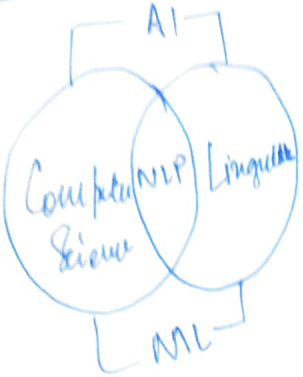
Natural Language Processing (NLP) (Day-1) 12-08-2023

Meenakshi.
UG - VIT
PG -
PHD - Research in ML, DL
CV, NLP, Adversary ML



→ NLP deals with only text data, emotion detection is not considered.

→ Phases in NLP.

* Lexical Analysis → Converting the sentences into tokens like individual words
  ↳ Morphology.
  ↳ Formulation of words.
* Syntactic Analysis → whether the language is following proper rules of the language grammar.
* Semantic Analysis (Syntax) → Understanding the meaning.
* Discourse Analysis → Analysing text in the paragraph level.
* Pragmatic Analysis → Let the system understand the general day to day conversation

ex:- I drink Taj → Taj is in context denoting Taj tea not Taj mahal, Taj hotel etc

Applications:

→ Some of the application.
→ Check Credit Analysis, Worthiness, Language Translation, Sentiment Analysis, Customer Support, Work Routing, Identify Similar Legal Cases. (Document Similarity)
Information Retrieval, Information Extraction, Q & A - Chatbot, Optical Character Recognition.
Text categorization (a classification), Word prediction (Auto completion), Speech recognition (Alexa), Machine Translation (Google Translate)

→ Ambiguity (Challenges of NLP) + Examples

* Word level.
  ex: Bank, Can, Bat.

* Sentence level.
  ex: - The man saw the girl with the telescope.
  - We saw his duck.

* Pragmatic level.
  ex: Tourist (checking out of the hotel): Waiter, go upstairs to my room and see if my sandals are there; do not be late; I have to catch the train in 15 minutes.
  ex: Waiter (running upstairs coming back panting): Yes sir, they are there.

* Discourse level.
  ex: The horse ran up the hill. It was very steep. It soon got tired
  (It → Hill; It → horse.)

→ Features in Textual processing.

→ NLP is based on:
* Prob. & Stats.
* ML
* Linguistics.
* Common sense.

→ Standard terms.
* Corpus: A body of text samples.
* Document: A text sample.
* Vocabulary: A list of words used in the corpus.
* Language model: How the words are supposed to be organised.

→ NER → Named Entity Recognition.

Ex: Sports → Cricket
→ Football
→ Tennis
} Entities.

→ NLP → NLU → Nat. Lang. Understanding.
→ NLG → Nat. Lang. Generation

→ GAN → Generative Adversarial Network.

→ Regular Expression :-

* Pattern matching.
* Purpose is to find the pattern within the corpus.

Regex.

Uses:
→ Web-search,
→ Retrieval applications.
→ Word-processing, computation of frequencies.

→ Basic Regular Patters.
→ Simple text search.

RE.
/"text"/

→ It is case-sensitive.

→ Disjunction of Characters.
Syntax.
/[word]/

→ Range of characters.
Syntax /[a-z]/

→ Caret - Negation.

→ The Caret ^ is the first symbol after the square bracket.

→ !

→ preceding character is ignored.

Ex: Color and Colour.

Colo u?r

↓

Ignored.

→ Kleene Operator.

The Kleene * means "Zero or more occurrences of the immediately previous Character or regular expression.

→ The Kleene + means "One or more occurances

→ Wild card expressions.

→ Syntax /./

Ex: /beg.n/

If we need to find only "." we can use /\./ or /[.]/

→ Anchors:

2 anchor expressions.

→ Caret - /^/ → specifies start of the sentence / Corpus.

→ Dollar - /$/ → specifies end of the sentence.

→ Word Boundary

- \b matches a word boundary while \B matches a non-boundary.

→ How to specify for ex: guppy and guppies. use "|" operator.

ex: /gappy/
/guppy|ies/

→ Repeaters.

ex: [a-z]{3}

→ Advanced operators.

\d [0-9]
\D [^0-9]
\w
\W
\s [\r\t\n\f]
\S [^\s]

→ ex: question.
Check the url file name @ gmail.com.

Spell checking:

→ Ex: Piece
peace

} Two types of errors.
① Word error. → when peace is written instead of piece.
② Non-word error.

word error.
→ Replace.
→ Insert.
→ Removal.

ex: tutor → misspelt.
tumour → correct.

2 solutions.

1st
| Replace t → m
| Insertion u
} optimal.

2nd
| Removal t
| Insertion m
| Insertion u
} General.

Minimum Edit Distance.

Dynamic
Greedy
Backtracking
} Algorithms.

Ex. Misspelled aecdg

Original abcfg

	N	a	b	c	f	g
N						
a						
e						
c						
d						
g						