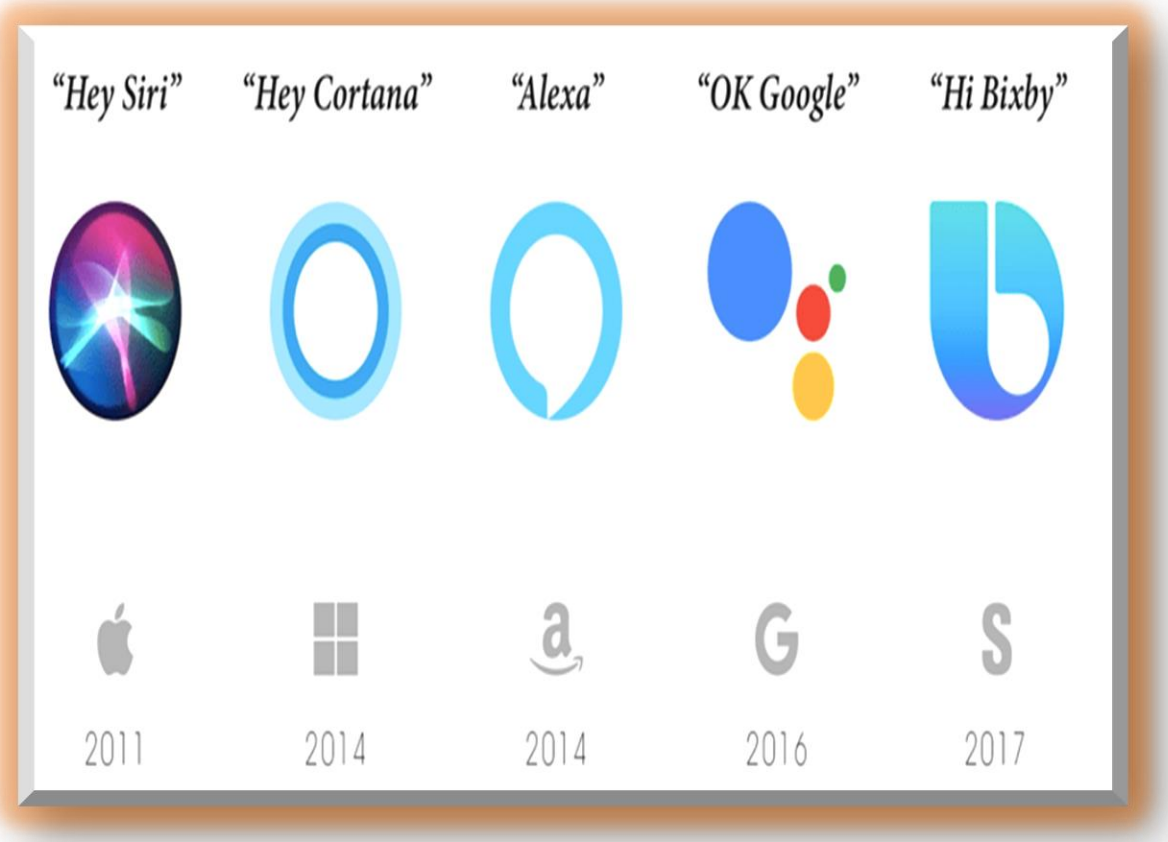
Text Classification is the process of understanding, analyzing, and categorizing unstructured text into organized groups using NLP and other AI technologies based on predefined tags and categories such as ‘sports articles’, ‘Technical reviews’,’ Political Essay’, and so forth

