Unit 6: Natural Language Processing

Lesson Title: Natural Language Processing	**Approach:** Session + Activity

Summary: Students will be introduced to the NLP and its importance. They will receive an overview of the various stages of NLP and test processing techniques used in Natural Language Processing (NLP). They will undertake activities to appreciate the distinction between the Code vs No-code NLP. They will understand the Bag of Words algorithm andthe concept of TFIDF.

Learning Objectives:

Students are introduced to

1. NLP and its importance.
2. Various applications of NLP.
3. Different stages of NLP.
4. Various text processing techniques used in NLP.
5. Different No-Code NLP tools.
6. The Bag of Words model.
7. The concept of TFIDF.
8. No code Orange Data Mining Tool

Learning Outcomes:

Students will be able to:

1. Describe the importance of pre-processing NLP data.
2. Recognize the different steps of NLP data pre-processing.
3. Learners will be able to list several applications of NLP which can be implemented without code.
4. Enlist different No-Code NLP tools.
5. Outline the concept of the Bag of Words algorithm.
6. Explain the process of TFIDF.
7. Explain Sentiment Analysis

Pre-requisites: None

Key-concepts:

1. Natural Language Processing
2. Applications of NLP
3. Different stages of NLP
4. Text Processing techniques used in NLP
5. Bag of Word
6. TFIDF
7. No code – Orange Data Mining Tool

6.1 Introduction

A natural language is a human language, such as French, Spanish, English, Japanese, etc.

Features of Natural Languages

- They are governed by set rules that include syntax, lexicon, and semantics.
- All natural languages are redundant, i.e., the information can be conveyed in multiple ways.
- All natural languages change over time.

Test Yourself:

Choose the right word:

1. I am so tired; I want to take a_____?

2. Let's_____her a letter.

Do you see how same-sounding words can have totally different meanings?

- **Different meanings in different contexts.**

Let's consider these three sentences:

His face turned red after he found out that he took the wrong bag.

What does this mean? Is he feeling ashamed because he took another person's bag insteadof his? Is he feeling angry because he did not manage to steal the bag that he has been targeting?

The red car zoomed past his nose.

Probably talking about the colour of the car

His face turns red after consuming the medicine.

Is he having an allergic reaction? Or was he ashamed because he lost a bet ("I will not fall sick because of this")? Or was he taking a medicine that dilates the artery?

Here we can see that context is important. We understand a sentence almost intuitively, depending on our history of using the language, and the memories that have been built within. In all three sentences, the word red has been used in three different ways which according to the context of the statement changes its meaning completely. Thus, in natural language, it is important to understand that a word can have multiple meanings and the meanings fit into the statement according to the context of it.

Think of some other words which can have multiple meanings and use them in sentences.

_____

_____

_____

_____

_____

Computer Language

Computer languages are languages used to interact with a computer, such as Python, C++, Java, HTML, etc.

Can computers understand our language?

Computers require a specific set of instructions to understand human input called programs. To talk toa computer, we convert natural language into a language that a computer understands. We need Natural Language Processing to help computers understand natural language.

Why is NLP important?

Computers can only process electronic signals in the form of binary language. Natural Language Processing facilitates this conversion to digital form from the natural form. Thus, the whole purpose of NLP is to make communication between computer systems and humans possible. This includes creating different tools and techniques that facilitate better communication of intent and context.

Demystify Natural Language Processing (NLP)

Natural Language Processing, or NLP, is the sub-field of AI that is focused on enabling computers to analyse, understand and process human languages to derive meaningful information from human language.

6.2 Applications of Natural Language Processing

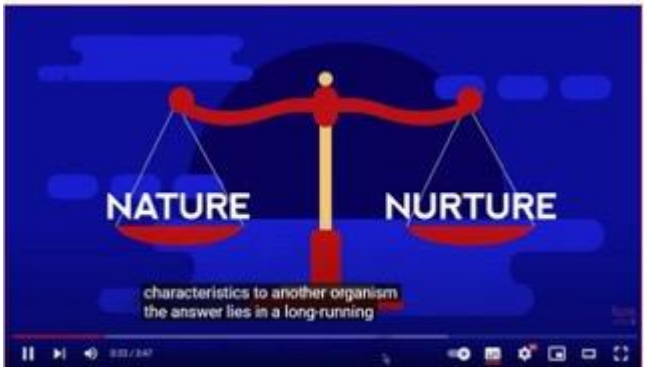
Since Artificial Intelligence nowadays is becoming an integral part of our lives, its applications are very commonly used by the majority of people in their daily lives. Here are some of the applications of Natural Language Processing which are used in the real-life scenario:

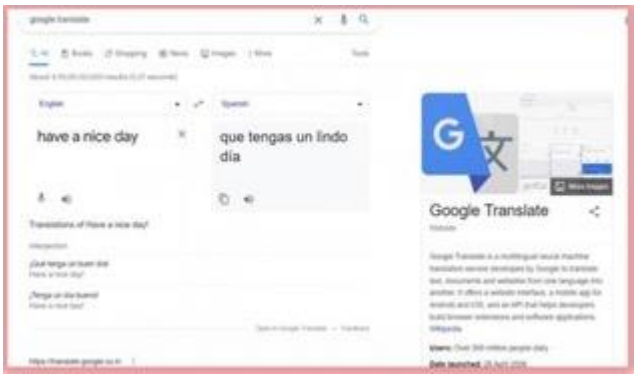
Voice assistants: Voice assistants take our natural speech, process it, and give us an output. These assistants leverage NLP to understand natural language and execute tasks efficiently.

For example:
Hey Google, set an alarm at 3.30 pm
Hey Alexa, play some music
Hey Siri, what's the weather today

Autogenerated captions: Captions are generated by turning natural speech into text in real-time. It is a valuable feature for enhancing the accessibility of video content.

For example:
Auto-generated captions on YouTube and Google Meet.

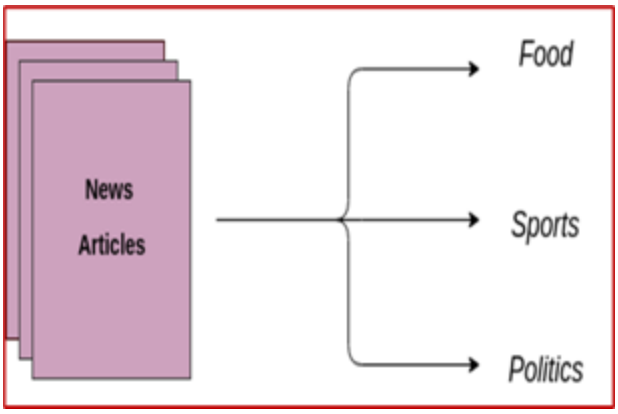




Language Translation: It incorporates the generation of translation from another language. This involves the conversion of text or speech from one language to another, facilitating cross-linguistic communication and fostering global connectivity.

For example:
Google Translate

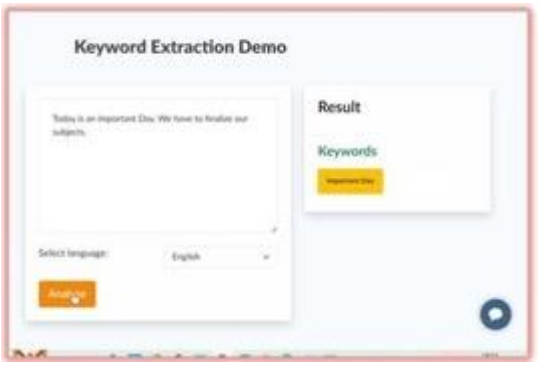
Sentiment Analysis: Sentiment Analysis is a tool to express an opinion, whether the underlying sentiment is positive, negative, or neutral. Customer sentiment analysis helps in the automatic detection of emotions when customers interact with the products, services, or brand



Text Classification: Text classification is a tool which classifies a sentence or document category-wise.

