MES Wadia College of Engineering Pune-01
Department of Computer Engineering

Name of Student:	Class:
Semester/Year:	Roll No:
Date of Performance:	Date of Submission:
Examined By:	Subject: LP_VI (EL-V NLP)

Assignment No. 1

Aim: Perform tokenization (Whitespace, Punctuation-based, Treebank, Tweet, MWE) using NLTK library. Use porter stemmer and snowball stemmer for stemming. Use any technique for lemmatization
Input / Dataset –use any sample sentence

Theory:

Tokenization
What is Tokenization?

Tokenization is a process of converting raw data into a useful data string. Tokenization is used in NLP for splitting paragraphs and sentences into smaller chunks that can be more easily assigned meaning.

Tokenization can be done to either at word level or sentence level. If the text is split into words it is called word tokenization and the separation done for sentences is called sentence tokenization.

Why is Tokenization required?

In tokenization, process unstructured data and natural language text is broken into chunks of information that can be understood by machine.

Tokenization converts an unstructured string (text document) into a numerical data structure suitable for machine learning. This allows the machines to understand each of the words by themselves, as well as how they function in the larger text. This is especially important for larger amounts of text as it allows the machine to count the frequencies of certain words as well as where they frequently appear.

Tokenization is the first crucial step of the NLP process as it converts sentences into understandable bits of data for the program to work with. Without proper / correct tokenization, the NLP process can quickly devolve into a chaotic task.

Challenges of Tokenization

Dealing with segment words when spaces or punctuation marks define the boundaries of the word. For example: donâ€™t
Dealing with symbols that might change the meaning of the word significantly. For example: ₹100 vs 100
Contractions such as ‘you’re’ and ‘I’m’ should be properly broken down into their respective parts. An improper tokenization of the sentence can lead to misunderstandings later in the NLP process.
In languages like English or French we can separate words by using white spaces, or punctuation marks to define the boundary of the sentences. But this method is not applicable for symbol based languages like Chinese, Japanese, Korean Thai, Hindi, Urdu, Tamil, and others. Hence a common tokenization tool that combines all languages is needed.
Types of Tokenization

Word Tokenization

Most common way of tokenization, uses natural breaks, like pauses in speech or spaces in text, and splits the data into its respective words using delimiters (characters like ‘,’ or ‘;’ or ‘“,”’).
Word tokenization’s accuracy is based on the vocabulary it is trained with. Unknown words or Out Of Vocabulary (OOV) words cannot be tokenized.

White Space Tokenization

Simplest technique, Uses white spaces as basis of splitting.
Works well for languages in which the white space breaks apart the sentence into meaningful words.

Rule Based Tokenization

Uses a set of rules that are created for the specific problem.
Rules are usually based on grammar for particular language or problem.
Regular Expression Tokenizer

Type of Rule based tokenizer

Uses regular expressions to control the tokenization of text into tokens.

Penn Treebank Tokenizer

Penn Treebank is a corpus maintained by the University of Pennsylvania containing over four million and eight hundred thousand annotated words in it, all corrected by humans
Uses regular expressions to tokenize text as in Penn Treebank
Perform tokenization (Whitespace, Punctuation-based, Treebank, Tweet, Multi-Word Expression - MWE) using NLTK library.

!pip install nltk

Objective:-

The objective of the experiment is to understand the fundamental concepts and techniques of natural language processing (NLP)

Procedure:-

STEP 1: Perform tokenization (Whitespace, Punctuation-based, Treebank, Tweet, MWE) using NLTK library.

STEP 2: Use porter stemmer and snowball stemmer for stemming.

STEP 3: Use any technique for lemmatization

Input / Dataset –use any sample sentenceSelect a word root.

Questions:

1. Explain Natural Language Processing. Why is it hard?
2. Differentiate between programming languages and natural languages.
3. Are natural languages regular? Explain in detail.
4. 4. Explain the terms:
   4.1. Finite Automata for NLP
   4.2. Stages of NLP
   4.3. Challenges and issues in nlp
5. What is the concept of tokenization, stemming, lemmatization and POS tagging. Explain all terms with suitable examples.