NATURAL LANGUAGE PROCESSSING

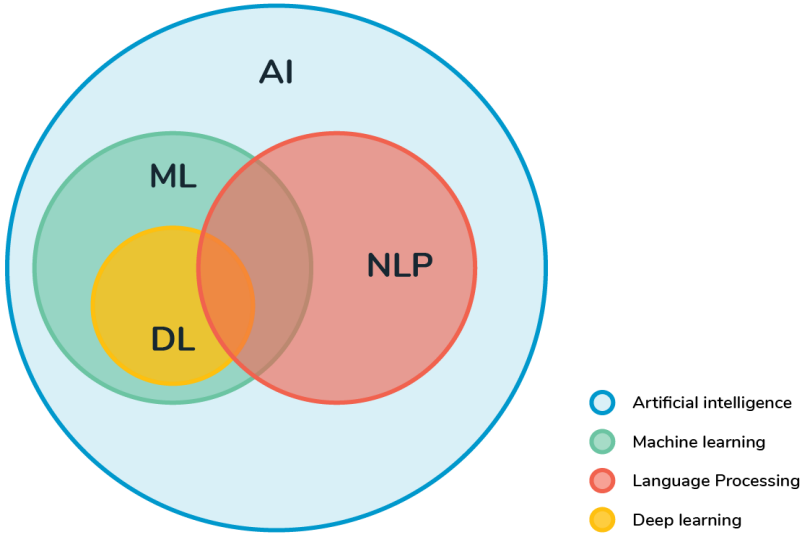
EFFORTS BY:-

DR. SHARANJEET KAUR DHAWAN

CONTENTS

INTRODUCTION TO NLP

- Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.



DATASET

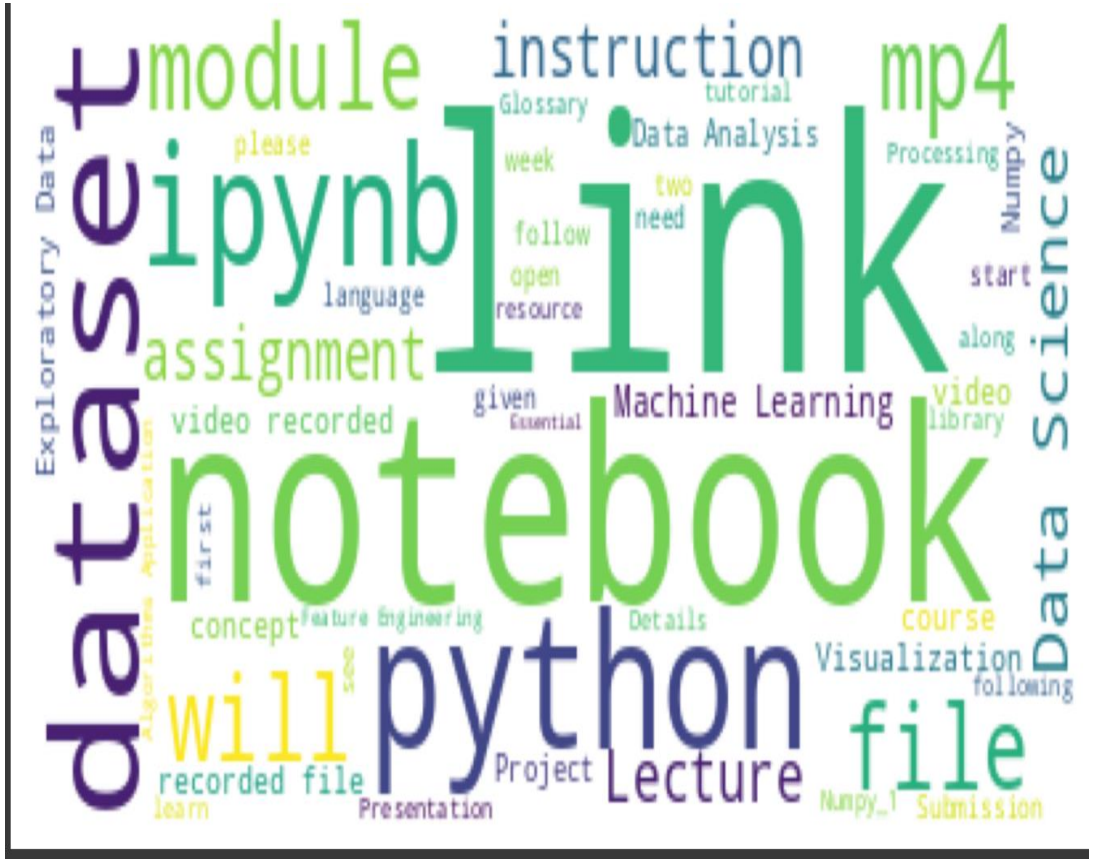
- "Essential Data Science Skill Course Guide Book_Ladakh.txt"