In the example, we can observe news articles containing information on various sectors, including Food, Sports, and Politics, being categorized through the text classification process. This process classifies the raw texts into predefined groups or categories.

Keyword Extraction: Keyword extraction is a tool that automatically extracts the most used, important words and expressions from a text. It can give valuable insights into people's opinions about any business on social media. Customer Service can be improved by using a Keyword extraction tool.

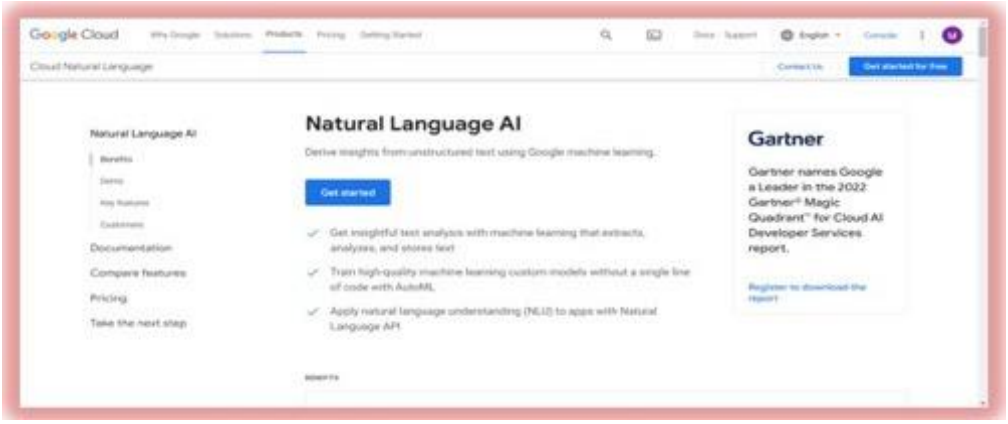


Activity 1: Keyword Extraction

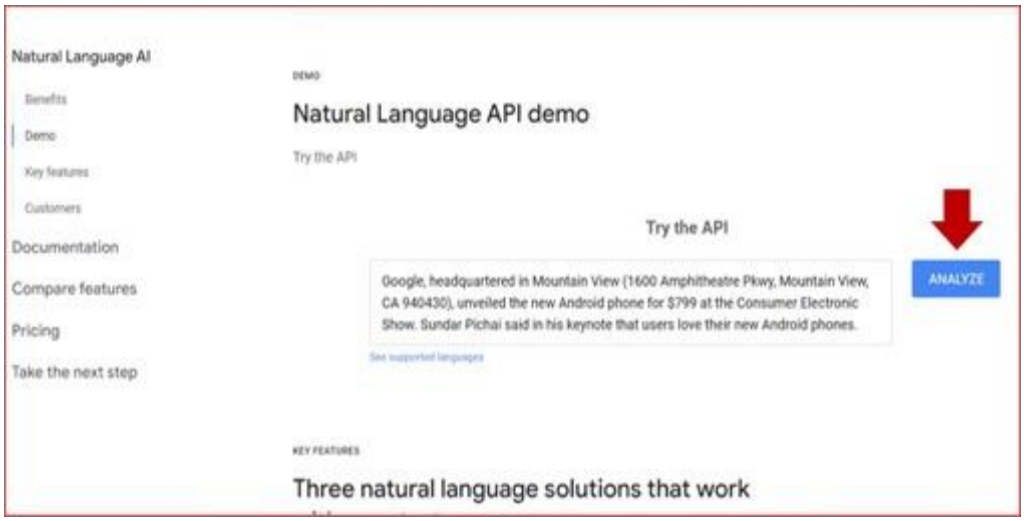
Purpose: To learn how to utilize an API for performing keyword extraction from a website.

Say: "Keyword extraction in NLP involves automatically identifying and extracting the most important words or phrases from a piece of text. These keywords represent the main topics or themes within the text and are useful for tasks like document summarization, information retrieval, and contentanalysis."

STEP – 1: Go to the given website: https://cloud.google.com/natural-language



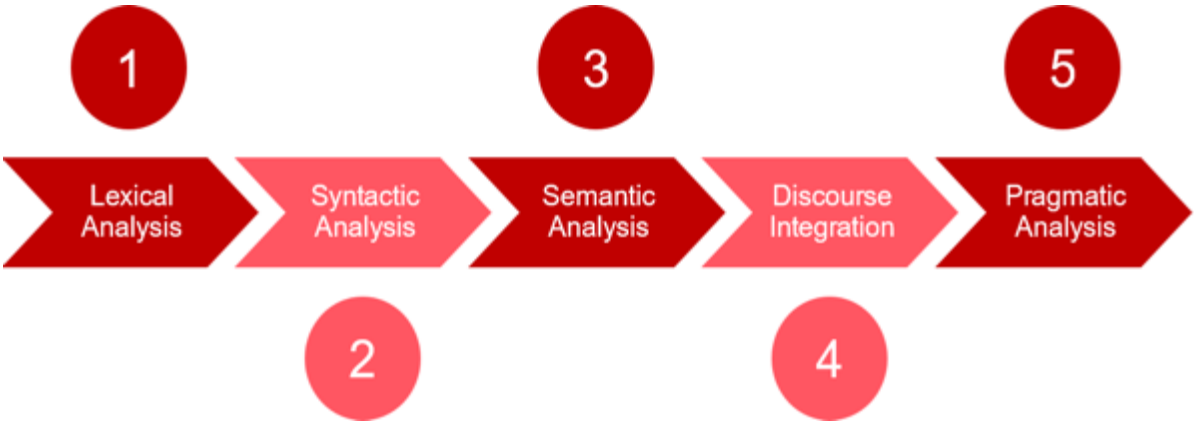
STEP – 2: Click on 'Analyze' and check the results.



- The keywords from the paragraph in the textbox have been highlighted in different colours e.g., Google, Mountain View, etc.
- Click on other options to check the output.
- Use your own text in the text box and observe the results.

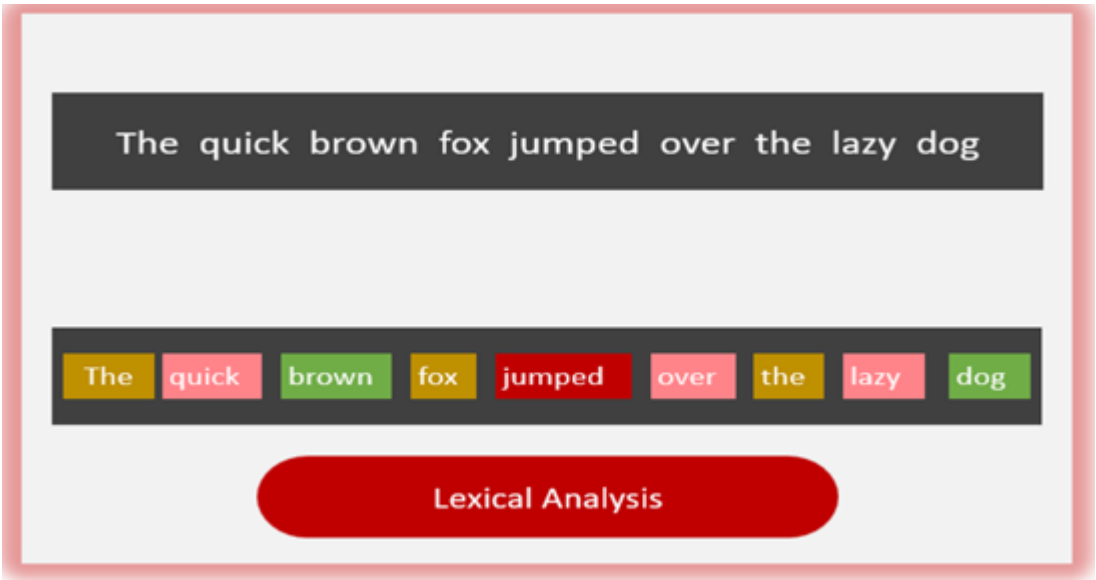
6.3 Stages of Natural Language Processing (NLP)

