ANISH ADNANI
D17B/01

EXPERIMENT 2

Aim: Apply various text processing techniques for any given text. Tokenization and filteration and script validation

Pre requisite: English morphology, grammar, finite automata

Theory: Natural language processing (NLP) is a branch of AI that helps computers to understand interpret and manipulate human language. NLP help developers to organize and structure knowledge to perform like translation, summarization, entity, recognition, relationship, extraction, speech recognition, etc.

NLP is a way of computers to analyze, understand and derive meaning from human languages like English, Hindi, Spanish, etc.

- Stages of NLP →

1) Morphological analysis: It depicts analyzing identifying and describing the structure of words. It includes dividing a text into paragraphs, words and sentences

2) Syntactic Analysis: Syntax refers to principles and rules that gives the sentence structure of any individual language

3) Semantic Analysis: It focus only on the literal

meaning of words, phrases and sentences. It abstracts the dictionary meaning from given context

4) Pragmatic Analysis: It deals with overall communication & social context and it effects on interpretation. It means abstracting or deriving the useful meaning of languages

5) Discourse Analysis: It means a sense of the context. It also considers meaning of the following sentence.

Applications

1) Machine Translation
2) Database access
3) Information retrival
4) Text categorization
5) Extracting data from text
6) Spoken language control system
7) Spelling, grammar checking

Tokenization is the process of splitting a string, text into list of tokens

Tokenization eg: Hey, How are you?
Tokens ['Hey', ',', 'How', 'are', 'you', '?']

The different steps in processing the text are

1) Tokenization
2) Filteration
3) stopword removal
4) stemming

- Filteration is the process of removing the special characters such as ',', ';', ':', '!' and to remove the words of different languages.

Eg: ~~ARE~~ ! How are you?

After tokenization - ['~~ARE~~', '!', 'How', 'are', 'you', '?']
Filteration : ['How', 'are', 'you']

- stopword Removal : In this step all stopwords like 'is', 'are', 'the', etc are removed

stopword removal of the above example will result into ['How', 'you']

- Tokenization is essential splitting a phrase, sentence, paragraph, entire text document into smaller units such as individual words or terms, each of these smaller units are called tokens.

- A type is the class of all tokens containing the same character sequence

Filtration: Tokenization can remove punctuation too, easing the path to a proper word segmentation but also triggering possible complications. In the case of periods that follows abbreviation 'eg-dr' the period following the abbreviation should be considered as part of the same token and not be removed. Tokenization process can be particularly problematic when dealing with biomedical text domains which contain hyphens, parentheses and other punctuation marks

Conclusion: Hence we have successfully implemented preprocessing with tokenization, filtration and script validation.

OUTPUT

Word Tokenize

```
import nltk
from nltk.tokenize import word_tokenize
text = "Hello, my name is Anish Adnani."
print(word_tokenize(text))
```

```
['Hello', ',', 'my', 'name', 'is', 'Anish', 'Adnani', '.']
```

Sentence Tokenize

```
from nltk.tokenize import sent_tokenize
text = "Hello, my name is Anish Adnani. I study in VESIT."
print(sent_tokenize(text))
```

```
['Hello, my name is Anish Adnani.', 'I study in VESIT.']
```

Special Character Elimination

```
import re
text = "Hello, I {Anish Adnani} is asking a ?"
print(re.sub('[^A-Za-z0-9]+',' ',text))
```

```
Hello I Anish Adnani is asking a
```

Trying out in Hindi language

Word Tokenization

```
import nltk
from nltk.tokenize import word_tokenize
text = "नमस्कार, मेरा नाम अनीश अदनानी है।"
print(word_tokenize(text))
```

```
['नमस्कार', ',', 'मेरा', 'नाम', 'अनीश', 'अदनानी', 'है।']
```

Word Tokenization using INLTK Library

```
from inltk.inltk import setup
setup('hi')
```

```
from inltk.inltk import tokenize

hindi_text = """मेरा नाम अनिश अदनानी है, मैं वर्तमान में विवेकानंद प्रौद्योगिकी संस्थान में कंप्यूटर इंजीनियरिंग की पढ़ाई कर रहा हूँ।"""

# tokenize(input text, language code)
tokenize(hindi_text, "hi")
```

```
['_मेरा',
 '_नाम',
 '_अ',
 'निश',
 '_अद',
 'नानी',
 '_है',
 ',',
 '_मैं',
 '_वर्तमान',
 '_में',
 '_विवेकानंद',
 '_प्रौद्योगिकी',
 '_संस्थान',
 '_में',
 '_कंप्यूटर',
 '_इंजीनियरिंग',
 '_की',
 '_पढ़ाई',
 '_कर',
 '_रहा',
 '_हूँ',
 '।']
```