UNIT-V

Natural Language processing:

Role of knowledge in language understanding:

* An Intelligent agent needs knowledge about the real world for taking decisions and reasoning to act efficiently.

* knowledge-based agents are those agents who have the capability of maintaining an internal state of knowledge, reason over that knowledge, update their knowledge after observations and take actions.

+ These agents can Represent the world with some formal representation and act Intelligentely.

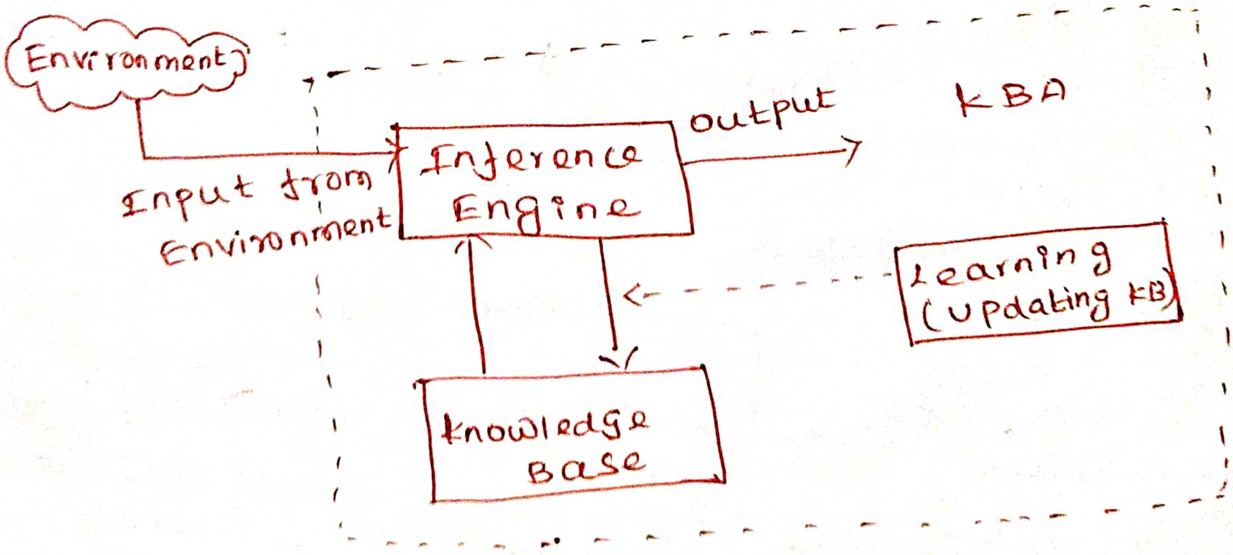
* knowledge-based agents are composed of two main parts:

  + knowledge-base and
  + Inference system.

* A knowledge-based agent must able to do the following:

- An Agent should be able to represent states, actions, etc.
- An agent should be able to Incorporate new percepts.
- An agent can update The Internal representation of the world.
- An agent can deduce the Internal representation of the world.
- An agent can deduce appropriate actions.

The architecture of knowledge-based agent:



* The above diagram is representing a generalized architecture for a knowledge- Based agent.

+ The knowledge based agent [KBA] take Input from the environment by perceiving the environment. The Input is taken by the Inference engine of the agent and which also communicate with KB to decide as per the knowledge store in KB.

* The Learning element of KBA Regularly updates the KB by Learning new Knowledge

Knowledge base: Knowledge-base is a central component of a knowledge -based agent. It is also known as KB.

* It is a collection of sentences (here 'sentence' is a technical term and It is not identical to sentence in English).

* These sentences are expressed in a language which is called a knowledge representation language.

* The knowledge-base of KBA stores fact about the world.

Why use a knowledge base?

* Knowledge-base is required for updating knowledge for an agent to learn with experiences and take action as per the knowledge.

Inference system:

* Inference means deriving new sentences from old. Inference system allows us to add a new sentence to the knowledge base.

* A sentence is a proposition about the world. Inference system applies logical rules to the KB to deduce new information.

* Inference system generates new facts So that an agent can update the KB.