The different stages of Natural Language Processing (NLP) serve various purposes in theoverall task of understanding and processing human language. The stages of Natural Language Processing (NLP) typically involve the following:



Lexical Analysis:

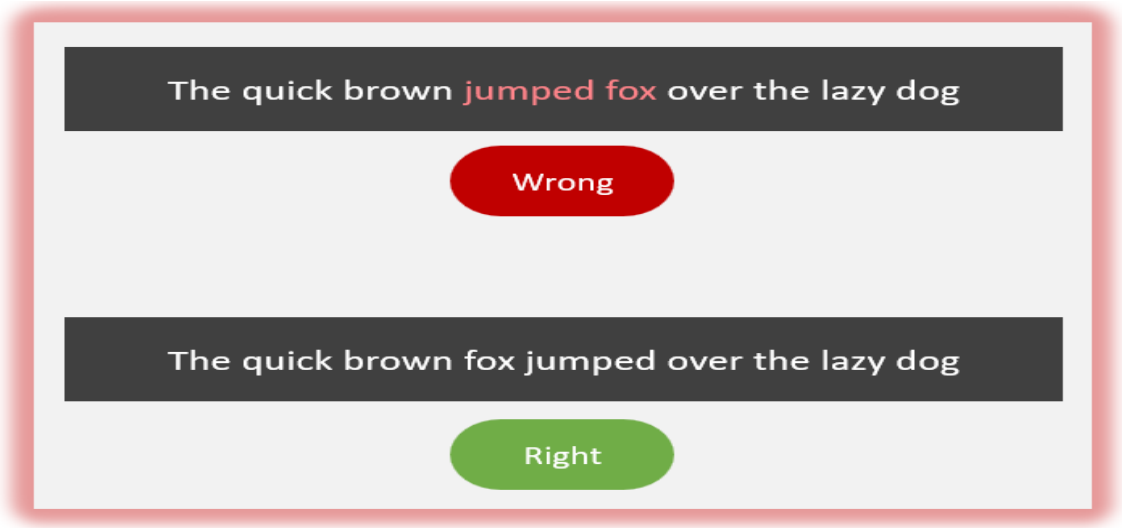
NLP starts with identifying the structure of input words. It is the process of dividing a large chunk of words into structural paragraphs, sentences, and words. Lexicon stands for a collection of the various words and phrases used in a language.



Lengthy text is broken down into chunks.

Syntactic Analysis / Parsing

It is the process of checking the grammar of sentences and phrases. It forms a relationship among words and eliminates logically incorrect sentences.



The grammar is correct!

Semantic Analysis

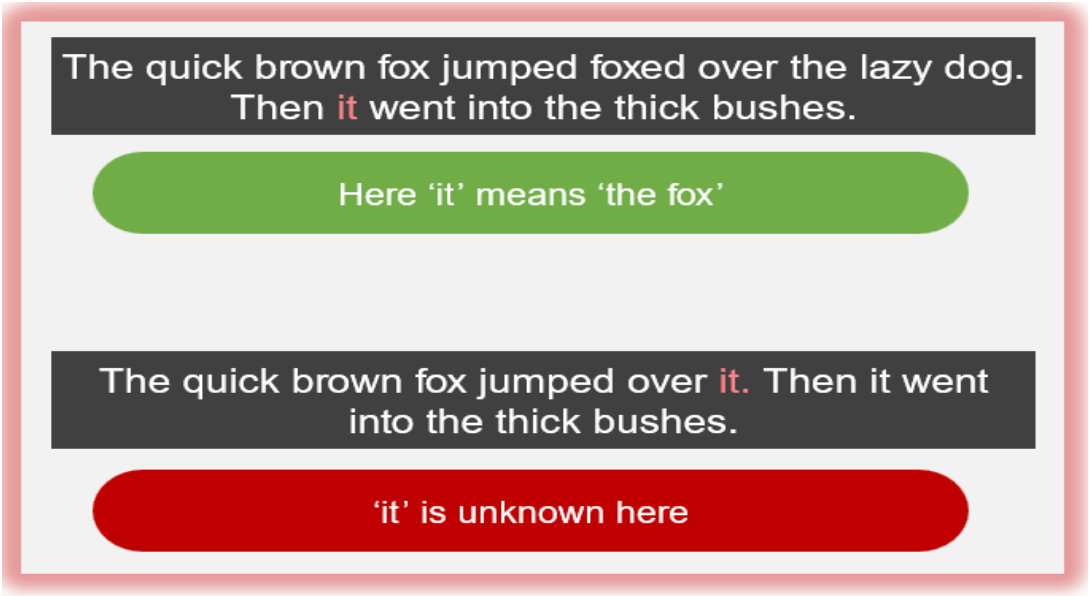
In this stage, the input text is now checked for meaning, and every word and phrase ischecked for meaningfulness.

For example:
It will reject a sentence that contains ‘hot ice cream’ in it.The
fox jumped into the dog.

Sentences make actual sense!

Discourse Integration

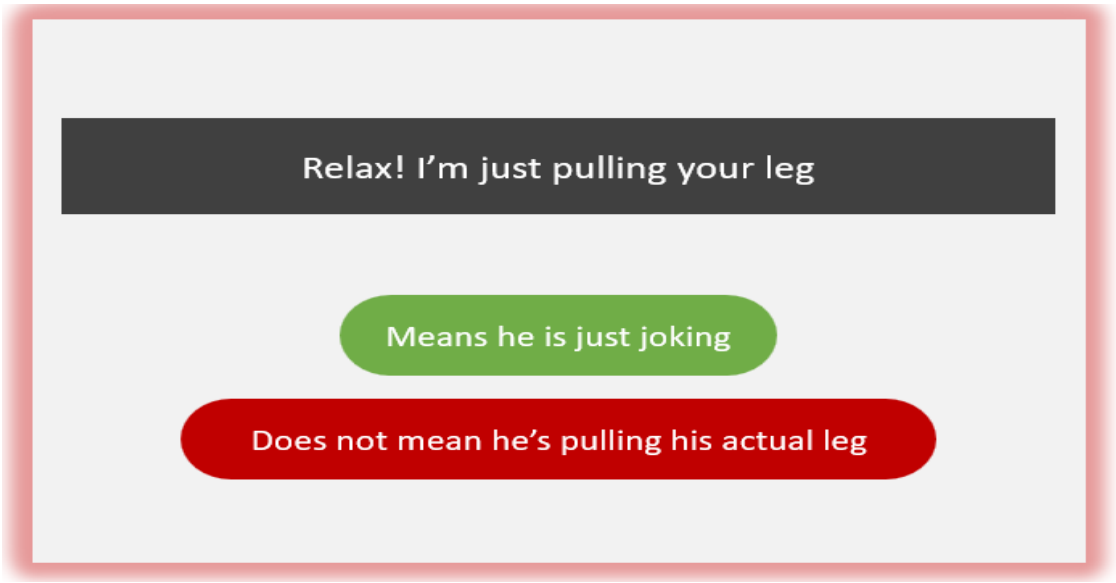
It is the process of forming the story of the sentence. Every sentence should have a relationship with its preceding and succeeding sentences.



The flow of words makes sense!

