Alok Bhawankar
PD09

Assignment 1

* Problem statement:

Program to read a paragraph from text File. Print the paragraph after removing the stop words. Identify part of speech of each words in the paragraph. Use NLTK

* objectivs:

1. To study and explore NLTX for text processing
2. To learn concepts of text processing in NLP

* Theory:

Q Explain Following concepts:

1. Text Processing concept:-

- Tokenization:- It is the process of breaking the given text i.e. the character sequence into smaller units called tokens. The tokens may be the words, numbers or punctuation marks. It is also called word Segmentation. the ending of a word and the begining of new word are called word boundries

- Stemming:- A lot of variations of words due to grammatical reasons. The concept of variations here means that to deal with different forms of the same words like democracy, democratic and demotradization.

- Lemmatization! Also extract the base form of words by lemmatization. Task with the use of a vocabulary and morphological analysis of words.

- POS Tagging - A POS Tag is a special label assigned to each token (word) in a text corpus to indicate the part of speech and often also other grammatical categories such as tense, number (plural / singular). case, etc.

- Stop word removal:- All stop words for example, common words, such as 'a' and 'the' are removed from multiple word queries to increase search performance.

- Bag of words (BoW) model, n-grams:

→ whenever we apply any algorithm in NLP, it works on number, we cannot directly feed our text into that algorithm. Hence, Bag of words model is used to preprocess the text by converting it into a bag of words, which keeps a count of the total occurences of most frequently used words. In this simple model, the syntax and even the order of words is ignored.

An n-gram is a contiguous sequence of n words, for example, in the sentence dog that barks does not bite, the n grams are

- unigrams (n=1): dog, that, barks, door, not, bite!
- bigrams (n=2): dog that, that barks, barks dog, doesnot, not bite.
- trigrams (n=3): dog that barks, that barks does, barks does not, does not bite

☆ Algorithm / Implementation:

1. Read a text file in python using read & open function.
2. Tokenize the file into sentences.
3. Tokenize each sentence in words and punctuation,
4. Remove the stopwords ('a', 'an', 'the', 'to')
5. Tag each word to indicate its part of speech

☆ Platform: 64 bit Linux, Jupyter Notebook.

☆ Input: Any text/doc file containing text paragraph in English Language.

☆ Output: Tokens, text after removing stop words tokens with POS tagging, stem form of text.

☆ Conclusion:- Hence learned the concepts of text processing in NLP and implemented using NLTK library.

☆ FAQ

1. Explain the difference stemming and lemmatization ?

=>

Stemming	Lemmatization
• Stemming is a process of reducing words to its root form even if the root has no dictionary meaning.	• Lemmatization. is a morphological analysis of a word is a normalized form of a set of morphologically related forms chosen by convention to represent that set.
• eg:- beautiful and beautifully will be stemmed to beauti which has no meaning in english dictionary.	eg:- beautiful and beautifully will be lemmatised to beautiful and beautifully respectively without changing the meaning of the words.

q2. What is sematic and syntactic analysis in NLP ?

→ semantic and syntactic analysis in NLP are two primary technique to understand natural language. Language is a set of value sentences, but only proper syntax and semantic can make a sentence valid.