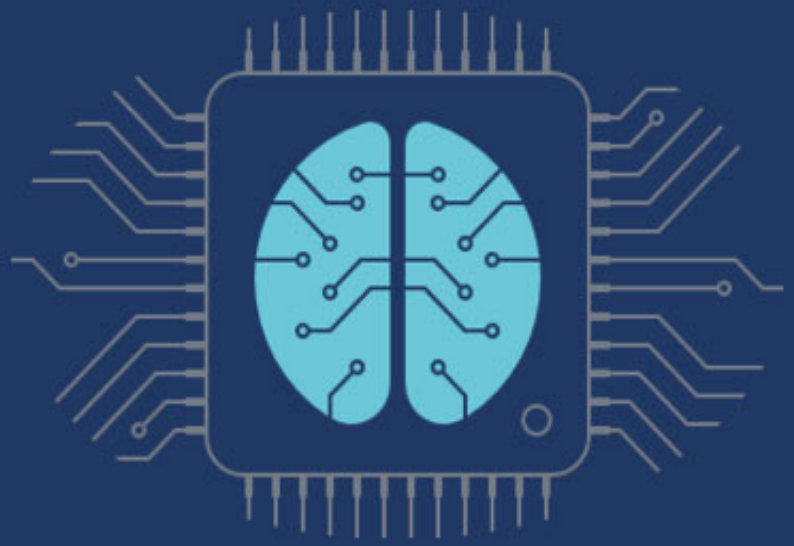
Unit 6



NLP (NATURAL LANGUAGE PROCESSING)

Includes Important Questions

Introduction to

Natural Language Processing - NLP

What is NLP ?

Natural Language Processing, or NLP, is the sub-field of AI that is ***focused on enabling computers to understand and process human languages***. AI is a subfield of *Linguistics*, *Computer Science*, *Information Engineering*, and *Artificial Intelligence concerned with the interactions between computers and human (natural) languages*, in particular *how to program computers to process and analyze large amounts of natural language data*.

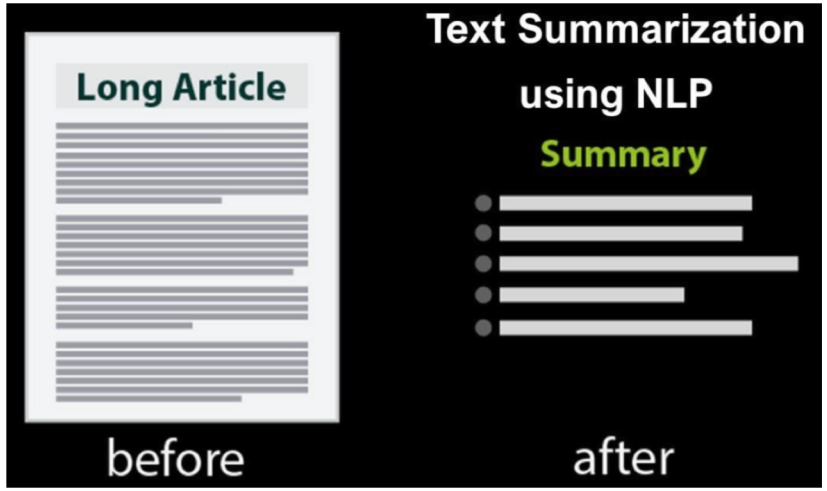
But how do computers do that?

How do they understand what we say in our language?

Applications of

Natural Language Processing

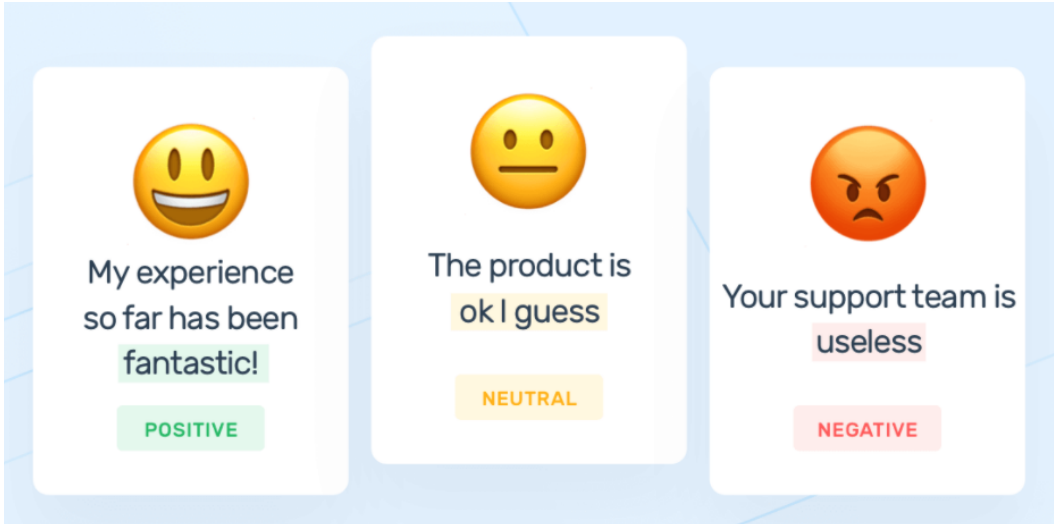
Automatic Summarization: *Information overload is a real problem when we need to access a specific, important piece of information from a huge knowledge base*. Automatic summarization is especially relevant when used to ***provide an overview of a news item or blog post***, ***while avoiding redundancy from multiple sources*** and ***maximizing the diversity of content obtained***.



