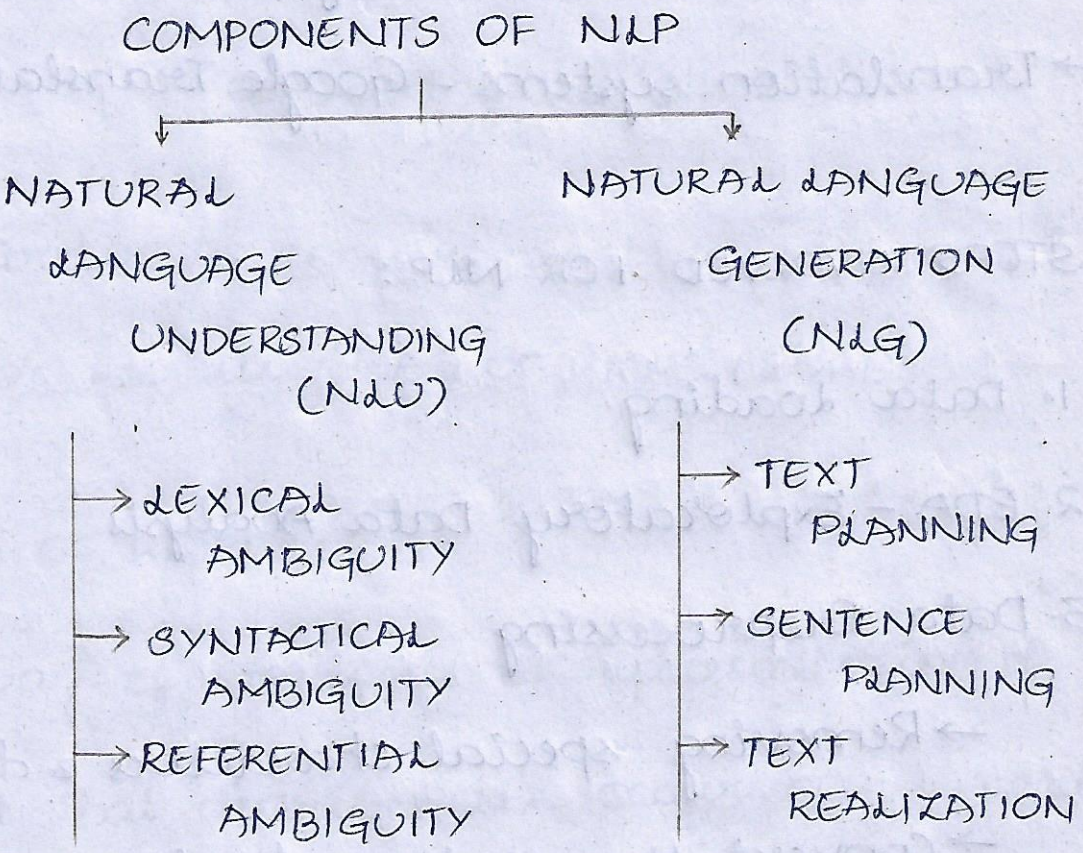
LECTURE - 57

COMPONENTS OF NLP:

NLP is basically divided into two components.



NOTE:

NLU is naturally harder than the NLG tasks.

APPLICATIONS OF NLP:

→ Grammarly, Microsoft word, Google Docs.

→ Search Engines like Google, DuckDuckGo.

→ Voice Assistants - Siri, Alexa.

→ News feed - Facebook, Google News.

→ Translation systems - Google Translate

STEPS INVOLVED FOR NLP:

1. Data loading.

2. EDA - Exploratory Data Analysis.

3. Data Preprocessing

   → Removing special characters & digits.

   → Convert the sentence to lower case.

   → Remove stop words

   → Stemming or lemmatization.

4. Data preparation.

   → Train-Test split

   → Text to numerical vector using Bag of words.

5. Training and Evaluation.

So, In NLP,

→ X labels - If categorical - use One Hot Encoding to convert into Numerical.

→ Y labels - Sentiment columns after converting to '0's & '1's.

In order to convert text to numerical vector we use "BAG OF WORS" (BOW).

BAG OF WORDS:

A bag-of-words is a representation of text that describes the occurance of words within a document.

→ It involves two things:

↳ A vocabulary of known words.

↳ A measure of the presence of known words.

Apart from BOW, we have TFIDF & W2V.

where,

TFIDF — Term Frequency-Inverse Document Frequency

W2V - Word2Vec.

The terms used in NLP are.

→ CORPUS - collection of text

→ DOCUMENT - Each row in a corpus.

→ DOCUMENT TERM FREQUENCY

EXAMPLE OF BAG-OF-WORD MODEL:

STEP-1: Collect the data.

It was the best of times,

It was the worst of times,

It was the age of wisdom,

It was the age of foolishness.

In the above example, treat each line as a seperate "DOCUMENT" and the 4 lines as entire corpus of documents.

STEP-2: Design the Vocabulary.

The unique words are:

It, was, the, best, of, times, worst, age, wisdom, foolishness.

In this, it has 10 words of vocabulary.

STEP-3: Create Document Vectors.

S.NO	IT	WAS	THE	BEST	OF	TIMES	WORST	AGE	WISDO-M	FOOLI-SHN-ESS
1	1	1	1	1	1	1	0	0	0	0
2	1	1	1	0	1	1	1	0	0	0
3	1	1	1	0	1	0	0	1	1	0
4	1	1	1	0	1	0	0	1	0	1

This is the Document term matrix & also the Numerical representation of text.

→ By default, it is called as sparse.

i.e., doesn't save '0's.

So, if we have fairly large number of documents then?

→ import countvectorizer - Bag-of-word.

While learning vocabulary, we take

* countvectorizer(ngram_range = (1, 2, 3))

NOTE:

if ngram increases - vocabulary increases.

In the previous example,

- when vocabulary is 10, ngram = 1; the dimensionality is '10'.

- After data cleaning we get the dimensionality as '6'.

NOTE:

Before going to Bag-of-words, we should perform the data cleaning.

* Instead of sparse, if we use dense, it is very hard to save in the memory

We should fit on LEARN VOCABULARY & transform on CREATE DOCUMENT

NOTE:

We should never apply fit. transform on the test data since we will never learn from test data.

SERIALIZATION: SAVING A FILE

This means we'll have translate its contents and structure into a format that can be saved like a file or a byte string.

DESERIALIZATION : LOADING TO MEMORY

It is the opposing process which takes data from a file, stream or network and rebuilts into an object.

→ Serialized objects can be structured in text such as XML, JSON or YAML.

* ↳ Serialization & Deserialization are safe, common processes in web applications.

PICKLE :

Pickling is a way to convert a python object (list, string, etc) into a character stream.

→ The idea is that, the character stream contains all the necessary information to reconstruct the object in another python script.