3. Text Classification



4.Smart Assistant

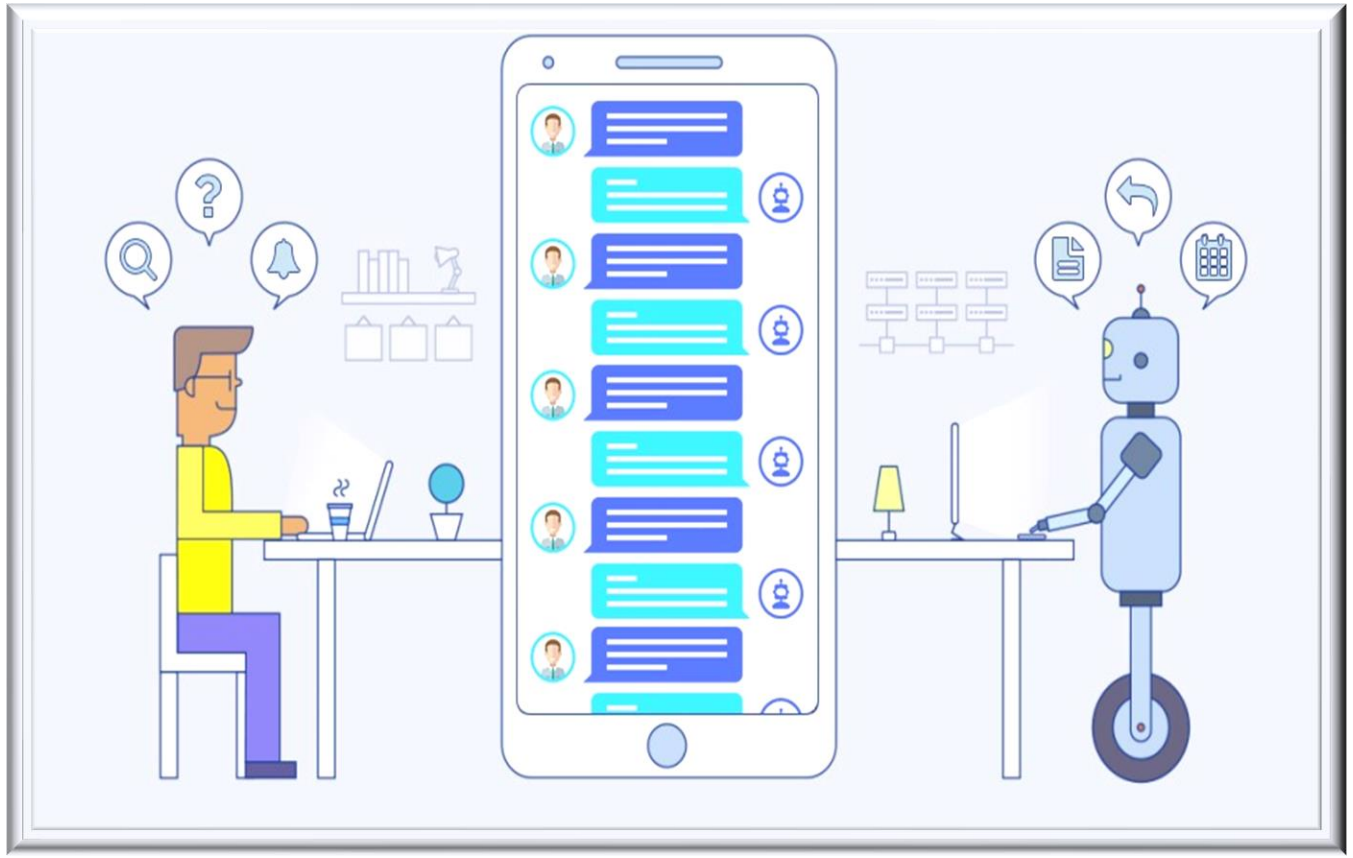
Virtual Assistants are NLP-based programs that are automated to communicate in the human voice, mimicking human interaction to help ease your day-to-day tasks, such as showing weather reports, creating reminders, making shopping lists, etc.



5.Digital Assistant

- ✓ We all hear "this call may be recorded for quality purposes," but rarely do we wonder what goes behind that. Other than for training, these recordings go into the database for an NLP system to learn from and improve in the future.
- ✓ Automated systems direct customer calls to service representatives or online chatbots, which respond to customer requests with helpful information.
- ✓ This is an NLP practice that many companies, including large telecommunications providers, have put to use.