WORDCLOUD

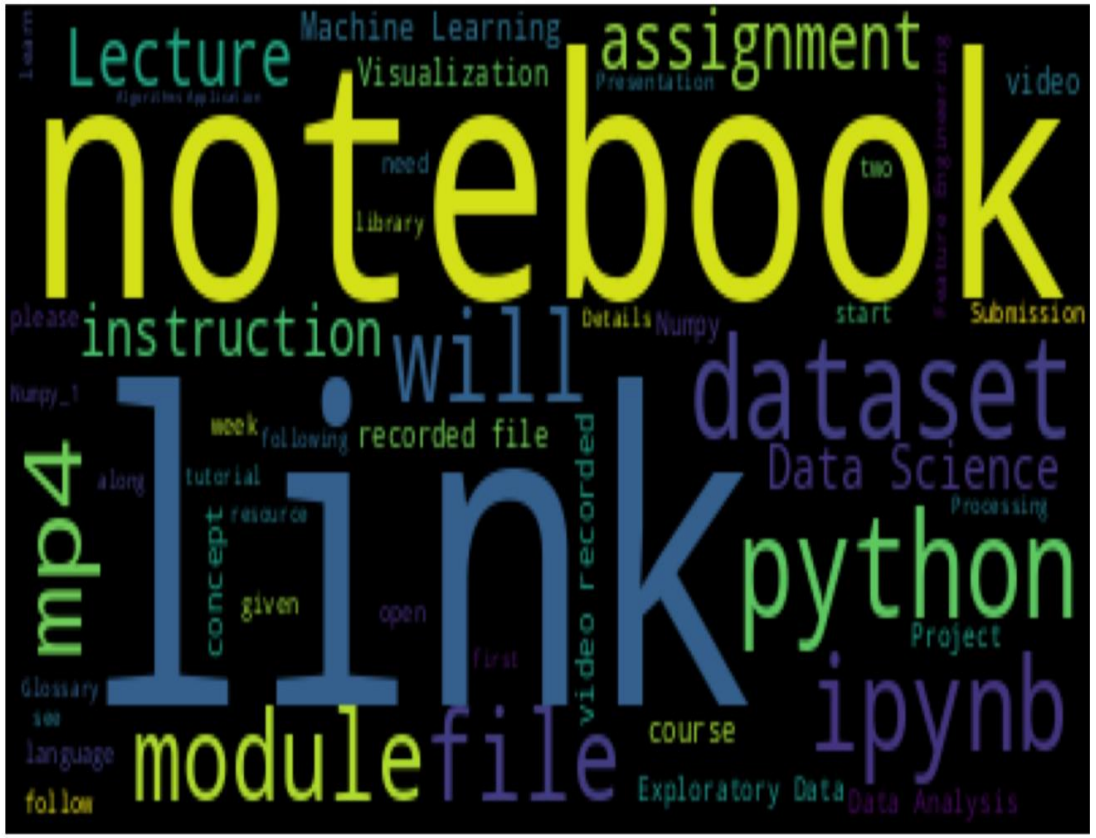
It is a visualization technique for text data wherein each word is picturized with its importance in the context or its frequency.