Applications of

Natural Language Processing

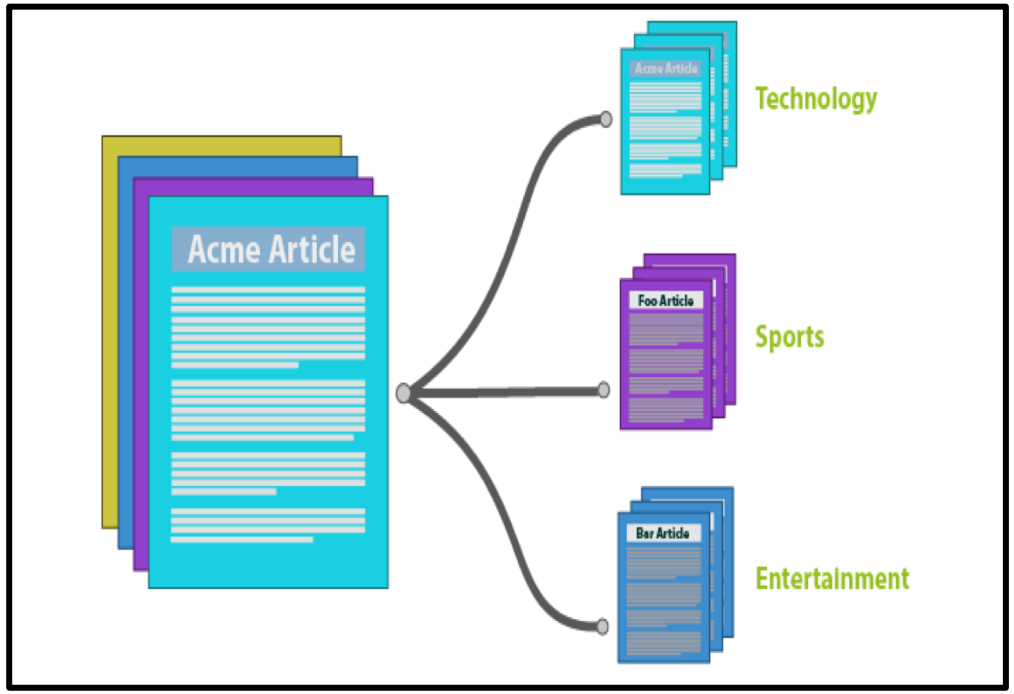
Sentiment Analysis: The goal of sentiment analysis is to identify sentiment among several posts or even in the same post where emotion is not always explicitly expressed. Companies use Natural Language Processing applications, such as sentiment analysis, to identify opinions and sentiment online to help them understand what customers think about their products and services (i.e., "***I love the new iPhone***" and, a few lines later "***But sometimes it doesn't work well***" where the person is still talking about the iPhone) and overall indicators of their reputation. Beyond determining simple polarity, sentiment analysis understands ***sentiment in context to help better understand what's behind an expressed opinion***, which can be ***extremely relevant in understanding and driving purchasing decisions***.



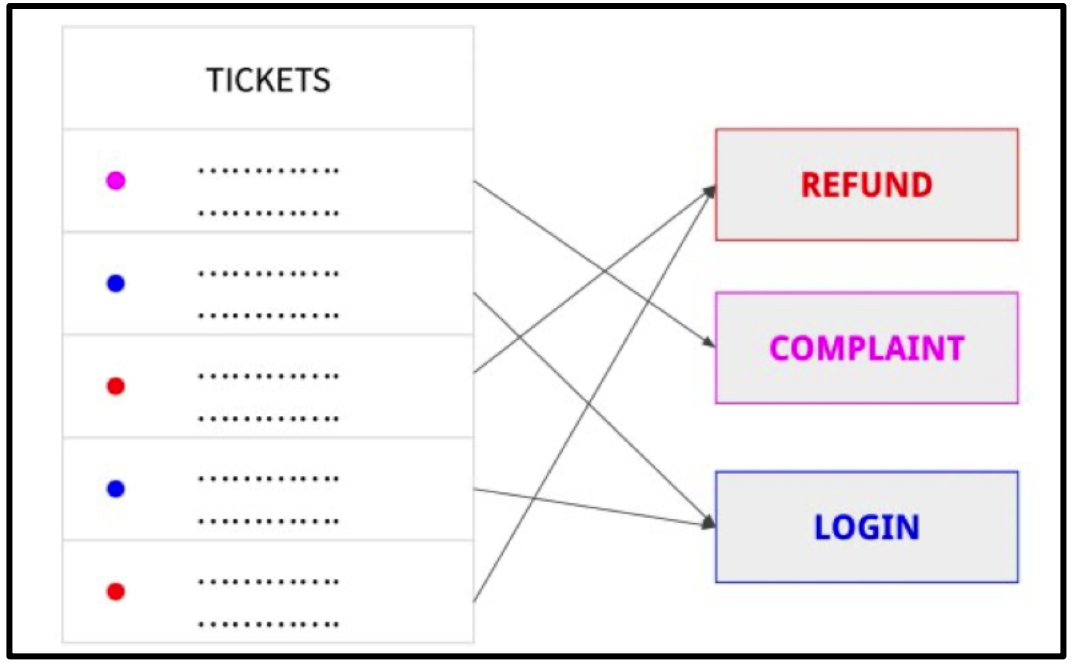
Applications of

Natural Language Processing

Text classification: ***Text classification makes it possible to assign predefined categories to a document*** and organize it to help you find the information you need or simplify some activities. For example, ***an application of text categorization is spam filtering in email***.