CHATBOTS

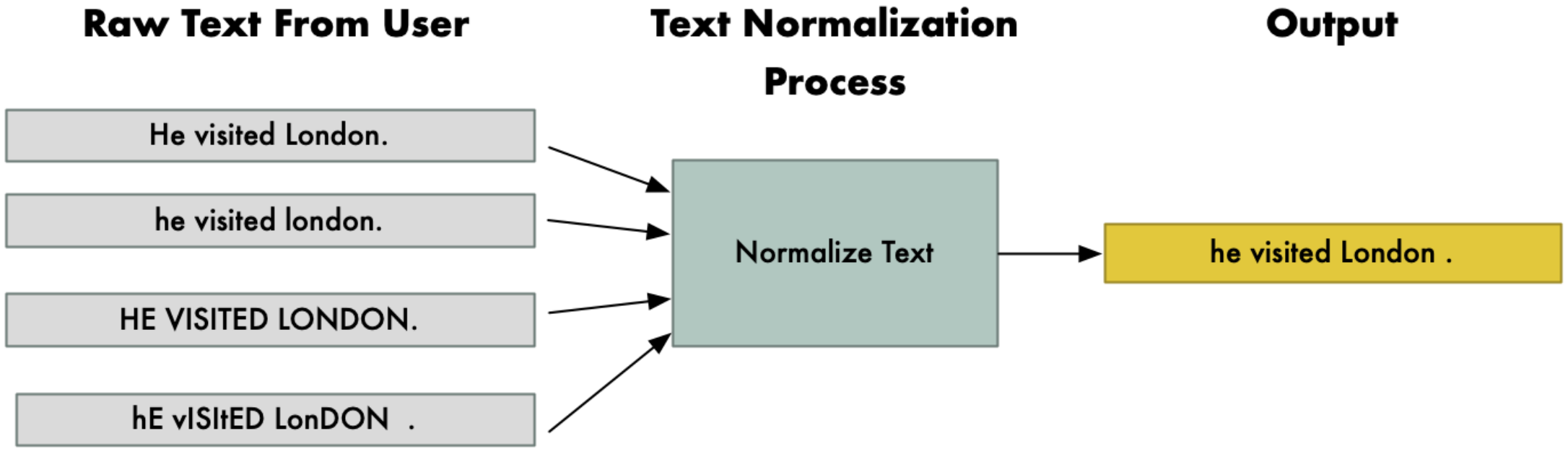


Modern-day chatbots that we often see in the form of a text box when we open a website or connect to customer service & interact with us in a text & based on the words used by us, they either connect us with customer support or redirect us to another webpage or link, are a classic example of NLP application today.

Concepts of NLP

Text Normalization

Text Normalization divides the text into smaller components called tokens (usually the words in the text) and groups related tokens together.



Steps of Text Normalisation

TEXT NORMALISATION

1. **Sentence Segmentation**
2. **Tokenization**
3. **Removing Stop Words, Special Characters, and numbers**
4. **Stemming**
5. **Converting Text to a Common Case**
6. **Lemmatisation**

1. Sentence Segmentation

Is the process of dividing the whole text into smaller components, i.e. individual sentences.

Example

Hello, world. AI is fun to know. It has started impacting our lives in many ways. Many more revolutionary technologies will soon evolve out of it.

- **Hello, world.**
- **Ai is fun to know.**
- **It has started impacting our lives in many ways.**
- **Many more revolutionary technologies will soon evolve out of it.**

Note – Whole Corpus gets reduced to sentences.

2. TOKENIZATION

Tokenization is the process of splitting up individual sentences into smaller units called tokens (any word, number or special character like (: , . , -).

My books are lying on the right side of my table. My books are covered with brown cover.

My	Books	Are
lying	On	the
right	side	of
My	Table	.
my	books	are
covered	with	brown
cover	.	

3. Removing Stop Words, Special Characters, and numbers

In this step, the unnecessary tokens are removed from the token list. i.e. removes the words that if removed, won't affect the meaning of the sentence. In other words, it removes the filler words that appear ' very frequently like 'and', 'the', and 'a. These words are often termed as "**stop words**" in NLP.

