Day22_jithin_Introduction to NLP

August 31, 2018

1 Introduction to NLP

We will use spacy library to do pos tagging, stemming and tokenization.

Exercise 1: Import spacy and load the module by using [nlp =spacy.load('en_core_web_sm')]

Exercise 2: Extract the data from the file War and peace and select a small range of data (Say 10 lists)

Exercise 3: Pass this data for tokenizing, stemming and POS tagging

Exercise 4: Pass this data to get a more detailed POS tag

2 Using NLTK Library

```
In [89]: import nltk
         from nltk.tokenize import sent_tokenize, word_tokenize
         from nltk.stem import PorterStemmer
```

```
In [90]: with open('War_And_Peace.txt', 'r') as myfile:
             doc = myfile.read()
```

```
In [91]: print("Type of the doc variable is :",type(doc)," and length is : ",len(doc))
         doc=doc.replace('\n'," ")
```

```
Type of the doc variable is : <class 'str'>  and length is :  3227580
```

```
In [92]: doc=doc[10000:15000]
```

```
In [93]: print("Type of the doc variable is :",type(doc)," and length is : ",len(doc))
         doc=doc.replace('\n'," ")
```

```
Type of the doc variable is : <class 'str'>  and length is :  5000
```

2.1 Tokenization

```
In [94]: # Sentence Tokenizer
         sent=sent_tokenize(doc)
         len(sent)
```