Example # 1

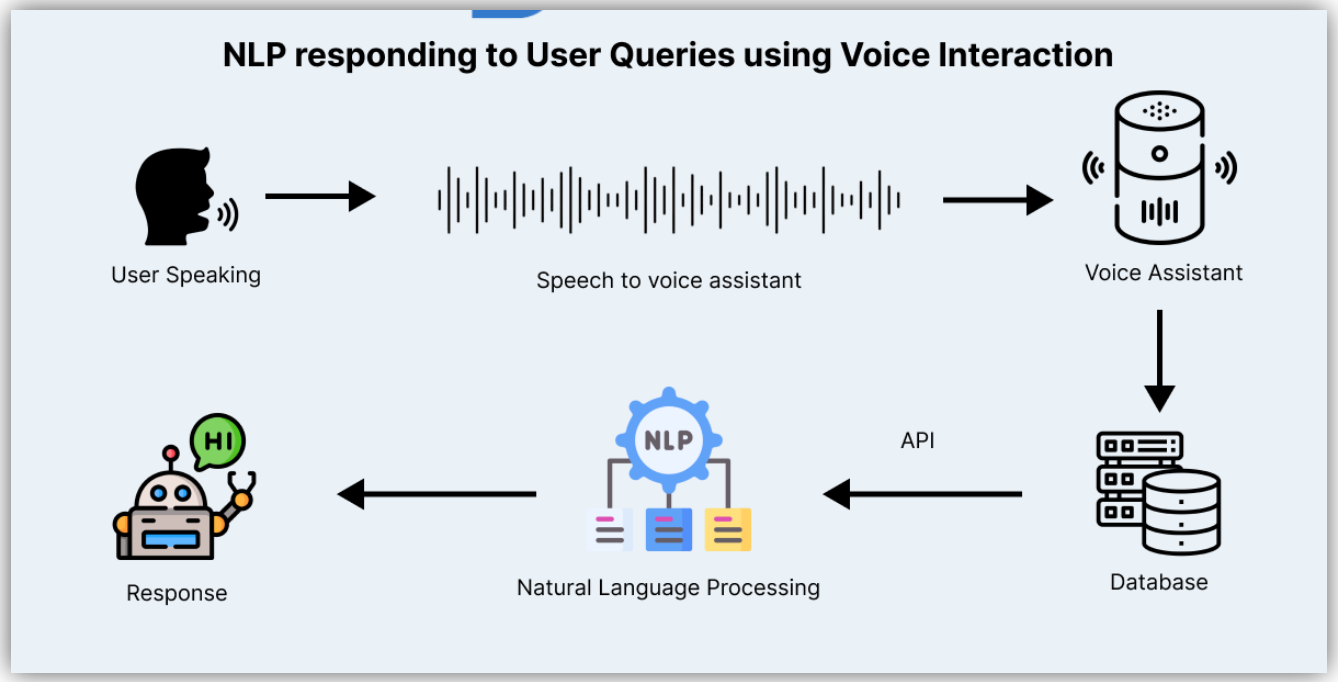


Example # 2

Applications of

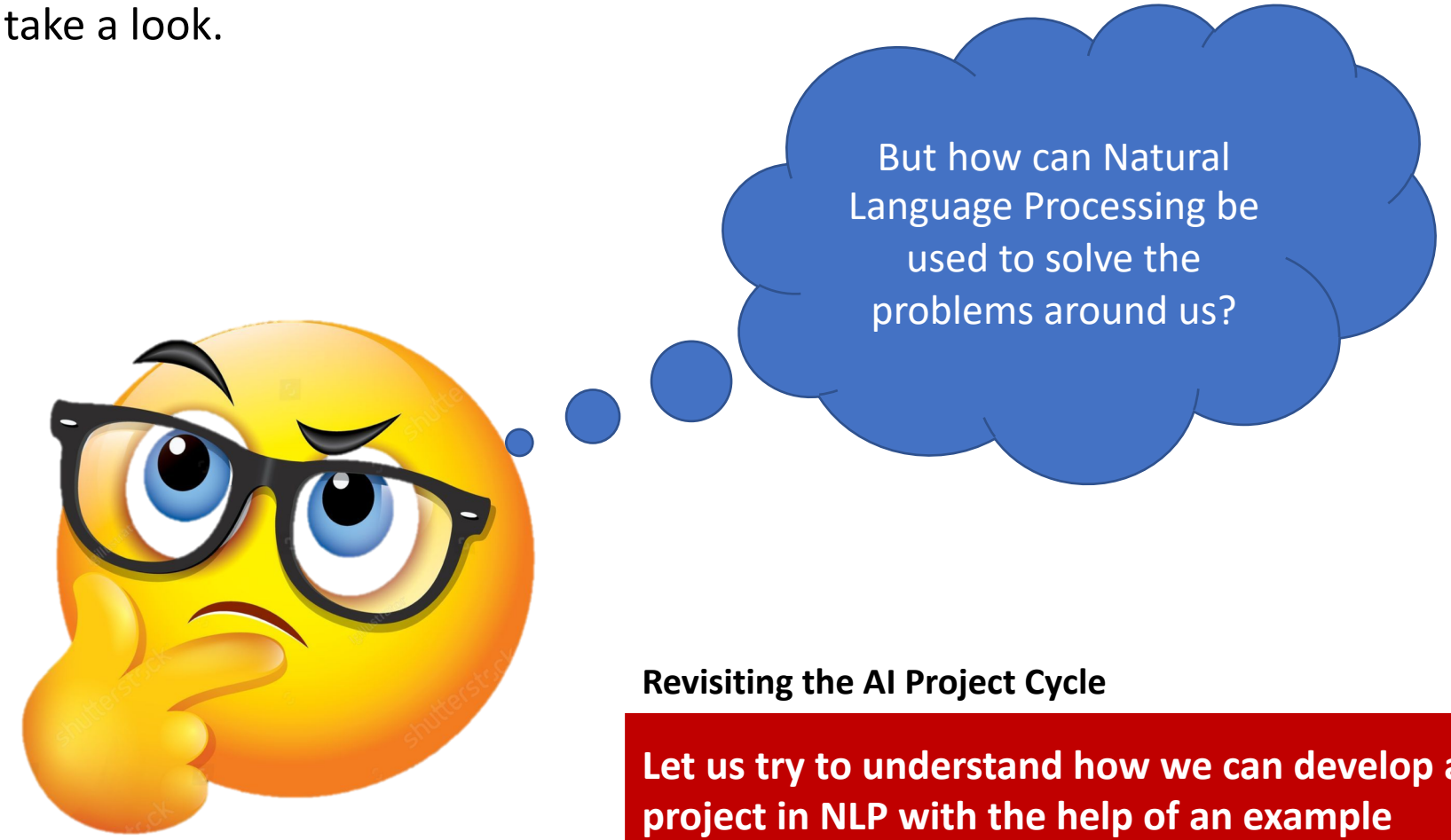
Natural Language Processing

Virtual Assistants: ***Google Assistant, Cortana, Siri, Alexa***, etc. have become an integral part of our lives. Not only can we talk to them but they also have the abilities to make our lives easier. By accessing our data, ***they can help us in keeping notes of our tasks, make calls for us, send messages and a lot more***. With the help of speech recognition, these assistants can not only detect our speech but can also make sense out of it.



Natural Language Processing

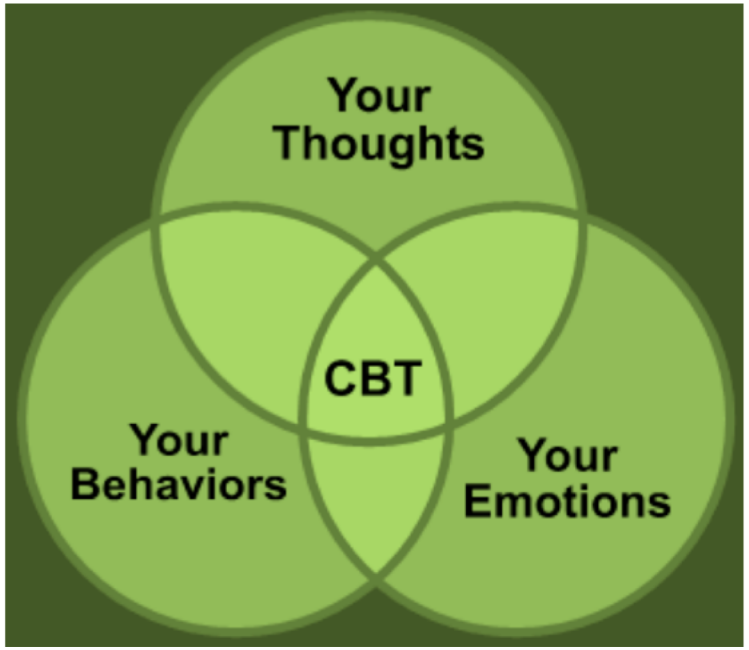
Natural Language Processing is all about ***how machines try to understand and interpret human language and operate accordingly***. Let us take a look.