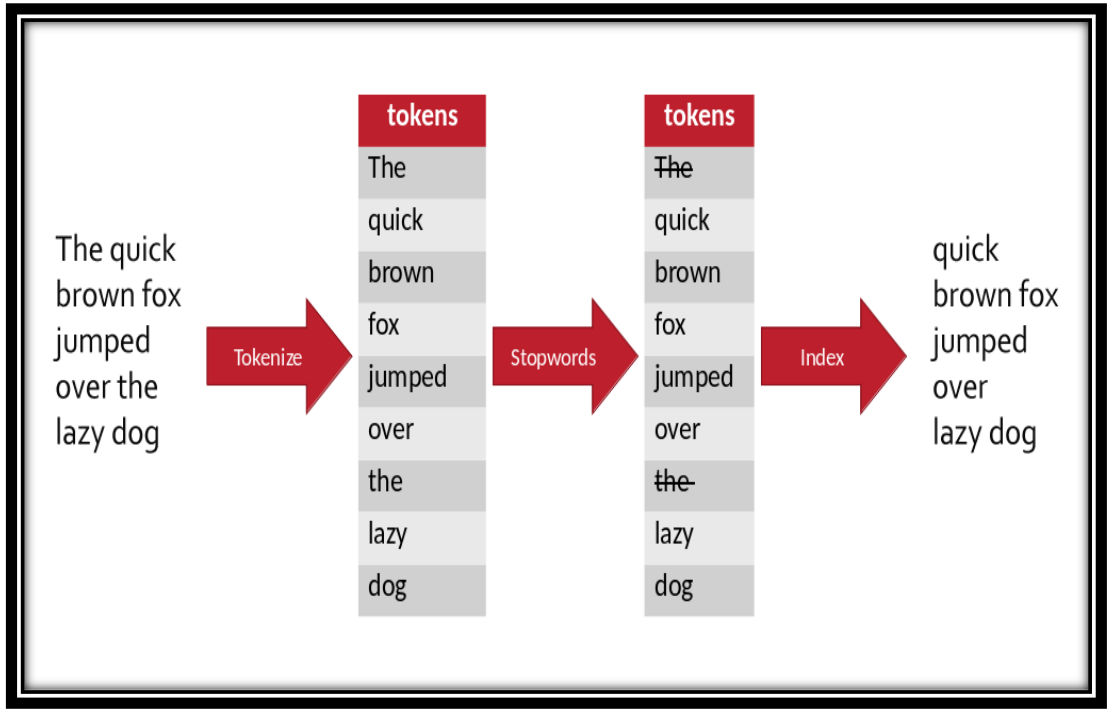
Advantages of removing stop words-

- ✓On removing the stop words, dataset size decreases as fewer tokens are now stored.
- ✓The time to train the model also decreases with the reduced size of the dataset.
- ✓It helps in improving performance, as there are fewer and only significant tokens left. Thus the classification accuracy of the model could be improved.

Stop Words

These words include:

- a
- I
- the
- in
- of
- for
- at
- to
- on
- with
- from



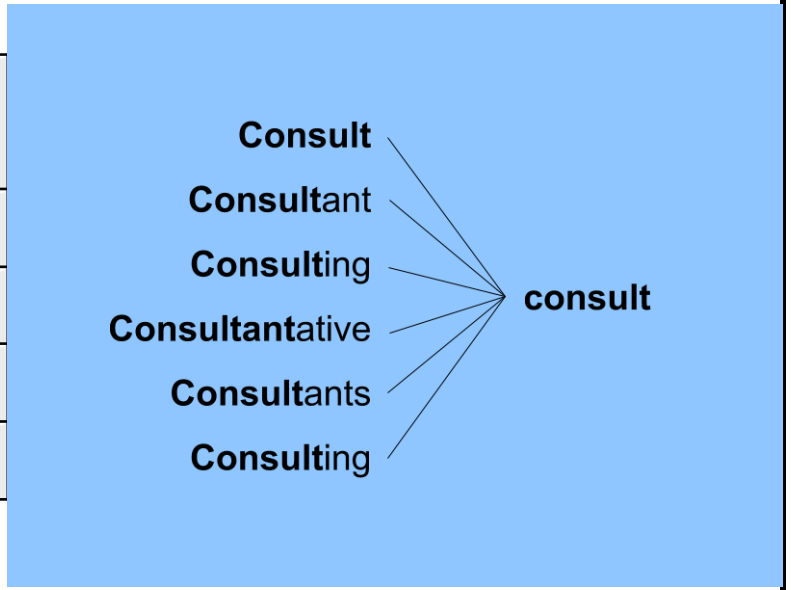
Stemming	Lemmatization
Stemming is a process that stems or removes last few characters from a word, often leading to incorrect meanings and spelling.	Lemmatization considers the context and converts the word to its meaningful base form, which is called Lemma.
For instance, stemming the word ‘Caring‘ would return ‘Car‘.	For instance, lemmatizing the word ‘Caring‘ would return ‘Care‘.
Stemming is fast process. Stemming just needs to get a base word and therefore takes less time.	Lemmatization takes more time as compared to stemming because it finds meaningful word/ representation.
Stemming has its application in **Sentiment Analysis.**	Lemmatization has its application in **Chatbots and human-answering.**
	

STEMMING

Stemming means converting a word into its stem (root form). In other words, the process of extracting the root form of the word is known as stemming.

For NLP, Stemming takes place by removing prefixes and suffixes at the beginning or end of each word. e.g.

Word in the sentence	Stem	Remarks
Studied	Studi	Affix **ed** removed
Standardize	Standard	Affix **ize** removed
Simplified	Simplifi	Affix **ed** removed
Drives	drive	Affix **s** removed



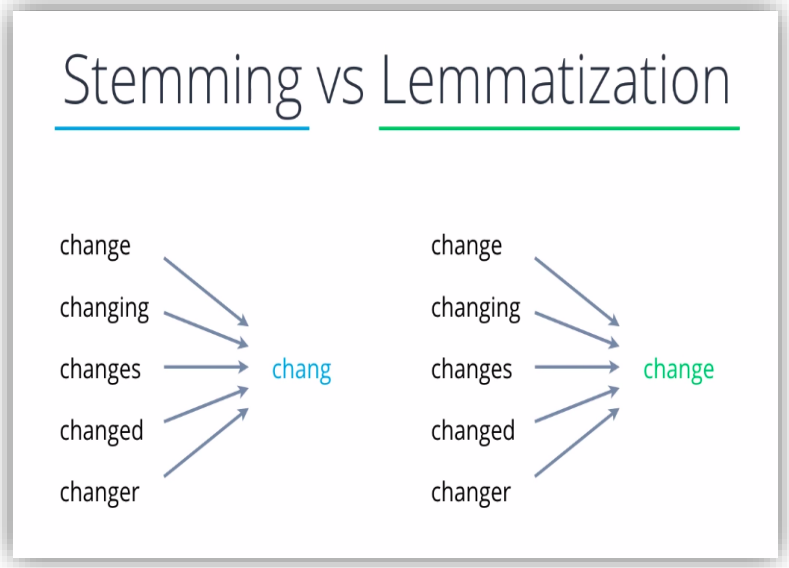
A Stem (root) is part of the word to which the inflectional (changing /deriving) affixes are added such as (ed, ize,s, mis, de) **A Stem may not be equal to a dictionary word.**

LEMMATISATION

Here in linguistic terms, the original word is taken e.g. drives, drove, driven, all are reduced as drive. i.e.