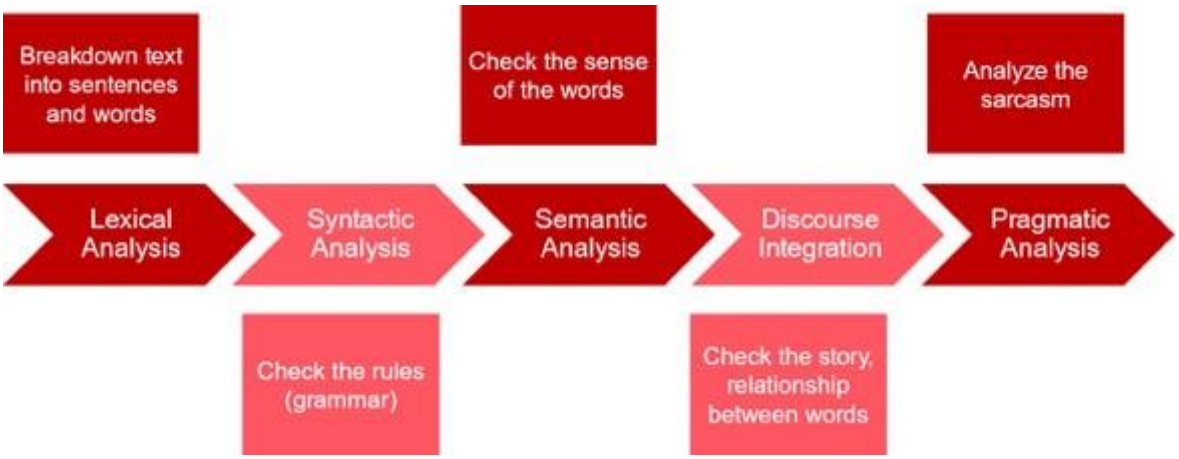
Pragmatic Analysis

In this stage, sentences are checked for their relevance in the real world. Pragmatic means practical or logical, i.e., this step requires knowledge of the intent in a sentence. It also means to discard the actual word meaning taken after semantic analysis and take theintended meaning.



The intended meaning has been achieved!

In summary,



Test Yourself:

Choose the right word:

1. Syntax refers to the grammatical structure of a sentence.

YES

NO

2. Which technique is used to assess the meaningfulness of the input text?

Pragmatic Analysis

Lexical Analysis

Semantic Analysis

Discourse Integration

6.4 Chatbots

Activity 2: Play with chatbots

Purpose: Identify and interact with different chatbots.

Say: “Have you ever wondered why chatbots are created? They are meant to make it feel as if we are talking to a real human as this is the way we are comfortable with. There are several chatbots here. I will split you into groups. Spend some time interacting with the chatbot, and then we will review our experience.”

One of the most common applications of Natural Language Processing is a chatbot. A chatbot is a computer program that's designed to simulate human conversation through voice commands or text chats or both. It can learn over time how to best interact with humans. It can answer questions and troubleshoot customer problems, evaluate and qualify prospects, generate sales leads and increase sales on an ecommerce site. There are a lot of chatbots available. Let us try some of the chatbots and see how they work.

ELIZA

Mitsuku

Cleverbot

Singtel

Elizabot - https://www.masswerk.at/elizabot/

Mitsuki - https://www.kuki.ai/

Cleverbot - https://www.cleverbot.com/

Singtel - https://www.singtel.com/personal/support

Let us discuss it!

• Which chatbot did you try? Name anyone.

• What is the purpose of this chatbot?

• How was the interaction with the chatbot?

• Did the chat feel like talking to a human or a robot? Why do you think so?

• Do you feel that the chatbot has a certain personality?

As you interact with more and more chatbots, you would realise that some of them are scripted or in other words are traditional chatbots while others are AI-powered and have 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 means in detail.

Script bot

- Script bots are easy to make.
- Script bots work around a script which is programmed in them.
- Mostly they are free and are easy to integrate to a messaging platform.
- No or little language processing skills.
- Limited functionality.

Smart bot

- Smart-bots are flexible and powerful.
- Smart bots work on bigger databases and other resources directly.
- Smart bots learn with more data.
- Coding is required to take this up on board.
- Wide functionality.

Quiz Time

1. Natural Language Processing majorly deals with_____ processing.

a. Numeric data

b. Textual data

c. Image data

d. Visual data

2. _____is an NLP tool to express an opinion, whether the underlying sentiment is positive, negative, or neutral.

a. Text Classification

b. Machine Translation

c. Sentiment Analysis

d. Automatic Text Summarization

3. What is the first stage of Natural Language Processing (NLP)?

a. Semantic Analysis

b. Pragmatic Analysis

c. Lexical Analysis

d. Syntactic Analysis

4. Words that we want to filter out before doing any analysis of the text are called_____.

a. Rare words

b. Stop words

c. Frequent words

d. Filter words

5. What does discourse integration involve in the context of sentence formation?

a. Identifying individual words in a sentence

b. Forming a coherent story within a sentence

c. Establishing relationships between preceding and succeeding sentences

d. Applying punctuation and grammar rules to a sentence

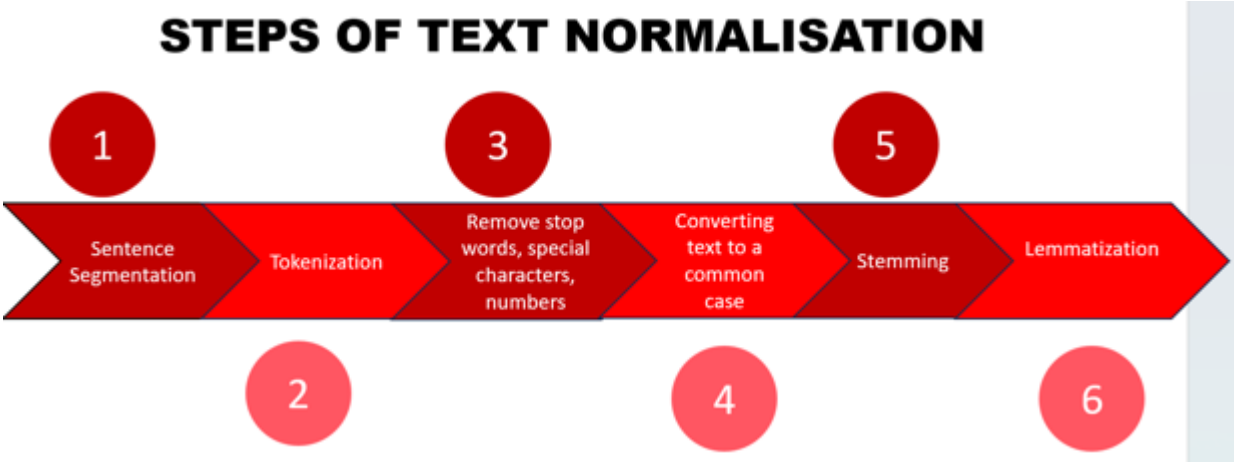
6.5 Text Processing

Humans interact with each other very easily. For us, the natural languages that we use are so convenient that we speak them easily and understand them well too. But for computers, our languages are very complex. As you have already gone through some of the complications in human languages above, now it is time to see how Natural Language Processing makes it possible for machines to understand and speak in Natural Languages just like humans.

Since we all know that the language of computers is Numerical, the very first step that comes to our mind is to convert our language to numbers. This conversion takes a few steps to happen. The first step to it is Text Normalisation. Since human languages are complex, we need to first of all simplify them in order to make sure that understanding becomes possible. Text Normalisation helps in cleaning up the textual data in such a way that it comes down to a level where its complexity is lower than the actual data. Let us go through Text Normalisation in detail.

