Page No.
Date

Experiment No: Group 1 - (1)

- **Title:**

To perform tokenization (whitespace, punctuation-based, Treebank, Tweet, MWE) using NLTK Library.

- **Problem statement:**

Perform tokenization (whitespace, punctuation-based, Treebank, Tweet, MWE) using NLTK Library. use Porter stemmer and snowball stemmer for stemming use any technique for ~~min~~ lemmatization

- **objective:**

To Execute tokenization using use porter stemmer and snowballstemmer.

- **Course outcome:**

CO2: Use tools and techniques in the area of software development to build projects.

- **softwares and hardware requirements:**

sr.no	components (software/ Hardwares)	specification.
1.	Laptop / Desktop	64-bits OS (8GB Ram)
2.	Jupyter Notebook	version 7.3.3.

Page No.
Date

- Theory :

What is tokenization :

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

- Tokenization can be done 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 the machine.

- Tokenization convert an unstructured string (text document) into a numerical data structure suitable for machine learning. This allows the machines to understand each of the word by themselves, as well as how they function in the larger text.

- This is especially important for larger amounts of texts as it allows the machine to count frequency

Date

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

- Challenges of Tokenization:

1. Dealing with segment words when spaces or punctuation marks define the boundaries of the word. for example : donaEtmt.

2. Dealing with symbols that might change the meaning of the word significantly ~~for~~ for example: ₹ 100 vs 100

3. 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 misunderstanding later in the NLP process.

4. In languages like English and french we can separate words by using white space or the punctuation marks to define the boundary of the sentences. But this method is not applicable for symbol based language like chinese, Japnese, korean. Thai, Hindi, urdu, Tamil and others.

Page No.
Date

Types of tokenization:

1. Word Tokenization:

Most common way of tokenization, use natural breaks, like pause in speech or space in text, and splits the data into its respective words using delimiters (characters like " or " or " ").
word tokenization accuracy is based on the vocabulary it is trained with. unknown words or out of vocabulary (OOV) words cannot be tokenize.

2. White Space Tokenization:

Simplest techniques, uses white spaces basis of splitting.
Words well for languages in which the white space breaks apart the sentence into meaningful words.

3. Rule Based Tokenization:

uses as set of rule that are created for the specific problem. Rules are usually based on grammar for particular language or problem.

4. Regular Expression Tokenizer:

Type of Rule based tokenizer. Uses regular expression to control the tokenization of text into tokens.

Page No.
Date

stemming :

stemming is a process of reducing inflectional words to their root form. it map the words a same stem even if the stem is not a valid word in the language.

Why is stemming required :

English language has several variants of a single term. the presence of these variance in a text corpus result in data redundancy when developing NLP or machine learning model such models may be ineffective.

To Build a robust model, it is especially to normalize text by removing repetition and the transforming words to their base from through stemming.

Types of stemmer in NLTK :

1. Porter stemmer
2. snowball stemmer
3. Lancaster stemmer
4. Regexp stemmer

Lemmatization :

Lemmatization is grouping together of different forms of the same words.

Page No.
Date

- **Conclusion :**

In this practical, I have executed tokenization using use porter stemmer and the snow-ball stemmer.