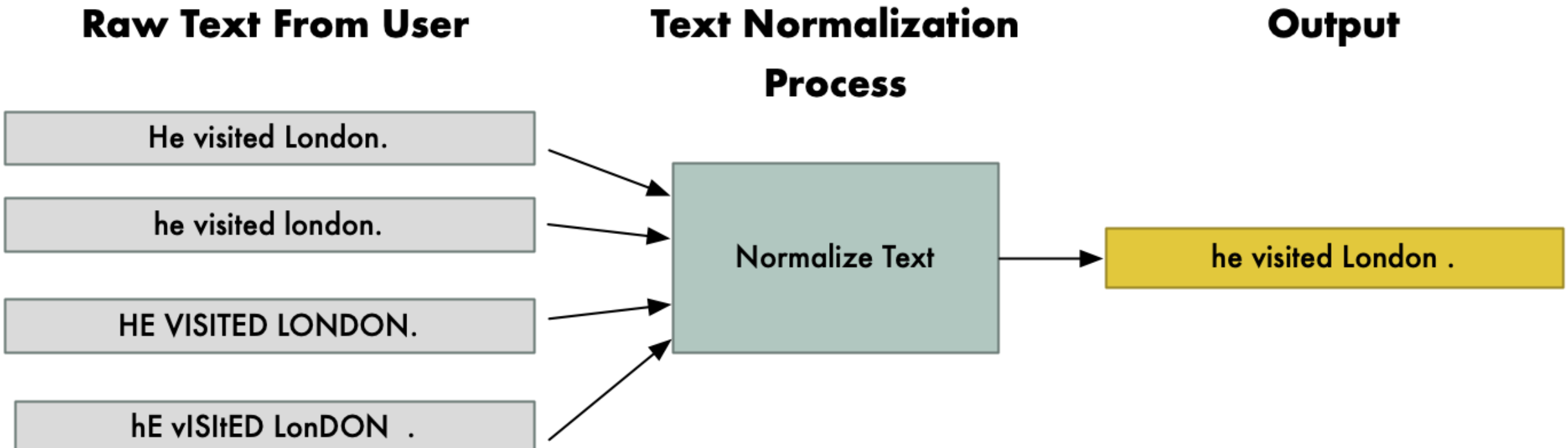
The process of converting a word to its actual root form linguistically (as per the language). The word extracted through lemmatization is called lemmas.

A Lemma is a base, root form of an inflectional word & that exists in a dictionary, unlike stems.



Case Normalization

Case Normalization refers to the conversion of all the words in the same case (often lowercase)