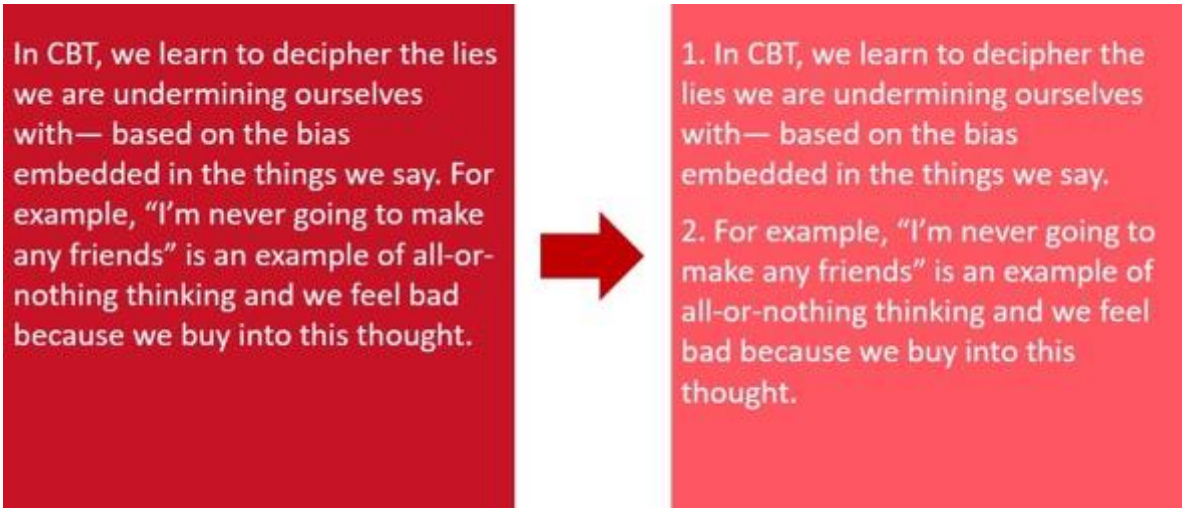
Text Normalisation

In Text Normalisation, we undergo several steps to normalise the text to a lower level. Before we begin, we need to understand that in this section, we will be working on a collection of written text. That is, we will be working on text from multiple documents andthe term used for the whole textual data from all the documents altogether is known as corpus. Not only would we go through all the steps of Text Normalisation, we would also work them out on a corpus. Let us take a look at the steps:



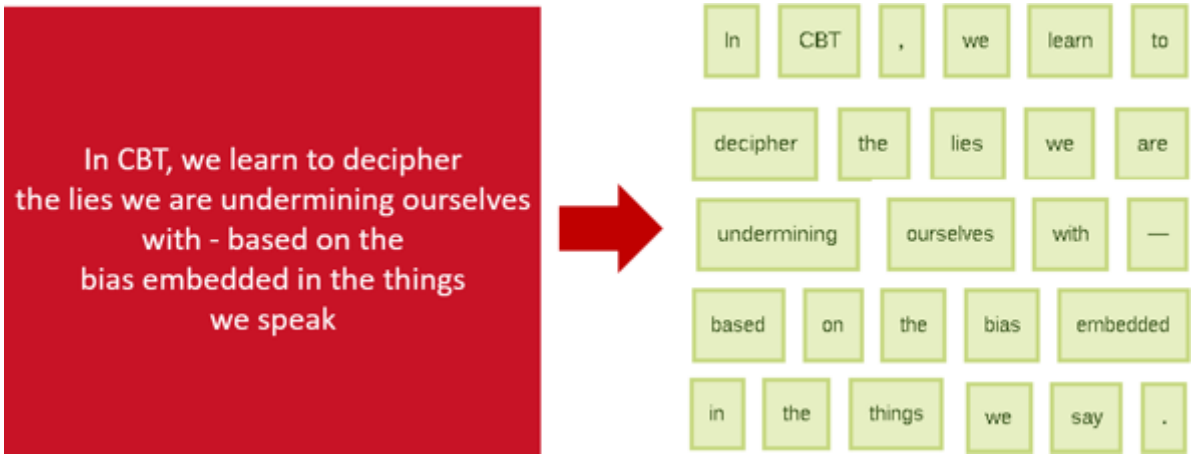
Sentence Segmentation

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.



Tokenization

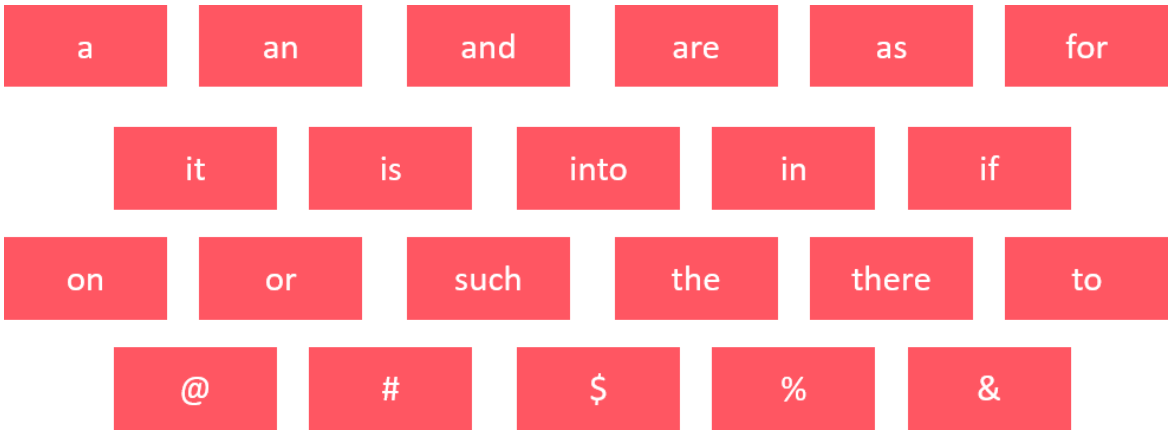
After segmenting the sentences, each sentence is then further divided into tokens. Tokens is a term used for any word or number or special character occurring in a sentence. Under tokenisation, every word, number and special character is considered separately and each of them is now a separate token.



Removing Stop words, Special Characters and Numbers

In this step, the tokens which are not necessary are removed from the token list. What arethe possible words which we might not require?

Stop words are the words which occur very frequently in the corpus but do not add any value to it. Humans use grammar to make their sentences meaningful for the other person to understand. But grammatical words do not add any essence to the information which is tobe transmitted through the statement hence they come under stop words. Some examples of stop words are:

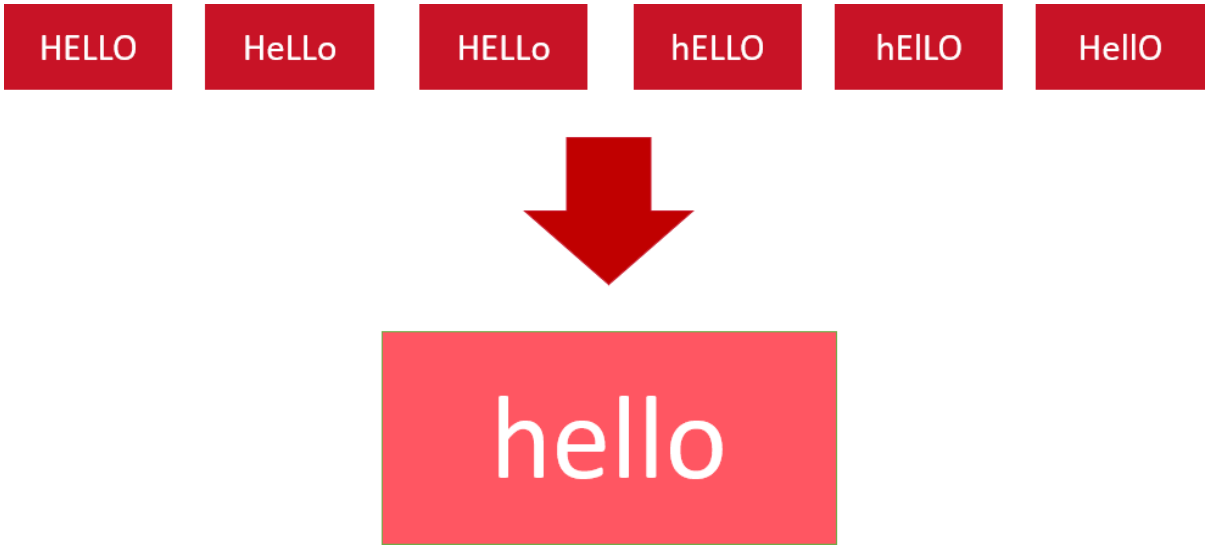


These words occur the most in any given corpus but talk very little or nothing about the context or the meaning of it. Hence, to make it easier for the computer to focus on meaningful terms, these words are removed.

