21CSE356T
Kanmani mam.

Natural Language Processing

Includes machine learning & python.

1) Unit - I → Basic preprocessing. → to prepare data.
   data cleaning.
   data transformation.
2) Unit - II → Compiler topics.
3) Unit - III → Analysis.
4) Unit - IV → Model.
5) Unit - V → Application.

~~Compilers~~ Learning & understanding of computers.
NLP used in understanding.

NLP :- Text, speech, music, images can be used.

Firstly,
Preprocessing - divide the data.

Ex: add chocolates and biscuits to my shopping list.

Secondly,
compiler - divide the data into tokens.
have to analyse vector to word or word to vector.

then,
put into model.

* we use text prediction ②, text generation ①, text summarization ③

* Applications of NLP :-
→ Google translate.
→ web search.
→ spam filtering.
→ Chat bot.

* Structured data - data stored in rows & columns.
data is organized.

* Unstructured data - images, music, videos; not organized.
different data types.

→ XML : Semi structured data.

<list>
<group> choco </group>

* Structured data → unstructured data.
NLG (generation).

* Unstructured data → Structured data.
NLU (understanding).

How.
NLP application used in summarization of text.
→ Claude - AI , Fusion : Generative Image.

Pre trained
birds.
llama

ML → Data preprocessing → Feature extraction → Model
(removing null values & selection ↙ training, testing.
etc).
Back propagation neural networks.

PCM → dimensionality reduction.

RNN → text based
LSTM → time based.

DL →
BERT } reads from both direction.
Roberta }
Transformers.

Trained for 100 GB of data.

Bert - taking data from dictionary.
Roberta - " " from whole library.

ML/DL → trains data
↓
NLP / LLM → understands the data.
↓
large language model.

* NLP → It is a branch of AI, that gives a machine the ability to read & understand and derive meaning from human language.

This is in order to interface in computer in both written and spoken context using natural human languages instead of computer language.

* Applications of NLP:

1) Google translator
2) Spam filtering.
3) Chatbots.
4) Sentimental analysis
5) text prediction
6) text generation.

* Converting unstructured data into a structured data is called natural language understanding. (NLU).

* Converting structured data into unstructured data is called natural language generation (NLG).

* NLU → Alexa.
* NLG → Alexa.

1) Tokenization
2) Stop words.
3) Normalization → Stemming, lemmatization.
4) Speech tagging.

* Steps in NLP:-

1) Tokenisation.
2) Lemmatisation.
3) Stop words.
4) stemming.
5) Bag of words.
6) TFIDF
7) N Grams.
8) word 2 vector
9) Avg word 2 vector.

Stages in NLP

1) Morphological analysis
2) Syntactic analysis
3) Semantic analysis
4) Discourse analysis.
5) Pragmatic analysis.

Tensorflow, keras package.

* Morphological analysis:-

It is a lowest level analysis which study the different forms of the words.

word level analysis.

Token: [Jhon] [at] ~~[the]~~ [pizza] ~~[.]~~

stop words : key words.

* Stemming:-

Proces of reducing words into base form.

2

studies - studi
studying - study.
cars → car.

Lemmatization.

studies & studying → study.
cars → car.

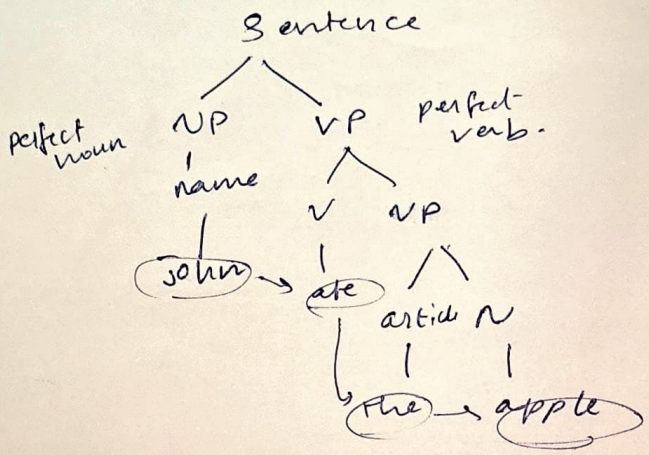
* N gram:-

It takes previous word and generates the next word. It is a sequence of n continuous words in a uni-gram, 1-gram. Finding the probability of word from next in next sequence of word.

* Syntactic analysis!-

It checks for the gramatical errors.

It constructs a parse tree



* Semantic analysis!-

checks the ~~sys~~ sentence meaningful or not.

→ analyse word by word.
→ construct graph.
→ Make understand the model.

The distance b/w 2 words is calculated in graph.

* Discourse analysis!-

ex: monkeys eat banana when they wake up.
who are they?

monkeys eat banana when they are ripe.

whole meaning of sentence has changed. 'They' should be understand by the machine.

* Pragmatic analysis!-

Close the door → ordu
Please close the door → request.

Based on emotions.