

20-11-2020

Natural Language Processing Fundamentals.

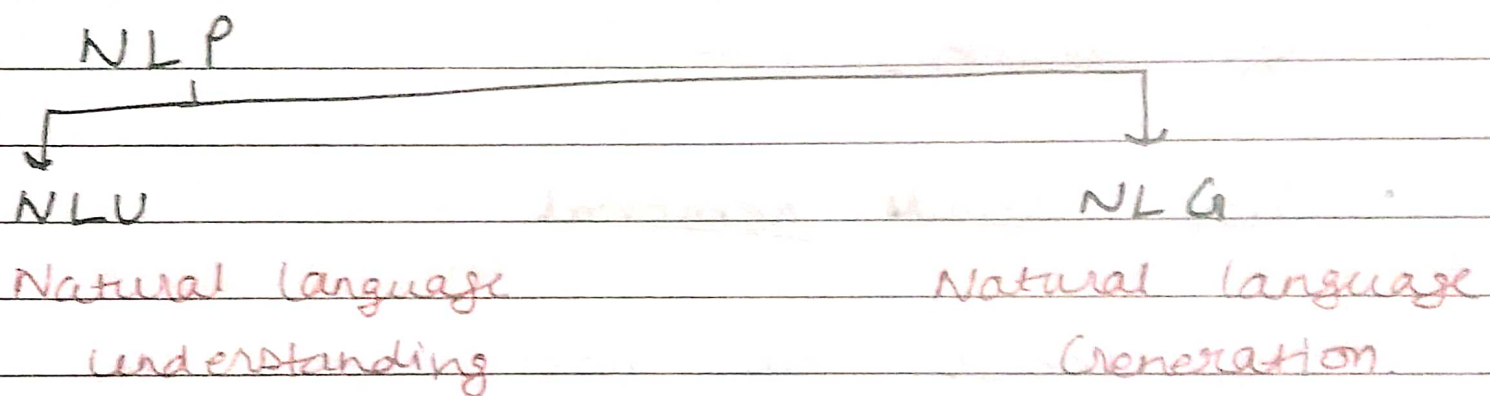
Chapter -1 \Rightarrow Introduction to NLP.

Objectives:

- i) Describe what NLP is all about
- ii) Describe the history of NLP
- iii) Differentiate b/w NLP & text analytics
- iv) Implement various preprocessing tasks
- v) Describe the various phases of NLP project.

Text analytics:

The art of extracting useful insights from any given text data



⇒ Various steps in NLP :

- Tokenization

Splitting a sentence into constituent words / tokens.

- Unigrams : one token represent one word.

"I am reading a book". \Rightarrow

"I" "am" "reading" "a" "book" "."

2 tokens \rightarrow bigrams

3 tokens \rightarrow trigrams

n tokens \rightarrow n grams.

- POS tagging

(Parts of speech tagging)

Process of tagging words within sentences into their respective parts of speech.

- Stop words removal

Stop words do not impact the meaning of sentences

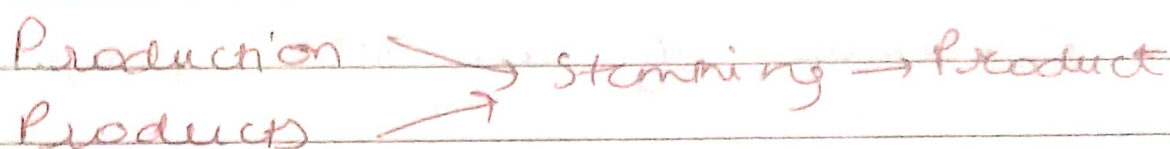
⇒ Text Normalization

- There are some words which are spelt, pronounced and represented differently, but they mean the same thing.
- Text Normalization is a process wherein different variations of text get converted into a standard form.

eg : Spelling correction, stemming, lemmatization etc.

- Stemming

Reduce words to their stems



- Lemmatization

Reduce words to their base form

- Slower than stemming but makes more sense.

⇒ Name Entity Recognition (NER)

Mapping words / tokens to categories like name of person, place and so on.

- Word Sense Disambiguation

Process of mapping a word to the correct sense it carries.

- Sentence boundary detection

Method of detecting where one sentence ends and other begins.

⇒ Structure of an NLP project :