* An Inference system works mainly in Two Rules which are given as:

- Forward chaining
- Backward chaining.

operations performed by KBA;

* Following are three operations which are performed by KBA in order to show the Intelligent behavior.

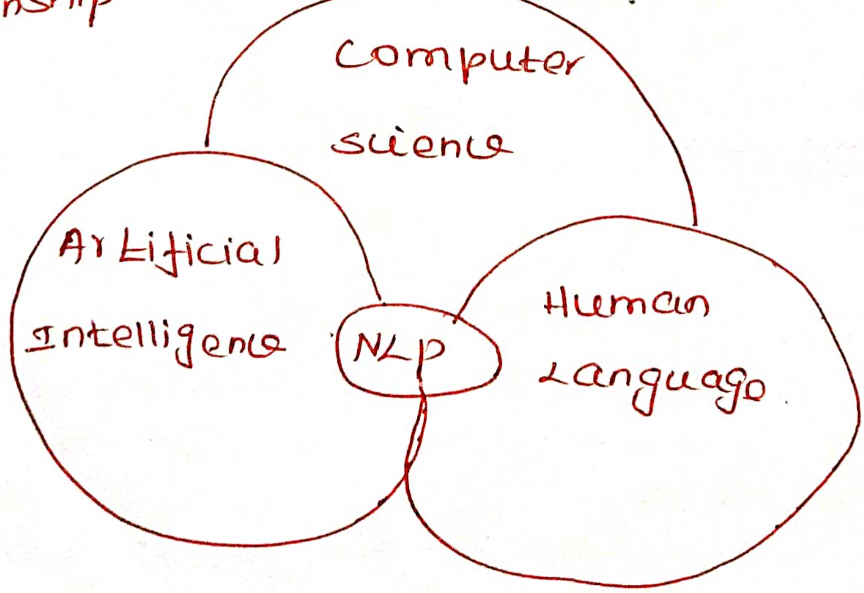
- TELL: This operation Tells the knowledge base what it perceives from the environment.
- ASK: This operation asks the knowledge base

Natural Language Processing

* NLP stands for Natural language processing, which is a part of computer Science, Human language, and Artificial Intelligence.

* It is the Technology that is used by machines to understand, analyse, manipulate and interpret human's languages.

* It helps developers to organize knowledge for performing tasks such as Translation, automatic Summarization, Named Entity Recognition [NER], speech recognition, Relationship extraction, and topic segmentation.



History of NLP:

* (1940-1960): Focused on Machine Translation (MT) Natural Language processing started in the year 1940's.

* 1948- In the year 1948, the first recognisable NLP application was introduced in Birkbeck college, London.

* 1950- In the year 1950's there was a conflicting view between linguistic and computer science.

* Now, chomsky developed his first Book syntactic structures and claimed that language is generative in nature.

* In 1957, chomsky also introduced the Idea of Generative Grammer, which is rule Based descriptions of syntactic structures.

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

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

Augmented Transition Networks [ATN]:

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

case Grammer:

* case Grammar was developed by Linguist Charles. J. fillmore in the year 1968.

* case Grammar uses languages such as English to express the relationship between nouns and verbs by using the prepositions.

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

Example; "Neha broke the Mirror with the hammar". In this example case grammar Identify Neha as an agent, Mirror as a theme, and hammer as an instrument.

Advantages of NLp: [Natural Language processing]

* NLp helps users to ask questions about any subject and get a direct response within seconds.

* NLp offers exact answers to the question Means It does not offer unnecessary and unwanted Information.

* NLp helps computers to communicate with humans in their languages.

* It is very time efficient.

* Most of the companies use NLp to Improve the efficiency of documentation processes, accuracy of documentation, and Identify the information from language databases.

Disadvantages of NLP:

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

There are following two components of NLP:

1. Natural Language Understanding (NLU):

* Natural Language understanding (NLU) helps the machine to understand and analyse human language by entraiting the meta data from content such as concepts, entities, keywords, emotion, relations and semantic roles.

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

NLU involved the following tasks:

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

2. Natural Language Generation [NLG]:

* Natural language Generation (NLG) acts as a translator that converts the computerized data into natural language representation.

* It mainly Involves Text planning, sentence planning, and Text Realization.

