Rohan Nyati
500075940
R177219148
BATCH – 5 (Ai & Ml)

ASSIGNMENT

Write short notes on:

1) Regular expressions in NLP
   A regular expression (RE) is a language for specifying text search strings. RE helps us to match or find other strings or sets of strings, using a specialized syntax held in a pattern. Regular expressions are used to search texts in UNIX as well as in MS WORD in identical way.

2) POS tagging

   Part-of-Speech (PoS) tagging, then it may be defined as the process of assigning one of the parts of speech to the given word. It is generally called POS tagging. In simple words, we can say that POS tagging is a task of labelling each word in a sentence with its appropriate part of speech. We already know that parts of speech include nouns, verb, adverbs, adjectives, pronouns, conjunction and their sub-categories.

   Most of the POS tagging falls under Rule Base POS tagging, Stochastic POS tagging and Transformation based tagging.

3) Lexicons
   If you have ever watched a football game, you might think that there's something wrong with your hearing. You probably heard the players yelling strange phrases and words like, "hut, hut, hike" or "blitz!"

   You might feel the same way when you hear lawyers on TV yelling, "I object" to a judge or, "you may cross-examine, counsel" to other lawyers. The vocabulary of a group of people, language, or field is called a lexicon.

   These groups of people—football players and lawyers—seem to have their own language or way of speaking.

   In the English language the word "hike" has a meaning: to take a walk in the wilderness or to raise up. So, hike is a part of the English vocabulary or lexicon. As we mentioned up above, hike is also a part of the special lexicon of football players. When a football player yells "hike," it means to be ready to catch the ball. Words or phrases that have meaning in a

lexicon are called lexemes. In this lesson, we'll look at how lexicons are used in the world around us.

4) Morphenes
   A "morpheme" is a short segment of language that meets three basic criteria: 1. It is a word or a part of a word that has meaning. 2. It cannot be divided into smaller meaningful segments without changing its meaning or leaving a meaningless remainder.

5) Natural language generation
   Natural language generation is a software process that produces natural language output. While it is widely agreed that the output of any NLG process is text, there is some disagreement on whether the inputs of an NLG system need to be non-linguistic.

6) Co-reference resolution
   In linguistics, coreference, sometimes written co-reference, occurs when two or more expressions refer to the same person or thing; they have the same referent.
   Coreference resolution is the task of finding all expressions that refer to the same entity in a text. It is an important step for a lot of higher level NLP tasks that involve natural language understanding such as document summarization, question answering, and information extraction.

7) NER
   Named-entity recognition is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.

8) Text summarization
   Text summarization in NLP is the process of summarizing the information in large texts for quicker consumption. In this article, I will walk you through the traditional extractive as well as the advanced generative methods to implement Text Summarization in Python. .Text summarization methods can be grouped into two main categories: **Extractive** and **Abstractive methods** * **Extractive Text Summarization** It is the traditional method developed first. The main objective is to identify the significant sentences of the text and add them to the summary. You need to note that the summary obtained contains exact sentences from the original

text. * **Abstractive Text Summarization** It is a more advanced method, many advancements keep coming out frequently(I will cover some of the best here). The approach is to identify the important sections, interpret the context and reproduce in a new way. This ensures that the core information is conveyed through shortest text possible.