Along with these words, a lot of times our corpus might have special characters and/or numbers. Now it depends on the type of corpus that we are working on whether we should keep them in it or not. For example, if you are working on a document containing email IDs, then you might not want to remove the special characters and numbers whereas in some other textual data if these characters do not make sense, then you can remove them along with the stop words.

Converting Text to a Common Case

After the stop words removal, we convert the whole text into a similar case, preferably lowercase. This ensures that the case sensitivity of the machine does not consider the same words as different just because of different cases.



Here in this example, all the 6 forms of hello would be converted to lowercase and hence would be treated as the same word by the machine.

Stemming

In this step, the remaining words are reduced to their root words. In other words, stemmingis the process in which the affixes of words are removed and the words are converted to their base form.

Word	Affixes	Stem
healed	-ed	heal
healing	-ing	heal
healer	-er	heal
studies	-es	studi
studying	-ing	study

Note that in stemming, the stemmed words (words that we get after removing the affixes) might not be meaningful. Here in this example as you can see: healed, healing and healer all were reduced to heal but studies was reduced to studi after the affix removal which is not a meaningful word. Stemming does not take into account whether the stemmed word is meaningful or not. It just removes the affixes hence it is faster.

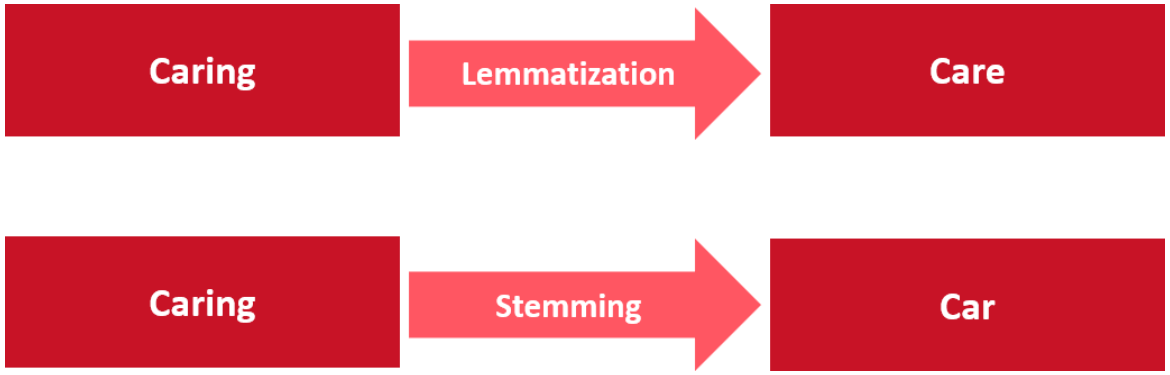
Lemmatization

Stemming and lemmatization both are alternative processes to each other as the role of both the processes is same – removal of affixes. But the difference between both of them isthat 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.

Word	Affixes	Stem
healed	-ed	heal
healing	-ing	heal
healer	-er	heal
studies	-es	**study**
studying	-ing	study

As you can see in the same example, the output for studies after affix removal has become study instead of studi.

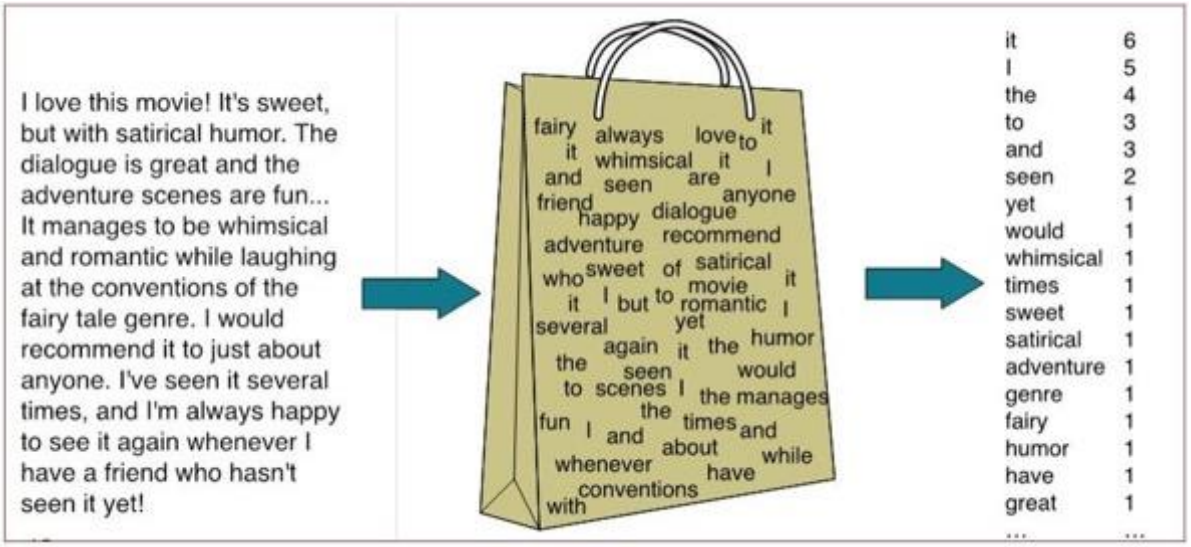
The difference between stemming and lemmatization can be summarized by this example:



With this, we have normalised our text to tokens which are the simplest form of words present in the corpus. Now it is time to convert the tokens into numbers. For this, we would use the Bag of Words algorithm

Bag of Words

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. In the bag of words, we get the occurrences of each word and construct the vocabulary for the corpus.



This image gives us a brief overview of how the bag of words works. Let us assume that the text on the left in this image is the normalised corpus which we have got after going through all the steps of text processing. Now, as we put this text into the bag of words algorithm, the algorithm returns to us the unique words out of the corpus and their occurrences in it. As you can see on the right, it shows us a list of words appearing in the corpus and the numbers corresponding to it show how many times the word has occurred in the text body. Thus, we can say that the bag of words gives us two things:

1. A vocabulary of words for the corpus
2. The frequency of these words (number of times it has occurred in the whole corpus).

Here calling this algorithm a “bag” of words symbolises that the sequence of sentences or tokens does not matter. In this case, all we need are the unique words and their frequency.

Here is the step-by-step approach to implementing the bag of words algorithm:

1. Text Processing: Collect data and pre-process it

2. Create a 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.

Let us go through all the steps with an example:

Step 1: Collecting data and pre-processing it.

Document 1: Aman and Avni are stressed

Document 2: Aman went to a therapist

Document 3: Avni went to download a health chatbot

Here are three documents having one sentence each. After text normalisation, the text becomes:

Document 1: [aman, and, avni, are, stressed]

Document 2: [aman, went, to, a, therapist]

Document 3: [avni, went, to, download, a, health, chatbot]

Step 2: Create a Dictionary

Go through all the steps and create a dictionary i.e., list down all the words which occur in all three documents:

Dictionary:

aman	and	avni	are	stressed	went
download	health	chatbot	therapist	a	to

Note that even though some words are repeated in different documents, they are all written just once as while creating the dictionary, we create the list of unique words.

Step 3: Create a document vector

In this step, the vocabulary is written in the top row. Now, for each word in the document, ifit matches the vocabulary, put a 1 under it. If the same word appears again, increment the previous value by 1. And if the word does not occur in that document, put a 0 under it.

aman	and	avni	are	stressed	went	to	a	therapist	download	health	chatbot
`1	1	1	1	1	0	0	0	0	0	0	0

Since, in the first document, we have words: aman, and, avni, are, stressed. So, all thesewords get a value of 1 and the rest of the words get a 0 value.