Difference between NLU and NLG:-

NLU	NLG
NLU is the process of reading and interpreting language	NLG is the process of writing or generating language.

NLU	NLG
It produce non-linguistic outputs from Natural language Inputs.	It produces constructing Natural language outputs from non-linguistic Inputs.

Applications of NLP:-

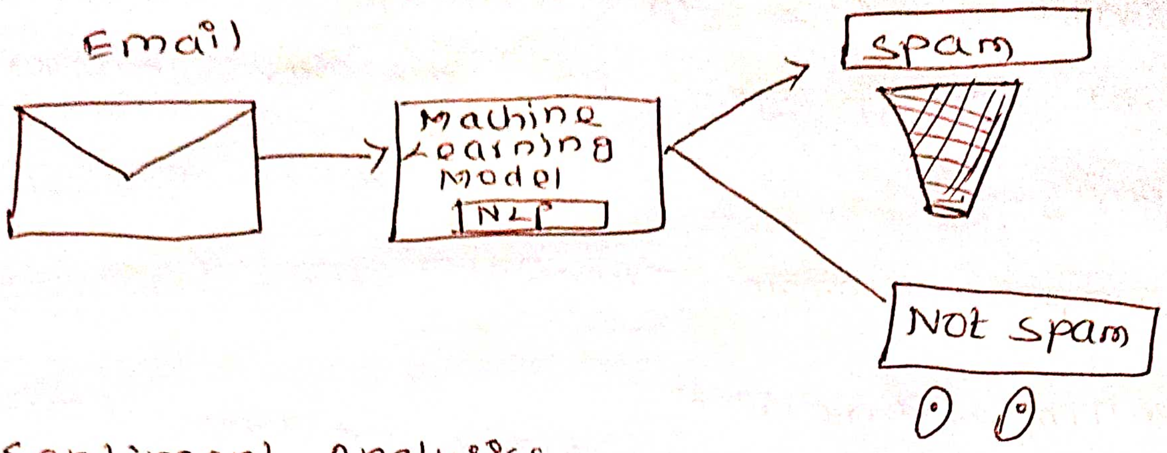
There are the following applications of NLP:-

1. Question Answering:

* question Answering focuses on Building systems that automatically answer the questions asked by humans in a natural language.

2. Spam Detection:

* Spam Detection is used to detect unwanted e-mails getting to a user's Inbox.



3. Sentiment Analysis:

* Sentiment Analysis is also known as opinion mining. It is used on the web to analyse the attitude, behaviour, and emotional state of the sender.

* This application is implemented through a combination of NLP (Natural Language processing) and statistics by assigning the values to the teat. (positive, negative or natural), identify the mood of the content. (happy, sad, angry).

4. Machine Translation:-

* Machine Translation is used to

Translate text or speech from one natural language to another natural language.

Example; Google Translator,

5. Spelling correction:

* Microsoft corporation provides word processor software like Ms-word, powerpoint for the spelling correction.

6. speech Recognition:

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

7. chat bot;

* Implementing the chatbot is one of the important applications of NLP.

* It is used by many companies to provide the customer's chat services.

8. Infoormation extraction

* Information extraction is one of the most important applications of NLP.

* It is used for extracting structured information from unstructured or semi structured machine readable format

How to build an NLP pipeline:

* There are the following steps to build an NLP pipe line.

* Sentence segmentation;

 * Sentence segment is the first step for Building the NLP pipeline.

 * It breaks the paragraph into separate sentences.

 Example; consider the following paragraph. Independence Day is one of the important festivels for every Indian citizen. It is celebrated on the 15th

August each year ever since India
got independence from the British Rule.
This day celebrates independence in the
True sense.

sentence segment produces the following
result;

1). "Independence Day is one of the
important festivals for every Indian citizen".

2) "It is celebrated on the 15th of
August each year ever since India got
independence from the British rule."

3) This day celebrates independence in
the True sense".

Word Tokenization;

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

Example;

Javatpoint offers Corporate Training,
Summer Training, online Training, and
Winter Training.

* Jav word Tokenizer generates the following result.

"Javat point", "offers", "corporate", "training", "Summer", "Training", "online", "training", "and", "winter", "training", "."

Stemming:

* Stemming is used to normalize words into its base form or root form.

