Semester : VI Subject : DAV Academic Year: 2023-2024

TEXT ANALYSIS STEPS

The text analysis steps consists the following:

- → (POS) Parts of speech Tagging
- → Parsing
- → Word Normalization.
  - ↳ Stemming
  - ↳ Lemmatization.

(POS) Part of Speech Tagging:

It is an linguistic activity in Natural Language Processing (NLP) wherein each word in a document is given a particular part of speech (adverb, adjective, verb etc) or grammatical category.

Part of Speech	Tag.
Noun	n
Verb	v
Adjective	a
Adverb	r

Example:

Consider the sentence: "The quick brown fox jumps over the lazy dog."

After performing POS Tagging:

"The" is tagged as determiner (DT)
"quick" is tagged as adjective (JJ)
"brown" is tagged as adjective (JJ)
"fox" is tagged as noun (NN)

Semester : VI Subject : DAV Academic Year: 2023 20 24

"jumps" is a tagged as ~~no~~ verb (VBZ)

"over" is a tagged as preposition (IN)

"the" is a tagged as a determiner (DT)

"lazy" is a tagged as adjective (JJ)

"dog" is a tagged as noun (NN).

Example in Phython:

```
#Importing the NLTK Library
 import nltk
 from nltk.tokenize import word tokenize.
 from nltk import pos_tag

#Sample Text
text = "NLTK is a powerful library for natural language
        processing".

#Performing Pos tagging
pos tags = pos tag(words)

#Displaying the Pos tagged result in separate lines:
print("Original Text:")
print(text)

print("\n POS Tagging Result:")
for word, pos tag in pos tags:
    print(f"{word}: {pos tag}")
```

Output:

```
Original Text:
NLTK is a powerful
library for natural
language processing:
Pos tagging Result:
NLTK: NNP          | for: IN
is : ~~DT~~ VBZ    | natural: JJ
a : DT             | language: NN
powerful : JJ      | processing: NN
library : NN       |
```

Semester : VI Subject : DAV Academic Year: 2023 2024

Parsing :-

- Parsing means dividing your text into multiple segments.
- Parsing helps to determine the relationship between different words in a sentence.

Example :

1. A handsome guy (Noun Phrase)
2. The blue umbrella (Noun phrase).

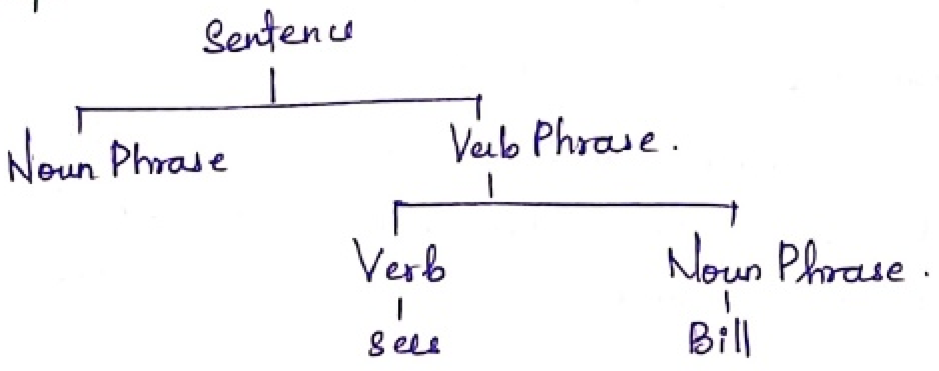
using POS tagging we may get the below result for "the blue umbrella".

The : Determiner
blue : Adjective
Umbrella : Noun.

Types of Parsing :

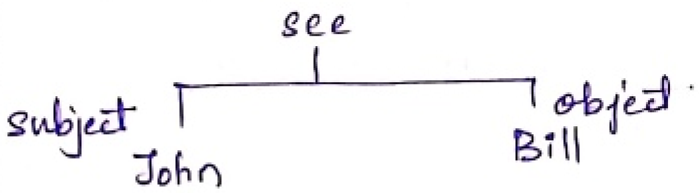
Synthetic parsing :- It uses rules to break the sentence in sub-phrases.

Example: John sees Bill.



Dependency Parsing : It aims to break the sentence depending on the relationship between the words rather than any predefined rules.

Example: John sees Bill.



Semantic parsing :- It aims to transform a sentence to logical, formal representation.

Example:

'How many runs did Dhoni scored in the match?' can be transformed as a SQL query (or any other formal representation) like SELECT runs from MATCH where player = 'DHONI'.

Word Normalization:

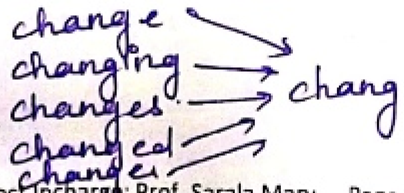
Word normalization is done using stemming or lemmatization.

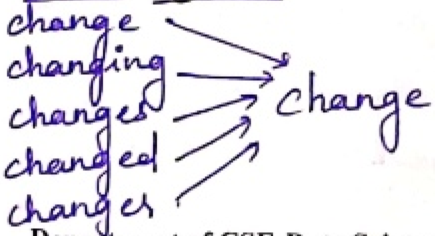
Stemming: It uses a rule based system to bring a word to its canonical form. Like removing 'ing' from dancing to form 'danc' or 'ticked' to 'tick'.

As you can see, stemming might not produce a dictionary word all the time after normalization.

Lemmatizer: It is a more intelligent system that keeps a dictionary on its side while normalizing words. Hence it will normalize 'dancing' to 'dance' and not 'danc' as done in stemming.

Example: Steming vs lemmatization.


change
changing
changes
changed
changes
chang


change
changing
changes
changed
changes
Change