WORDCLOUD FROM DATASET

Orignal wordcloud from dataset

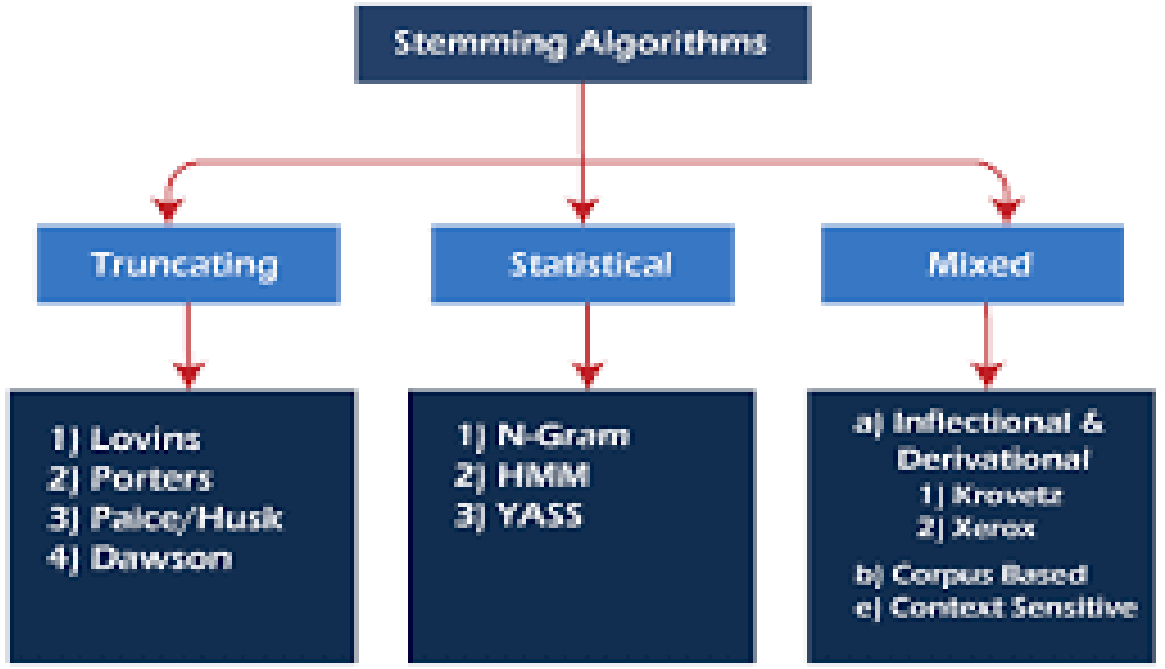


Upgraded wordcloud from same dataset



STEMMING

- Stemming is the process of reducing a word to its stem that affixes to suffixes and prefixes or to the roots of words known as "lemmas". Stemming is important in natural language understanding (NLU) and natural language processing (NLP).