* Example; celebrates, celebrated and celebrating all these words are originated with a single root word "celebrate".

* The Big problem with stemming is that sometimes It produces the root word which may not have any meaning.

Lemmatization:-

* Lemmatization is quite similar to the stamming. It is used to group defferent Inflected forms of the word, called lemma.

* The main difference between Stemming and Lemmatization is that it produces the root word, which has a meaning.

Example; In Lemmatization, the words intelligence, intelligent and intelligently has a root word Intelligent, which has a meaning.

Identifying stop words;

* In engish, there are a lot of words that appear very frequently like "is", "and", "the" and "a".

* NLP pipelines will flag these words as stop words.

* Stop words might be filtered out before doing any statistical analysis.

Example;

He is a good boy.

is, a; Stopping words.

Dependency parsing:

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

pos tags:

* pos stands for parts of speech, which includes noun, verb, advorb, and Adjective.
* It Indicates that how a word functions with its meaning as well as gramatically within the sentences.
* A word has one or more parts of speech Based on the context in which it is used.

Example; "Google" Something on the Internet.

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

Named Entity Recognition [NER]

* Named Entity Recognition is the process of detecting the named entity such as person name, movie name, Organization name, or location.

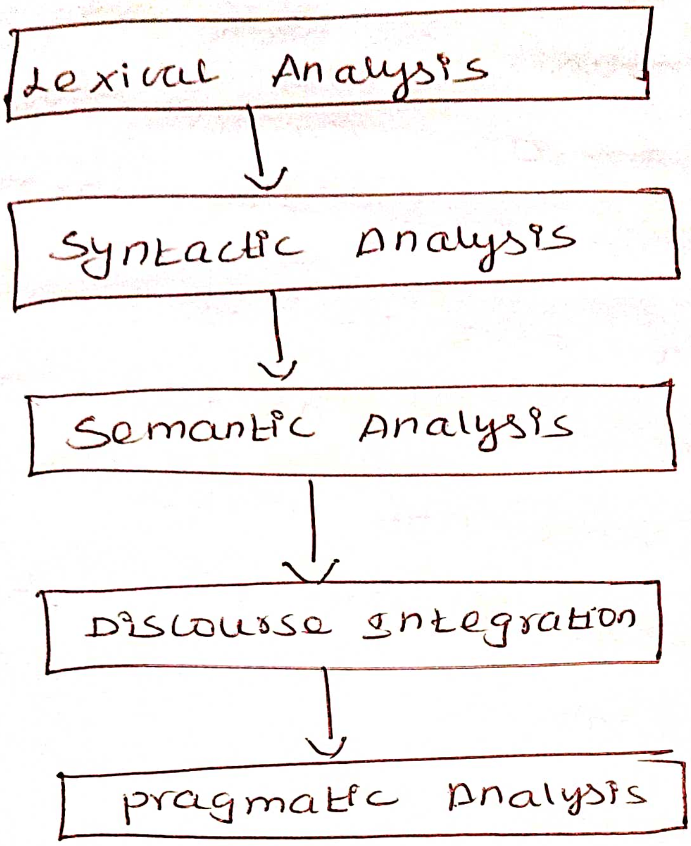
Example; Steve Jobs Introduced Iphone at the Macworld conference in San Francisco california.

Chunking:-

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

Phases of NLP:-

* There are the following five phase of NLP.



1) Lexical Analysis and Morphological

* The first phase of NLP is the lexical Analysis. This phase scans the source code as a stream of characters and converts in into meaningful lexems.

* It divides the whole text into paragraphs, sentences, and words.

2. Syntactic Analysis (parsing):

* Syntatic Analysis is used to check grammar, word arrangements, and shows the relationship among the woras.

Example; Agra goes to the poonam.

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

3. Semantic Analysis:-

* semantic analysis is concerned with the meaning representation. It mainly focuses on the literal meaning of words, ~~phar~~ phrases, and sentences.

4. Discousa Integration;

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

* pragmatic Analysis:

* pragmatic is the fifth and last phase of NLP.

* It helps you to discover the intended effect by applying a set of Rules that characterize cooperative dialogues.

Example:

" open the door" is interpreted as a request instead of an order.