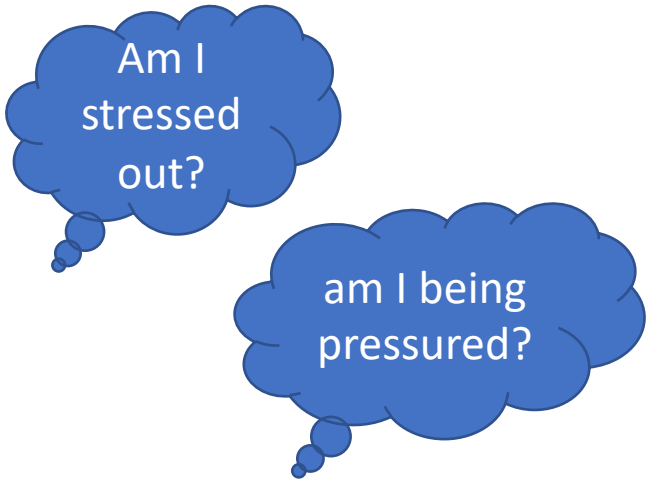


Revisiting the AI Project Cycle

Let us try to understand how we can develop a project in NLP with the help of an example

Scenario





People face competition in even the tiniest tasks and are expected to give their best at every point in time. When people are unable to meet these expectations, they get stressed and could even go into depression. We get to hear a lot of cases where people are ***depressed due to reasons like peer pressure, studies, family issues, relationships***, etc. and they eventually get into something that is bad for them as well as for others. So, to overcome this, ***cognitive behavioral therapy (CBT)*** is considered to be one of the ***best methods to address stress*** as it is easy to implement on people and also gives good results. This therapy includes understanding the behavior and mindset of a person in their normal life.

To understand more about the concept of this therapy, visit this link:
https://en.wikipedia.org/wiki/Cognitive_behavioral_therapy

Problem Scoping

CBT is a technique used by most therapists to cure patients out of stress and depression. ***But it has been observed that people do not wish to seek the help of a psychiatrist willingly. They try to avoid such interactions as much as possible. Thus, there is a need to bridge the gap between a person who needs help and the psychiatrist***. Let us look at various factors around this problem through the 4Ws problem canvas.

Who Canvas – Who has the Problem?

Who are the stakeholders?	○ People who suffer from stress and are at the onset of depression.
What do we know about them?	○ People who are going through stress are reluctant to consult a psychiatrist.

What Canvas – What is the nature of the problem?

What is the problem?	○ People who need help are reluctant to consult a psychiatrist and hence live miserably.
How do you know it is a problem?	○ Studies around mental stress and depression available on various authentic sources.

Where Canvas – Where does the problem arise?

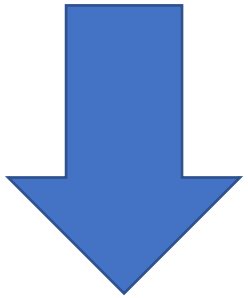
What is the context/situation in which the stakeholders experience this problem?	○ When they are going through a stressful period of time ○ Due to some unpleasant experiences

Where Canvas – Where does the problem arise?

What would be of key value to the stakeholders?	○ People get a platform where they can talk and vent out their feelings anonymously ○ People get a medium that can interact with them and applies primitive CBT on them and can suggest help whenever needed
How would it improve their situation?	○ People would be able to vent out their stress ○ They would consider going to a psychiatrist whenever required

Problem Summary / Problem Template

Our	People undergoing stress	Who?
Have a problem of	Not being able to share their feelings	What?
While	They need help in venting out their emotions	Where?
An ideal solution would	Provide them a platform to share their thoughts anonymously and suggest help whenever required	Why



"To create a chatbot which can interact with people, help them to vent out their feelings and take them through primitive CBT."

Data Acquisition

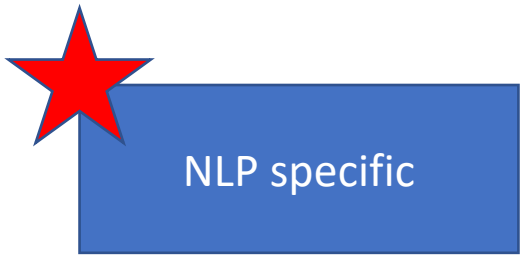
To understand the *sentiments of people*, we need to ***collect their conversational data*** so that machine can interpret the words that they use and understand their meaning. Such data can be collected from various means:

Data Exploration

Once the ***textual data has been collected***, it needs to be processed and cleaned so that an easier version can be sent to the machine. Thus, the text is normalized through various steps and is lowered to minimum vocabulary since the machine does not require grammatically correct statements but the essence of it.