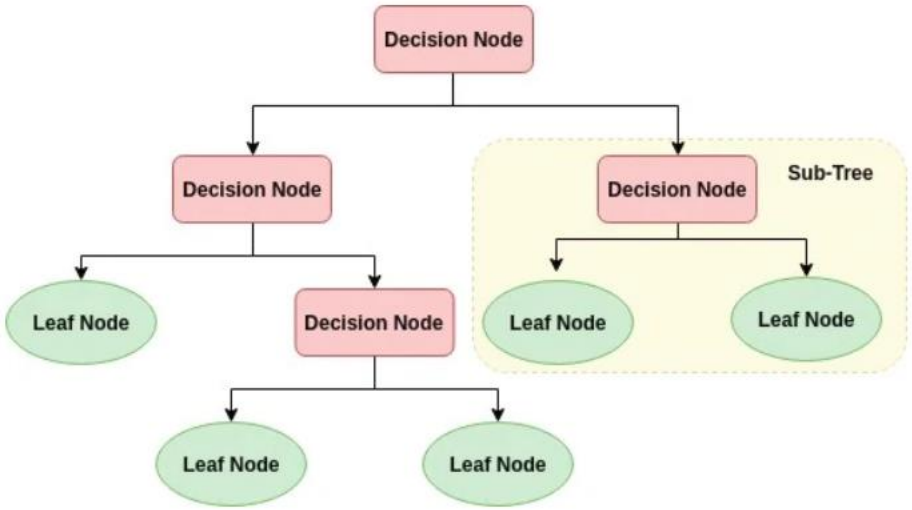
Classification of Stemming Algorithms

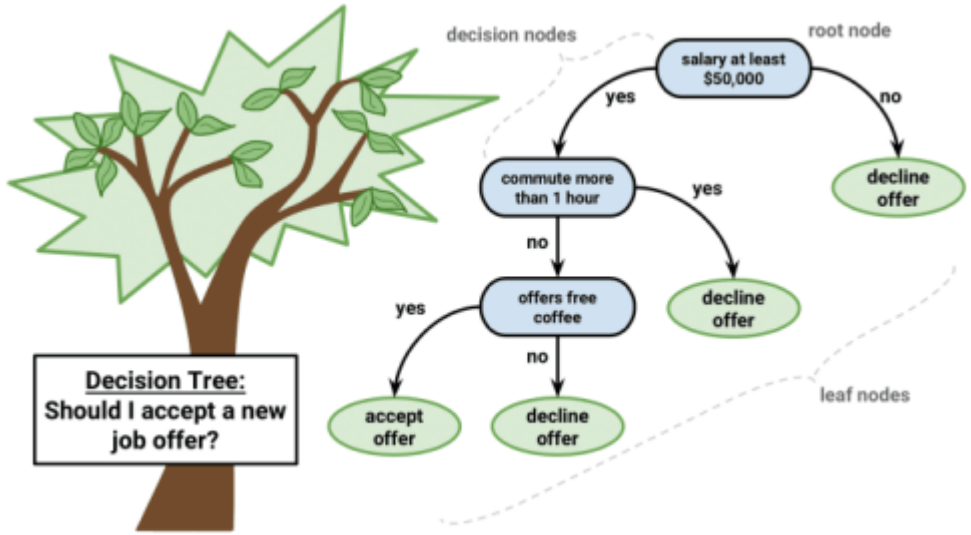
LEMMATIZATION

- Lemmatization deals with reducing the word or the search query to its canonical dictionary form. The root word is called a 'lemma' ad the method is called lemmatization. This approach takes a part of the word into consideration in a way that it is recognized as a single element.

DECISION TREE ALGORITM USED

- A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.





N-Gram

- N-grams are continuous sequences of words or symbols, or tokens in a document. In technical terms, they can be defined as the neighboring sequences of items in a document. They come into play when we deal with text data in NLP (Natural Language Processing) tasks.

N-grams

- $P(w_n \mid w_1 w_2 \ldots w_{n-1})$ is called a parameter of the language model
- To estimating the values of the parameters of an N-gram model from the training data:

Unigram $P(w_i) = \frac{C(w_i)}{N}$

$C(w_i)$ = count of occurrence of w_i

N = total number of words in the training data

Bigram $P(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i)}{C(w_{i-1})}$

Trigram $P(w_i|w_{i-2}w_{i-1}) = \frac{C(w_{i-2}w_{i-1}w_i)}{C(w_{i-2}w_{i-1})}$

SENTIMENT ANALYSIS

- Sentiment analysis is the process of analyzing digital text to determine if the emotional tone of the message is positive, negative, or neutral. Today, companies have large volumes of text data like emails, customer support chat transcripts, social media comments, and reviews.

THANK YOU