Why NLP is difficult?

* NLP is difficult because Ambiguity and uncertainty exist in the language.

Ambiguity:-

* There are the following three ambiguity.

* Lexical Ambiguity:-

• Lexical Ambiguity exists in the presense of two or more possible meanings of the sentences within a single word.

Example:-

Manya is looking for a match.

In above example the word match refers to that either Mayna is looking for a partner or mayna is looking for a match. [Cricket or other match].

Syntactic Ambiguity;

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

Example;

I saw the girl with the binocular.

- In above example, did I have the binoculars? or did the girl have the binoculars?

* Rejerential Ambiguity-

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

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

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

NLP APIs:

* Natural Language processing APIs Allow developers to integrate human-to-machine communications and complete several useful tasks such as Speech recognition, chatbots, spelling correction, Sentiment analysis etc,

The list of NLP API is the given below:-

IBM watson API;

* IBM watson API combines different sophisticated machine learning techniques to enable developers to classify text into various custom categories.

* It supports multiple languages, such as English, French, Spanish, German, chinese etc.,

* with the help of IBM watson API, you can extract insights from texts add automation in workflows, enhance search, and understand the sentiment.

Chatbot API;

* chatbot API allows you to create Intelligent chatbots for any service.

* It supports unicode characters, classifies text, multiple languages, etc.,

* It is very easy to use. It helps you to create a chatbot for your

web applications.

Speech To text Api :

* Speech To Text Api is used to convert speech to Text.

Sentiment Analysis Api ;

* Sentiment Analysis Api is also called as "opinion mining" which is used to identify the Tone of a user (positive, negative or neutral).

Translation Api by SYSTRAN:-

* The Translation Api by SYSTRAN is used To Translate The Txt from the

* Source Language.

Text Analysis Api by AYLIEN;

* Text Analysis Api by AYLIEN is used to derive meaning and Insight from the textual content.

* Cloud NLP Api;

* The cloud NLP API is used to Improve the capabilities of the application

Using natural Language processing Technology.

Google cloud Natural Language API;

* Google cloud Natural Language API allows you to extract beneficial Insights from unstructured Text.

Bag of words model;

* Bag of words model. Whenever we apply Bag of words is a Natural Language processing Technique of Text modeling.

* Through this bag, we will learn more about why Bag of words is used, we

Bag of words:

* Bag of words is a natural language processing Technique of Text modelling.

* In Technical Terms, we can say that is a method of feature extractation with Text data.

Example:- Let us see an example of how the bag of words technique converts text into vectors.

Example (1); without preprocessing.

Sentence 1; welcome to Great learning, Now start learning.

sentence 2; "Learning is a good practice".

sentence 1	sentence 2.
welcome	learning
to	is
Great	a
learning	good

practice.

Now

start

learning.

Step1; Go through all the words in the above Text and make a list of all of the words in our model Vocabulary

- welcome
- TO
- Great
- learning
- /
- Now
- start
- learning
- is
- a
- good
- practice.

* Note that the words 'Learning' and 'learning' are not the same here because of the difference in their cases and hence are repeated.

* Because we know the vocabulary has 12 words, we can use a fixed length document representation of 12, with one position in the vector to score each word.

* The scoring of sentence 1 would look as follows.

word	frequency
welcome	1
to	1
Great	1
Learning	1
,	1
Now	1
start	1
learning	1
is	0
a	0

good	0
practice	0

writing the above frequencies in the vector.

Sentence 1 → [1, 1, 1, 1, 1, 1, 1, 1, 0 0 0]

Sentence 2, The scoring would like

word	Frequency
welcome	0
to	0
Great	0
learning	1
,	
Now	0
start	0
learning	0
is	1
a	1
good	1
practice	1

* A vector space represents each word by a vector of Real numbers.

Text vectorizer;

* In natural Language processing (NLP) we often Talk about Text vectorization -Representing words, sentences or even larger units of Text as vector.

* other data Types likes, Images, sound and videos.

TF-IDF;

* TF- IDF stands for Term frequency of records. Inverse Document Frequency

Terminologies:-

- Term Frequency: In document d, the frequency represents the number of Instances of a given word t.