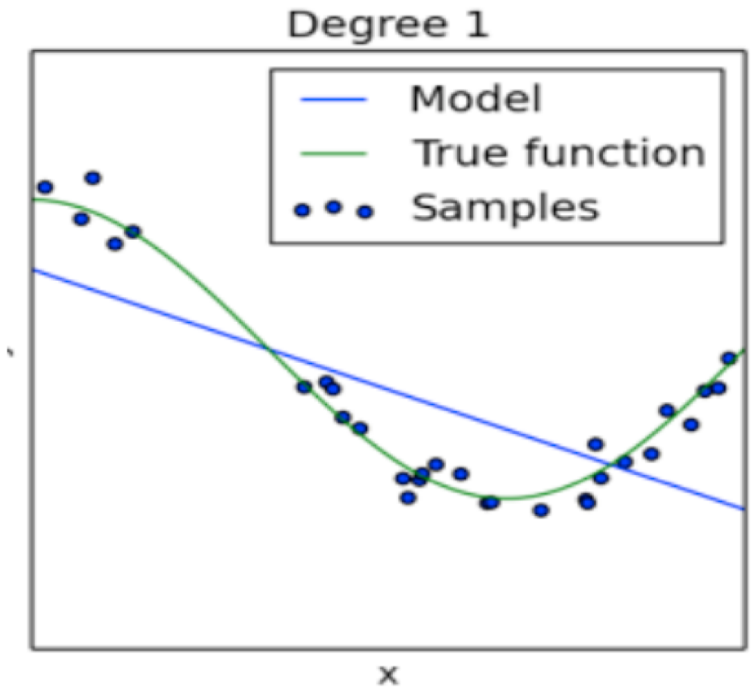
Modelling

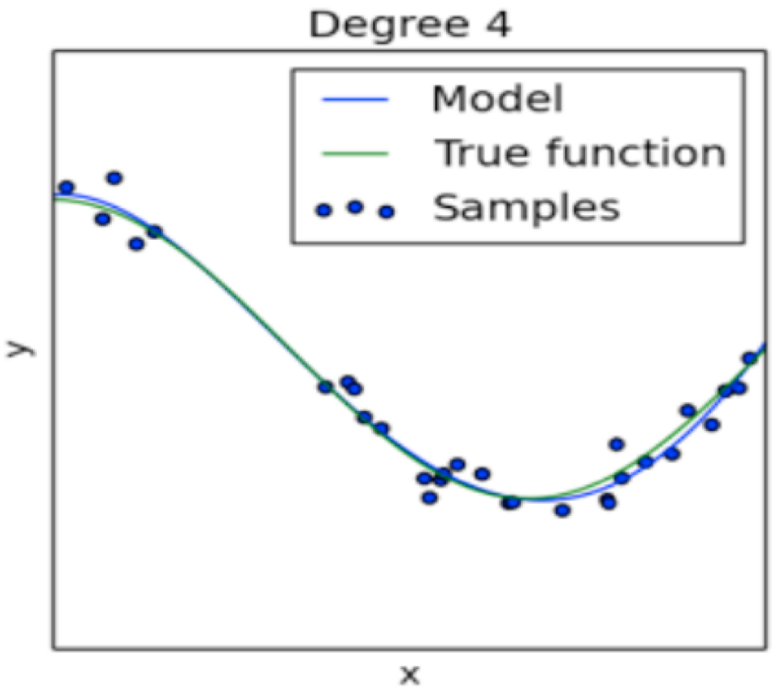
Once the text has been ***normalized***, it is then fed to an NLP based AI model. Note that in NLP, modelling requires data pre-processing only after which the data is fed to the machine. Depending upon the type of ***chatbot*** we try to make, there are a lot of AI models available which help us build the foundation of our project.

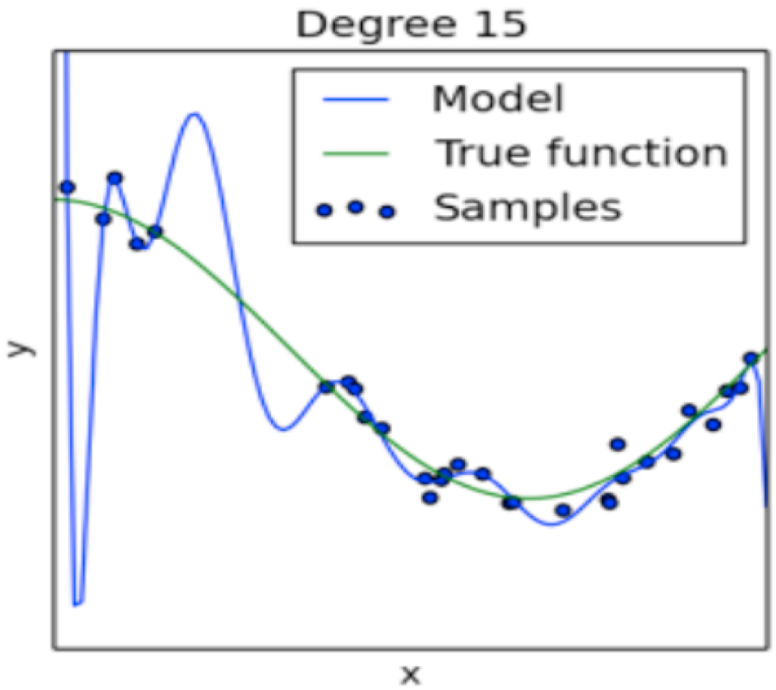


Evaluation

The model trained is then ***evaluated*** and the ***accuracy*** for the same is generated on the basis of the relevance of the answers which the ***machine*** gives to the user's ***responses***. To understand the ***efficiency of the model***, the ***suggested answers*** by the chatbot are compared to the ***actual answers***.







The model's output does not match the true function at all. Hence the model is said to be ***underfitting*** and its accuracy is lower.

The model's performance matches well with the true function which states that the model has optimum accuracy and the model is called a **perfect fit**.

The model performance is trying to cover all the data samples even if they are out of alignment to the true function. This model is said to be **overfitting** and this too has a lower accuracy.

Chatbots

One of the most common ***applications*** of NLP is a ***chatbot***. There are a lot of chatbots available and many of them use the same approach as we used in the scenario above. Let us try some of the chatbots and see how they work.

	• **Mitsuku Bot*** https://www.pandorabots.com/mitsuku/
	• **CleverBot*** https://www.cleverbot.com/
	• **Jabberwacky*** http://www.jabberwacky.com/
	• **Haptik*** https://haptik.ai/contact-us
	• **Rose*** http://ec2-54-215-197-164.us-west-1.compute.amazonaws.com/speech.php
	• **Ochatbot*** https://www.ometrics.com/blog/list-of-fun-chatbots/