Step 4: Create document vectors for all the documents.

The same exercise has to be done for all the documents. Hence, the table becomes:

aman	and	avni	are	stressed	went	to	a	therapist	download	health	chatbot
`1	1	1	1	1	0	0	0	0	0	0	0
1	0	0	0	0	1	1	1	1	0	0	0
0	0	1	0	0	1	1	1	0	1	1	1

In this table, the header row contains the vocabulary of the corpus and three rows correspond to three different documents. Take a look at this table and analyse the positioning of 0s and 1s in it.

Finally, this gives us the **document vector table** for our corpus. However, the tokens havestill not been converted to numbers. This leads us to the final step of our algorithm: TFIDF.

TFIDF: Term Frequency & Inverse Document Frequency

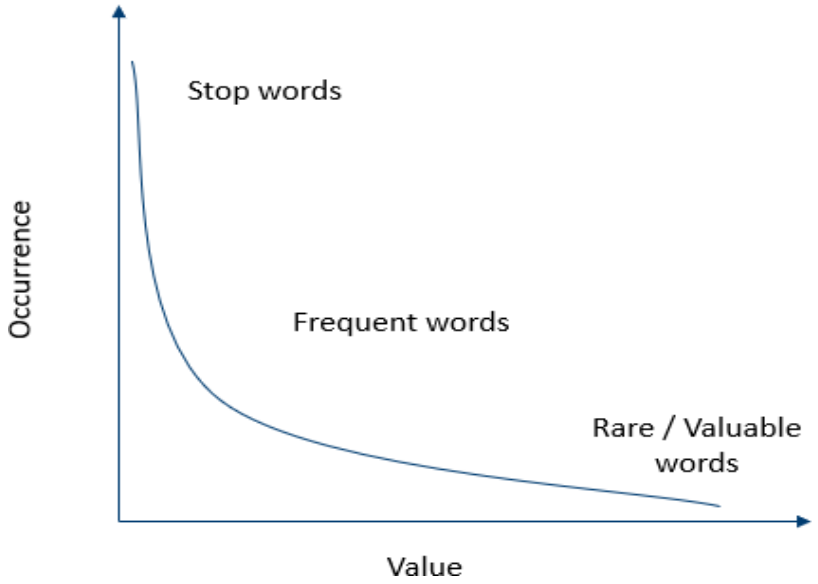
Suppose you have a book. Which characters or words do you think would occur the most init?

_____

_____

The bag of words algorithm gives us the frequency of words in each document we have in our corpus. It gives us an idea that if the word is occurring more in a document, its value is more for that document. For example, if I have a document on air pollution, air and pollutionwould be the words which occur many times in it. And these words are valuable too as they give us some context around the document. But let us suppose we have 10 documents and all of them talk about different issues. One is on women's empowerment; the other is on unemployment and so on. Do you think air and pollution would still be one of the most occurring words in the whole corpus? If not, then which words do you think would have the highest frequency in all of them?

And, this, is, the, etc. are the words which occur the most in almost all the documents. But these words do not talk about the corpus at all. Though they are important for humans as they make the statements understandable to us, for the machine they are a complete waste as they do not provide us with any information regarding the corpus. Hence, these are termed as stop words and are mostly removed at the pre-processing stage only.



Take a look at this graph. It is a plot of the occurrence of words versus their value. As you can see, if the words have the highest occurrence in all the documents of the corpus, they are said to have negligible value hence they are termed as stop words. These words are mostly removed at the pre-processing stage only. Now as we move ahead from the stop words, the occurrence level drops drastically and the words which have adequate occurrence in the corpus are said to have some amount of value and are termed as frequent words. These words mostly talk about the document's subject and their occurrence is adequate in the corpus. Then as the occurrence of words drops further, the value of such words rises. These words are termed as rare or valuable words. These words occur the least but add the most value to the corpus. Hence, when we look at the text, we consider frequent and rare words.

Let us now demystify TFIDF. TFIDF stands for Term Frequency and Inverse Document Frequency. TFIDF helps us identify the value of each word. Let us understand each term one by one.

Term Frequency

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

aman	and	avni	are	stressed	went	to	a	therapist	download	health	chatbot
`1	1	1	1	1	0	0	0	0	0	0	0
1	0	0	0	0	1	1	1	1	0	0	0
0	0	1	0	0	1	1	1	0	1	1	1

Here, you can see that the frequency of each word for each document has been recorded in the table. These numbers are nothing but the Term Frequencies!

Inverse Document Frequency

Now, let us look at the other half of TFIDF which is Inverse Document Frequency. For this, let us first understand what document frequency means. Document Frequency is the number of documents in which the word occurs irrespective of how many times it has occurred in those documents. The document frequency for the exemplar vocabulary would be:

aman	and	avni	are	stressed	went	to	a	therapist	download	health	chatbot
`2	1	2	1	1	2	2	2	1	1	1	1

Here, you can see that the document frequency of ‘aman’, ‘avni’, ‘went’, ‘to’ and ‘a’ is 2 asthey have occurred in two documents. The rest of them occurred in just one document hence the document frequency for them is one.

Talking about inverse document frequency, we need to put the document frequency in the denominator while the total number of documents is the numerator. Here, the total number of documents is 3, hence inverse document frequency becomes:

aman	and	avni	are	stressed	went	to	a	therapist	download	health	chatbot
`3/2	3/1	3/2	3/1	3/1	3/2	3/2	3/2	3/1	3/1	3/1	3/1

Finally, the formula of TFIDF for any word W becomes:

TFIDF(W) = TF(W) * log(IDF(W))

Here, log is to the base of 10. Don’t worry! You don’t need to calculate the log values by yourself. Simply use the log function in the calculator and find out!

Now, let’s multiply the IDF values by the TF values. Note that the TF values are for each document while the IDF values are for the whole corpus. Hence, we need to multiply the IDF values to each row of the document vector table.

aman	and	avni	are	stressed	went	to	a	therapist	download	health	chatbot
`1*log(3/2)	1*log(3)	1*log(3/2)	1*log(3)	1*log(3)	0*log(3/2)	0*log(3/2)	0*log(3/2)	0*log(3)	0*log(3)	0*log(3)	0*log(3)
1*log(3/2)	0*log(3)	0*log(3/2)	0*log(3)	0*log(3)	1*log(3/2)	1*log(3/2)	1*log(3/2)	1*log(3)	0*log(3)	0*log(3)	0*log(3)
0*log(3/2)	0*log(3)	1*log(3/2)	0*log(3)	0*log(3)	1*log(3/2)	1*log(3/2)	1*log(3/2)	0*log(3)	1*log(3)	1*log(3)	1*log(3)

Here, you can see that the IDF values for Aman in each row are the same and a similar pattern is followed for all the words of the vocabulary. After calculating all the values, weget:

aman	and	avni	are	stressed	went	to	a	therapist	download	health	chatbot
`0.176	0.477	0.176	0.477	0.477	0	0	0	0	0	0	0
0.176	0	0	0	0	0.176	0.176	0.176	0.477	0	0	0
0	0	0.176	0	0	0.176	0.176	0.176	0	0.477	0.477	0.477

Finally, the words have been converted to numbers. These numbers are the values of each for each document. Here, you can see that since we have less amount of data, words like ‘are’ and ‘and’ also have a high value. But as the IDF value increases, the value of that word decreases. That is, for example:

Total Number of documents: 10

Number of documents in which ‘and’ occurs: 10

Therefore, IDF(and) = 10/10 = 1

Which means: log(1) = 0. Hence, the value of ‘and’ becomes 0.

On the other hand, the number of documents in which ‘pollution’ occurs: 3 IDF(pollution) = 10/3 = 3.3333…

This means log(3.3333) = 0.522; which shows that the word ‘pollution’ has considerablevalue in the corpus.