Sentence 2 -> [0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1]

Sentence	Welcome	to	Great	Learning	,	Now	start	Learning	is	a	good	practice
Sentence 1 :	1	1	1	1	1	1	1	1	0	0	0	0
Sentence 2	0	0	0	0	0	0	0	1	1	1	1	1.

Word vectorization:

* Word vectorization is a methodology in NLP to map words or phrases from vocabulary to a corresponding vector of real numbers which used to find word predictions, word similarities/ semantics.

Word vectorizer;

* Converting words to vectors or word vectorization is a Natural Language processing.

* The process uses language models to map words into vector space.

* The weight of a term that occurs in a document is simply proportional to the term frequency

tf(t,d) = count of t in d / number of words in d.

Document Frequency: This tests the meaning of the text, which is very similar to TF, In the whole corpus collection.

* The only difference is that in document d, TF is the frequency counter for a term t, while df is the number of occurrences in the document set N of the term.

Df(t) = occurence of t in documents

Recommendation System: ⊗⊗ 2mark.

* Recommendation system widely use unsupervised learning Techniques for building recommendation applications for

different web appiications and e- commerce websites

Classification and filtering:

- * Classification algorithm is a supervised learning Technique that is used to identify the category of new observations on the basis of Training data.

- * In classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or, groups.

- * Such as, yes or NO, 0 or 1, Spam or not spam, cat or dog etc.,

- * Classes can be labelled as Targets / labels or categories.

* unlike regression, the output variable of classification is a category not a value, such as "Green or Blue", "fruit or animal" etc,

* since the classification algorithm is a supervised learning technique, hence it take labeled input data, which means it contains input with the corresponding output.

* In classification algorithm, a descrete output function (y) is mapped to input variable (x).

$\boxed{y = f(x), \text{ where } y = \text{categorical output}}$

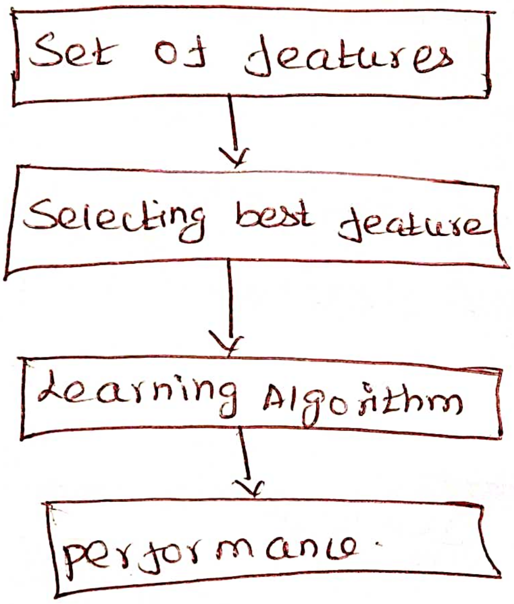
Filtering:

* In filter method, features are selected on the basis of statistics measures.

* This method does not depend on the learning algorithm and chooses the features as a pre-processing step.

* The filter method filters out the irrelevant feature and redudant columns from the model by using different metrics through ranking.

* The advantage of using filter methods is that It needs low computational time and does not overfit the Data.



Some common Techniques of filter methods are as follows;

* Information Gain

* Chi- square Test

* Fisher's Score

* Missing value Ratio.

Information Gain;

* Information gain determines the reduction in entropy while transforming the dataset.

* It can be used as a feature selection Technique by calculating the Information gain of each variable with to the Target variable.

Chi- Square Test:

* Chi- square Test is a Technique to determine the Relationship between the categorical variable.

* The chi- square value is calculated between each feature and the Target variable and the desired number of features with the best chi- square value is selected.

Fisher's Score;

* Fisher's Score is one of the popular supervised Technique of

feature selection.

* It Returns the rank of the variable on the fisher's criteria in desending order.

Missing value Ratio;

* The value of the missing value ratio can be used for evaluating the feature set against the the Threshold value.

* The formula for obtaining the missing values in each column divided by the total number of observations.

* The variable is having more than the threshold value can be dropped.

$$\text{missing value Ratio} = \frac{\text{Number of missing values} = 100}{\text{Total number of observations}}$$