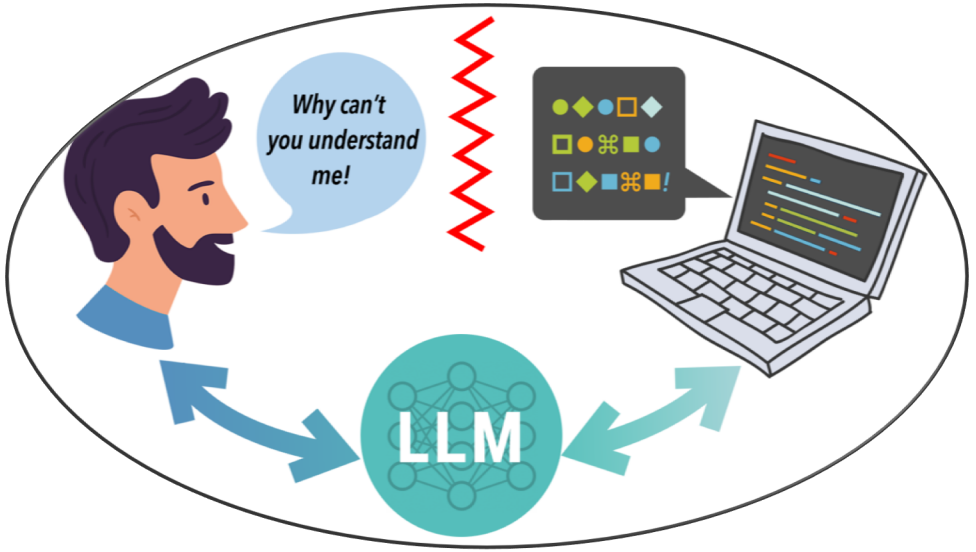
Chatbots

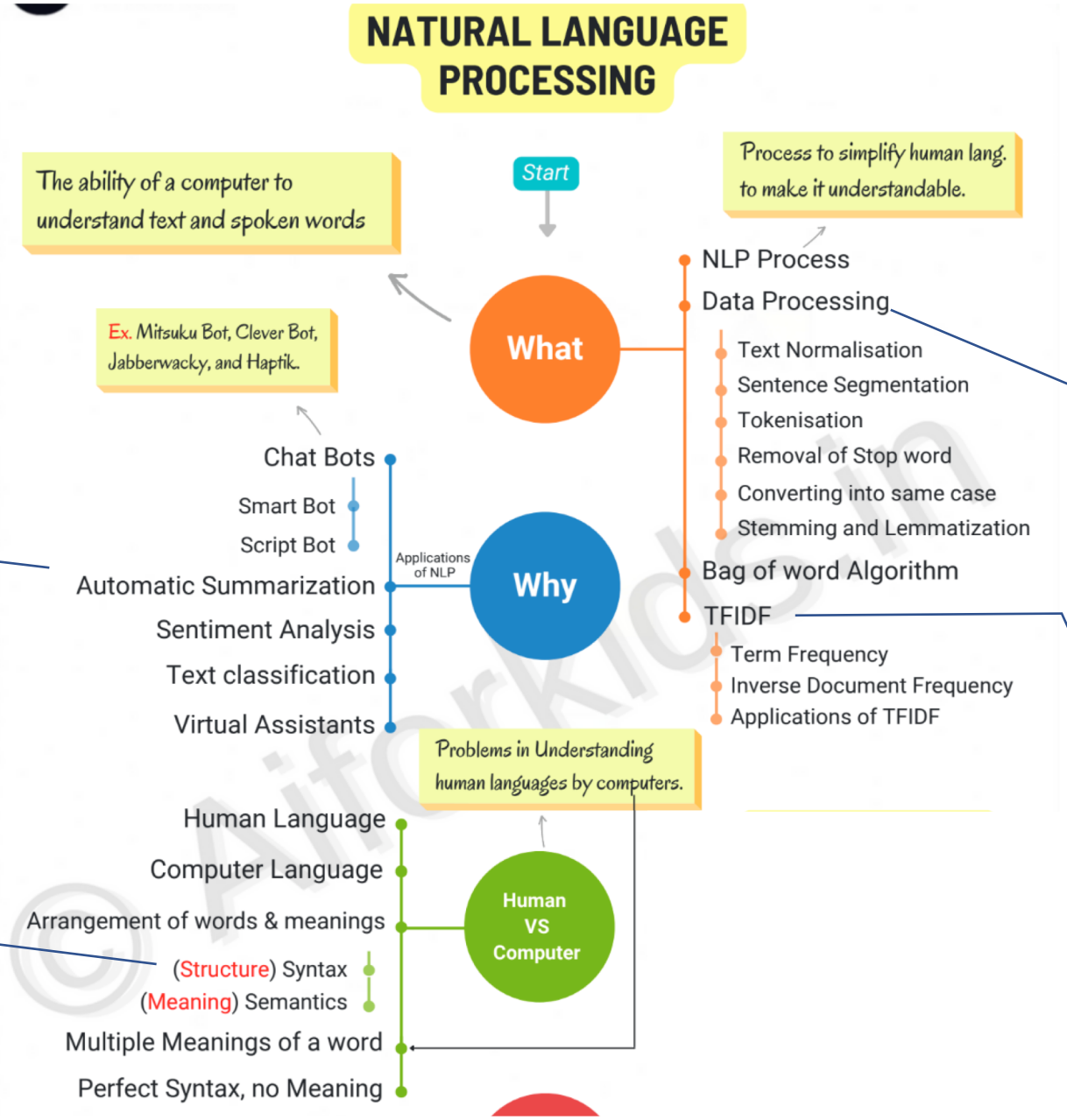
As you interact with more and more chatbots, you would realize that some of them are ***scripted*** or in other words are ***traditional chatbots*** while others were ***AI-powered and had more knowledge***. With the help of this experience, we can understand that there are 2 types of chatbots around us: ***Script- bot*** and ***Smart-bot***. Let us understand what each of them mean in detail:

Script-bot	Smart-bot
Script bots are easy to make	Smart-bots are flexible and powerful
Script bots work around a script which is programmed in them	Smart bots work on bigger databases and other resources directly
Mostly they are free and are easy to integrate to a messaging platform	Smart bots learn with more data
No or little language processing skills	Coding is required to take this up on board
Limited functionality	Wide functionality

Human Language Vs Computer Language

- Humans need language to communicate, which we constantly process. ***Our brain continuously processes the sounds it hears around us and works to make sense of them***. Our brain continuously processes and stores everything, even as the teacher is delivering the lesson in the classroom.
- The Computer Language is understood by the computer, on the other hand. ***All input must be transformed to numbers before being sent to the machine***. And if a single error is made while typing, the machine throws an error and skips over that area. Machines only use extremely simple and elementary forms of communication.





Stages of Data Processing

Applications of NLP

Widely used Statistical Method for NLP

Human Language elements

Data Processing

- Data Processing is a *method of manipulation of data*. It means the conversion of raw data into meaningful and machine-readable content.
- Since human languages are complex, we need to simplify them in order to make sure that the understanding becomes possible.