Summarising the concept, we can say that:

1. Words that occur in all the documents with high term frequencies have the lowest values and are considered to be the stop words.

2. For a word to have a high TFIDF value, the word needs to have a high term frequency but less document frequency which shows that the word is important for one document but is not a common word for all documents.

3. These values help the computer understand which words are to be considered while processing the natural language. The higher the value, the more important the word is for a given corpus.

Applications of TFIDF

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

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 unnecessary words from a text body.**

6.6 Natural Language Processing: Use Case Walkthrough

Purpose: Students are introduced to the No-code tools for Natural Language Processing.

They will learn about Sentiment Analysis, one of the applications of NLP with the No-code tool Orange Data Mining. Learners will be able to understand this application with use cases

Examples of Code and No-code NLP Tools

Code NLP	No-Code NLP
NLTK package: Natural Language Tool Kit or NLTK is a package readily available for text processing in Python. The package contains functions and modules which can be used for Natural Language Processing.	**Orange Data Mining:** It is a machine learning tool for data analysis through Python and visual programming. We can perform operations on data through simple drag-and-drop steps.
SpaCy: SpaCy is an open-source natural language processing (NLP) library designed to build NLP applications. It offers various features such as tokenization, part-of-speech tagging, named entity recognition, dependency parsing, and more.	**MonkeyLearn:** MonkeyLearn is a text analysis platform that offers NLP tools and machine learning models for text analysis, supporting tasks such as classification, sentiment analysis, and entity recognition. Users can create custom models or use pre-trained ones for tasks like social media monitoring and customer feedback analysis.

Applications of NLP

Introduction to Sentiment Analysis

Sentiment (opinion)

Analysis

Sentiment analysis is a natural language processing (NLP) technique used to analyzewhether a given textual data is positive, negative, or neutral.

Applications of Sentiment Analysis-Customer Service

Collecting customer data through live chat

Customer feedback about service

Customer sentiment analysis helps in the automatic detection of emotions when customers interact with products, services, or brands.

Applications of Sentiment Analysis –Voice of the Customer

- Voice of the customer analysis helps to analyze customer feedback and gain actionable insights from it.
- It measures the gap between what customers expect and what they actually experience when they use the products or services,

Now, let's understand Sentiment Analysis in detail using the Orange Data Mining tool.

Follow the given link to understand the technique.

Case Walkthrough – Steps involved in project development

Short Link - https://bit.ly/OrangeNLP

Long Link - https://drive.google.com/drive/u/2/folders/1geFLXxV5890kfcakMfEg_KsH1LPcS_Iz

Or Scan the QR code provided below.

Test Yourself:

1. What is the primary challenge faced by computers in understanding human languages?

A) Complexity of human languages

B) Lack of computational power

C) Incompatibility with numerical data

D) Limited vocabulary

2. How do voice assistants utilize NLP?

A) To analyze visual data

B) To process numerical data

C) To understand natural language

D) To execute tasks based on computer code

3. Which of the following is NOT a step in Text Normalisation?

A) Tokenization

B) Lemmatization

C) Punctuation removal

D) Document summarization

4. In the context of text processing, what is the purpose of tokenisation?

A) To convert text into numerical data

B) To segment sentences into smaller units

C) To translate text into multiple languages

D) To summarize documents for analysis

5. What distinguishes lemmatization from stemming?

A) Lemmatization produces meaningful words after affix removal, while stemming does not.

B) Lemmatization is faster than stemming.

C) Stemming ensures the accuracy of the final word.

D) Stemming generates shorter words compared to lemmatization.

6. What is the primary purpose of the Bag of Words model in Natural Language Processing?

A) To translate text into multiple languages

B) To extract features from text for machine learning algorithms

C) To summarize documents for analysis

D) To remove punctuation marks from text

7. In the context of text processing, what are stop words?

A) Words with the frequent occurrence in the corpus

B) Words with negligible value that are often removed during preprocessing

C) Words with the lowest occurrence in the corpus

D) Words with the most value added to the corpus

8. What is the characteristic of rare or valuable words in the described plot?

A) They have the highest occurrence in the corpus

B) They are often considered stop words

C) They occur the least but add the most value to the corpus

D) They are typically removed during preprocessing

9. What information does the document vector table provide?

A) The frequency of each word across all documents

B) The frequency of each word in a single document

C) The total number of words in the entire corpus

D) The average word length in the entire corpus

10. What is the primary purpose of TFIDF in text processing?

A) To identify the presence of stop words in documents

B) To remove punctuation marks from text

C) To identify the value of each word in a document

D) To translate text into multiple languages

11. Assertion: Pragmatic analysis in natural language processing (NLP) involves assessing sentences for their practical applicability in real-world scenarios.

Reasoning: Pragmatic analysis requires understanding the intended meaning behind sentences and considering their practical or logical implications, rather than solely relying on literal word meanings obtained from semantic analysis.

A) Both Assertion and Reasoning are true, and Reasoning is the correct explanation of the Assertion.

B) Assertion is true, but Reasoning is false.

C) Both Assertion and Reasoning are true, but Reasoning is not the correct explanation of the Assertion.

D) Assertion is false, but Reasoning is true.

12. Assertion: Converting the entire text into lowercase following stop word removal is a crucial preprocessing step in natural language processing.

Reasoning: This process ensures uniformity in word representation, preventing the machine from treating words with different cases as distinct entities, thereby enhancing the accuracy of subsequent text analysis.

A) Both Assertion and Reasoning are true, and Reasoning is the correct explanation of the Assertion.

B) Assertion is true, but Reasoning is false.

C) Both Assertion and Reasoning are true, but Reasoning is not the correct explanation of the Assertion.

D) Assertion is false, but Reasoning is true.

Reflection Time:

1. Mention a few features of natural languages.

2. What is the significance of NLP?

3. What do you mean by lexical analysis in NLP?

4. What do you mean by a chatbot?

5. What does the term "Bag of Words" refer to in Natural Language Processing (NLP)?

6. Describe two practical uses of Natural Language Processing in real-world scenarios.

7. Explain the process of stemming and lemmatization in text processing, supported by an example.

8. Describe any four applications of TFIDF.

9. Samiksha, a student of class X was exploring the Natural Language Processing domain. She

got stuck while performing the text normalisation. Help her to normalise 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.

10. Through a step-by-step process, calculate TFIDF for the given corpus

Document 1: Johny Johny Yes Papa,

Document 2: Eating sugar? No Papa

Document 3: Telling lies? No Papa

Document 4: Open your mouth, Ha! Ha! Ha!

CLASS – X ANSWER KEY (Unit-End Exercise)

UNIT – 1: Ethical Frameworks for AI

TEST YOURSELF

1. B 2. C 3. B 4. D 5. C 6. B 7. B 8. C 9. A 10. B 11. A 12. A

UNIT – 2: Advanced Concepts of Modelling in AI

TEST YOURSELF

1. A 2.A 3. B 4. C 5. C 6. B 7. D 8. B 9. C 10. A 11. B 12. B

UNIT – 3: Evaluating Models

TEST YOURSELF

1. B 2.A 3. A 4. B 5. A 6. B 7. A 8. B 9. B 10. C

UNIT – 4.1: Statistical Data

TEST YOURSELF

1. A 2. D 3. D 4. A 5. A

UNIT – 4.2: Statistical Data

TEST YOURSELF

1. B 2. B 3. A 4. C 5. A

UNIT – 5: Computer Vision

TEST YOURSELF

1. B 2. C 3. A 4. C 5. C 6. A 7. B 8. D 9. B 10. B 11. A 12. C

UNIT – 6: Natural Language Processing

Test Yourself

1. A 2. C 3. D 4. B 5. A 6. B 7. B 8. C 9. A 10. C 11. A 12. A