Finally Convert to Numbers

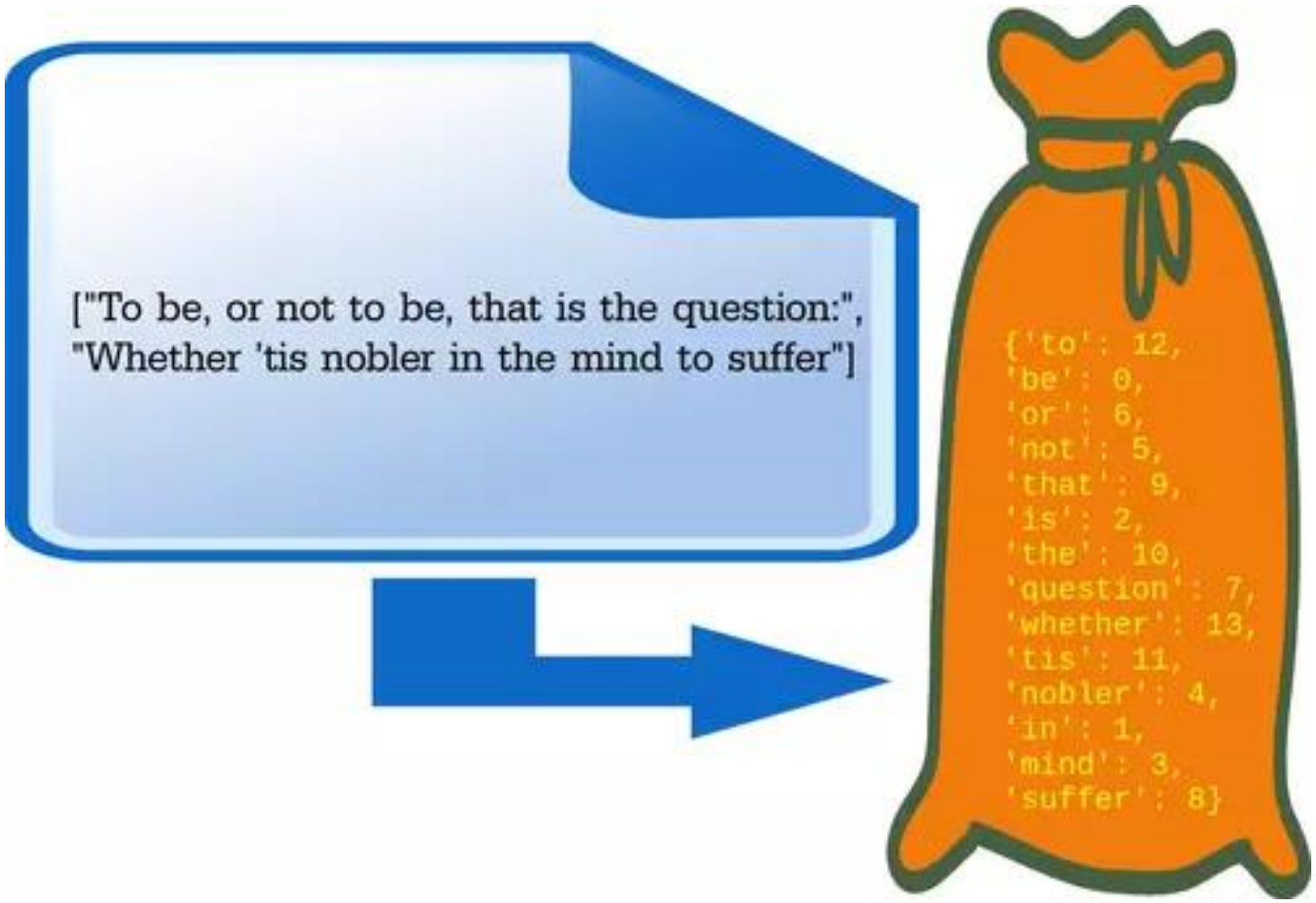
As Computer language understands Numbers better than alphabets and words, we have to convert the reduced information into numbers to enable computer understanding. Here, it is important to know which words are more important and carry more weight. To get to the reduced information and the most important words, there are many algorithms available.

A simple algorithm used for the same is Bag of Words algorithm.

Q- Samiksha, a class X student, was exploring the Natural Language Processing domain. She got stuck while performing the text normalization. Help her to normalize the text on the segmented sentences given below:
Document 1: Akash and Ajay are best friends.
Document 2: Akash likes to play football but Ajay prefers to play online games.

1. Tokenisation Akash, and, Ajay, are, best, friends Akash, likes, to, play, football, but, Ajay, prefers, to, play, online, games
2. Removal of stopwords Akash, Ajay, best, friends Akash, likes, play, football, Ajay, prefers, play, online, games
3. converting text to a common case akash, ajay, best, friends akash, likes, play, football, ajay, prefers, play, online, games
4. Stemming/Lemmatisation akash, ajay, best, friend akash, like, play, football, ajay, prefer, play, online, game

BAG OF WORDS (BoW)



- ✓ A **bag-of-Words** is a representation of text that describes the occurrence of words within a document.
- ✓ A bag of words contains two things :
  - (i) A vocabulary of known words.
  - (ii) A measure of the presence of known words (such as frequency of words)
- ✓ It is called a "**bag" of words** because it contains just the collection of words without any information about the order or structure of words in the document. It only tells that the known words occur in the document, not their position in the document.

A **Bag-of-words (BoW)** model is a model used for extracting features from the text for use in modeling with many AI algorithm.

Steps of implementing Bag-of-words Model

1. **Text Normalization**
2. **Design the vocabulary or Dictionary**
3. **Create Document Vector**
4. **Calculate TF-IDF**