1. Text Normalization

- The process of converting a text into a ***Canonical*** (Standard) from known as ***text normalization***. For instance, the canonical form of the word “good” can be created from the words “goooood” and “gud”.

2.Sentence Segmentation

- Under sentence segmentation, the whole corpus is divided into sentences. Each sentence is taken as a different data so now the whole corpus gets reduced to sentences.

3.Tokenization

- ***Sentences*** are first broken into ***segments***, and then each segment is further divided into ***tokens***. Any ***word, number, or special character*** that appears in the sentence is referred to as a token.

4.Removing Stopwords, Special Characters and Numbers

- In this step, the tokens which are not necessary are removed from the token list. Stopwords are words that are used frequently in a corpus but provide nothing useful. Stopwords include ***a, an, and, or, for, it, is***

5.Converting text to a common case

- After eliminating the stopwords, we change the text's case throughout, preferably to lower case.

6.Stemming – "stem" (technique)

- Removes prefix / suffixes from the words to find their base form. Stemming is quick and efficient, but it can sometimes produce unmeaningful base forms. For example, "swimmer," "swimming," and "swims" all have the root word "swim". However, stemming can sometimes produce words that don't exist in the dictionary.

6.1.Lemmatization – "lemma" (technique)

- Analyzes the context of a word in a sentence to reduce it to its root form. Lemmatization is more accurate than stemming because it considers the word's use in the larger text. For example, the verb "running" would be identified as "run". Lemmatization is more complex and computationally intensive than stemming.

Bag of Words – “BoW” (is a model technique)

- A bag-of-words is a textual illustration that shows where words appear in a document. There are two components: A collection of well-known words. A metric for the amount of well-known words.
- **NLP** model called Bag of Words aids in the extraction of textual information that can be used by machine learning techniques. That gather the instances of each term from the bag of words and create the corpus’s vocabulary.

Here is the step-by-step approach to implement bag of words algorithm.

1. ***Text Normalization:*** *Collect data and pre-process it.*
2. ***Create Dictionary:*** *Make a list of all the unique words occurring in the corpus (Vocabulary)*
3. ***Create document vectors:*** *For each document in the corpus, find out how many times the word from the unique list of words has occurred.*
4. *Create document vectors for all the documents.*

Term Frequency - TF

- The measurement of a term's frequency inside a document is called term frequency. The simplest calculation is to count the instances of each word.

Inverse Document Frequency - IDF

- A key concept used in information retrieval and natural language processing to evaluate the importance of a term (or word) in a collection of documents.

Applications of TFIDF

- TFIDF is commonly used in the NLP domain.

Document Classification	Topic Modelling	Information Retrieval System	Stop word filtering
Helps in classifying the type and genre of a document.	It helps in predicting the topic for a corpus.	To extract the important information out of a corpus.	Helps in removing the unnecessary words out of a text body.

Unit 6 - NLP

Important Questions

What is a Chabot?
A chatbot is a computer program that's designed to simulate human conversation through voice commands or text chats or both. Eg: Mitsuku Bot, Jabberwacky etc. (OR)

A chatbot is a computer program that can learn over time how to best interact with humans. It can answer questions, troubleshoot customer problems, evaluates and (OR)

A chatbot is a computer program designed to simulate conversation with human users. A chatbot is also known as an artificial conversational entity (ACE), chat robot, talk bot, chatterbot, or chatterbox. (OR)

A chatbot is a software application used to conduct an online chat conversation via text or text-to-speech, in lieu of providing direct contact with a live human agent.

Which package is used for Natural Language Processing in Python programming?
Natural Language Toolkit (NLTK). NLTK is one of the leading platforms for building Python programs that can work with human language data.

Important Questions

What is the full form of NLP?
Natural Language Processing

While working with NLP what is the meaning of Syntax and Semantics
Syntax: Syntax refers to the grammatical structure of a sentence.
Semantics: It refers to the meaning of the sentence

What is the full form of TFIDF?
Term Frequency and Inverse Document Frequency

What is meant by a dictionary in NLP?
Dictionary in NLP means a list of all the unique words occurring in the corpus. If some words are repeated in different documents, they are all written just once a while creating the dictionary.

What is term frequency?
Term frequency is the frequency of a word in one document. Term frequency can easily be found from the document vector table as in that table we mention the frequency of each word of the vocabulary in each document.

Important Questions

What is the difference between stemming and lemmatization?

Stemming is a technique used to extract the base form of the words by removing affixes from them. It is just like cutting down the branches of a tree to its stems. For example, the stem of the words eating, eats, eaten is eaten.

Lemmatization is the grouping together of different forms of the same word. In search queries, lemmatization allows end-users to query any version of a base word and get relevant results. ***(OR)***

Stemming is the process in which the affixes of words are removed and the words are converted to their base form.

In lemmatization, the word we get after affix removal (also known as lemma) is a meaningful one. Lemmatization makes sure that a lemma is a word with meaning and hence it takes a longer time to execute than stemming. ***(OR)***

Stemming algorithms work by cutting off the end of the beginning of the word, taking into account a list of common prefixes and suffixes that can be found in an inflected word.

Lemmatization, on the other hand, takes into consideration the morphological analysis of the words. To do so, it is necessary to have detailed dictionaries which the algorithm can look through to link the form back to its lemma.

Important Questions

What is a document vector table?

Document Vector Table is used while implementing the Bag of Words algorithm. In a document vector table, the header row contains the vocabulary of the corpus and other rows correspond to different documents. If the document contains a particular word it is represented by 1 and the absence of a word is represented by 0 value. ***(OR)***

Document Vector Table is a table containing the frequency of each word of the vocabulary in each document.

What do you mean by corpus?

In-Text Normalization, we undergo several steps to normalize the text to a lower level. That is, we will be working on text from multiple documents, and the term used for the whole-textual data from all the documents altogether is known as corpus. ***(OR)***

A corpus is a large and structured set of machine-readable texts that have been produced in a natural communicative setting. ***(OR)***

A corpus can be defined as a collection of text documents. It can be thought of as just a bunch of text files in a directory, often alongside many other directories of text files.

What are the types of data used for Natural Language Processing applications?

Natural Language Processing takes in the data of Natural Languages in the form of written words and spoken words which humans use in their daily lives and operate on this.

Differentiate between a script-bot and a smart-bot

Script Bot	Smart Bot
A scripted chatbot doesn't carry even a glimpse of A.I	Smart bots are built on NLP and ML.
Script bots are easy to make	Smart bots are comparatively difficult to make.
Script bot functioning is very limited as they are less powerful.	Smart bots are flexible and powerful.
Script bots work around a script that is programmed into them	Smart bots work on bigger databases and other resources directly
No or little language processing skills	NLP and Machine learning skills are required.
Limited functionality	Wide functionality

Give an example of the following:

A. Multiple meanings of a word *B. Perfect syntax, no meaning*

Example of Multiple meanings of a word –
His face turns red after consuming the medicine Meaning – Is he having an allergic reaction? Or is he not able to bear the taste of that medicine?

Example of Perfect syntax, no meaning-
Chickens feed extravagantly while the moon drinks tea. This statement is correct grammatically but it does not make any sense. In Human language, a perfect balance of syntax and semantics is important for better understanding.

What is inverse document frequency?

To understand inverse document frequency, first, we need to understand document frequency.

1. Document Frequency is the number of documents in which the word occurs irrespective of how many times it has occurred in those documents.
2. In the case of inverse document frequency, we need to put the document frequency in the denominator while the total number of documents is the numerator.
    - For example, if the document frequency of the word “AMAN” is 2 in a particular document then its inverse document frequency will be 3/2. (Here no. of documents is 3)

Mention some applications of Natural Language Processing.

Natural Language Processing Applications-

- Sentiment Analysis.
- Chatbots & Virtual Assistants.
- Text Classification.
- Text Extraction.
- Machine Translation
- Text Summarization
- Market Intelligence
- Auto-Correct

Explain the concept of Bag of Words.

1. Bag of Words is a Natural Language Processing model which helps in extracting features out of the text which can be helpful in machine learning algorithms.
2. In a bag of words, we get the occurrences of each word and construct the vocabulary for the corpus.
3. Bag of Words just creates a set of vectors containing the count of word occurrences in the document (reviews). Bag of Words vectors is easy to interpret.

What are the applications of TFIDF?

TFIDF is commonly used in the Natural Language Processing domain. Some of its applications are:

- Document Classification – Helps in classifying the type and genre of a document.
- Topic Modelling – It helps in predicting the topic for a corpus.
- Information Retrieval System – To extract the important information out of a corpus.
- Stop word filtering – Helps in removing the unnecessary words out of a text body.

What are stop words? Explain with the help of examples.

- “Stop words” are the most common words in a language like “the”, “a”, “on”, “is”, “all”. These words do not carry important meaning and are usually removed from texts.

- It is possible to remove stop words using Natural Language Toolkit (NLTK), a suite of libraries and programs for symbolic and statistical natural language processing.

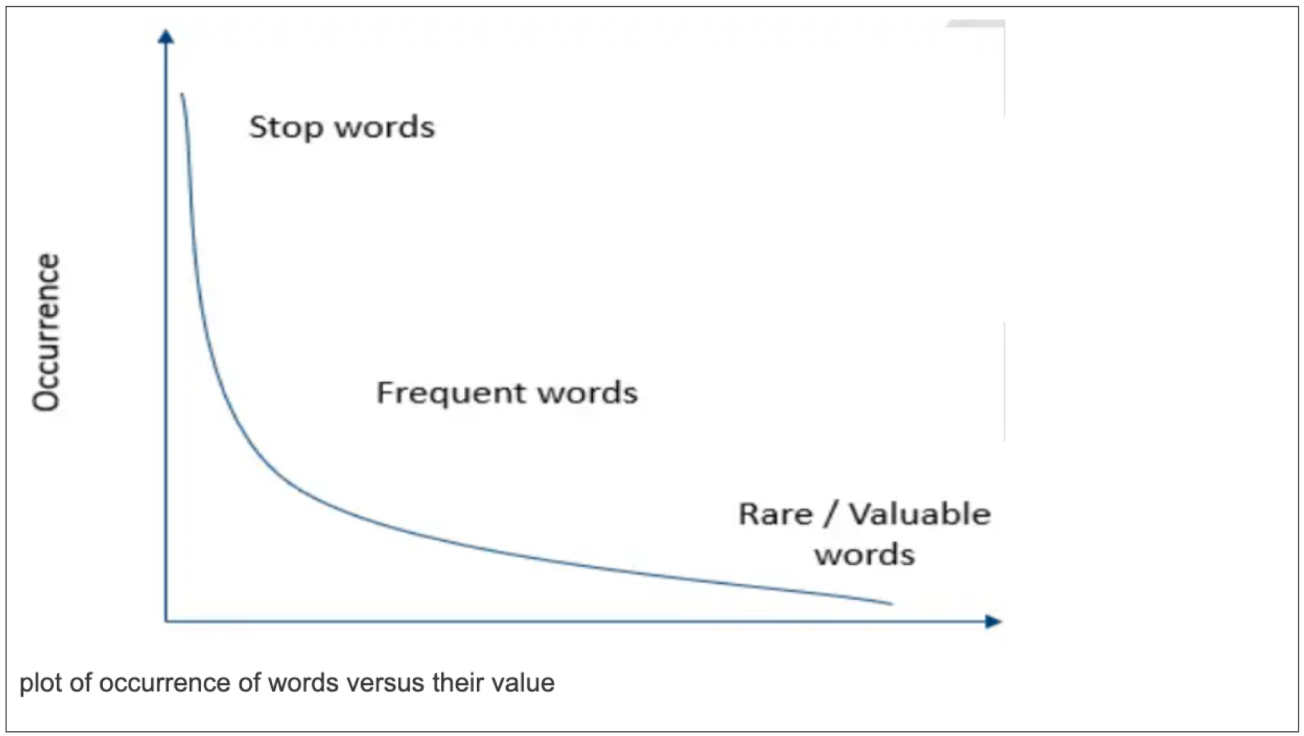
Differentiate between Human Language and Computer Language.

Human Language	Computer Language
Humans communicate through language which we process all the time.	The computer understands the language of numbers.
Our brain keeps on processing the sounds that it hears around itself and tries to make sense of them all the time.	Everything that is sent to the machine has to be converted to numbers.
–	While typing, if a single mistake is made, the computer throws an error and does not process that part.
–	The communications made by the machines are very basic and simple.

Explain the relation between occurrence and value of a word

As shown in the graph, the occurrence and value of a word are inversely proportional.

1. The words which occur most (like stop words) have negligible value.
2. As the occurrence of words drops, the value of such words rises. These words are termed rare or valuable words.
3. Rare words occur the least but add the most value to the corpus.



plot of occurrence of words versus their value