**NLP Assignment**

text='''

Look for data to help you address the question. Governments are good

sources because data from public research is often freely available. Good

places to start include http://www.data.gov/, and http://www.science.

gov/, and in the United Kingdom, http://data.gov.uk/.

Two of my favorite data sets are the General Social Survey at http://www3.norc.org/gss+website/,

and the European Social Survey at http://www.europeansocialsurvey.org/.

Using tokenization , Extract all money transaction from below sentence along with currency. Output should be,

wo $

500 €

2.

1.Use stemming for following docs

doc = nlp("Mando talked for 3 hours although talking isn't his thing")

doc = nlp("eating eats eat ate adjustable rafting ability meeting better")

2. convert these list of words into base form using Stemming and Lemmatization and observe the transformations

#using stemming in nltk

lst\_words = ['running', 'painting', 'walking', 'dressing', 'likely', 'children', 'whom', 'good', 'ate', 'fishing']

#using lemmatization in spacy

doc = nlp("running painting walking dressing likely children who good ate fishing")

3.convert the given text into it's base form using both stemming and lemmatization

text = """Latha is very multi talented girl.She is good at many skills like dancing, running, singing, playing.She also likes eating Pav Bhagi. she has a

habit of fishing and swimming too.Besides all this, she is a wonderful at cooking too.

"""

3.You are parsing a news story from cnbc.com. News story is stores in [news\_story.txt](https://github.com/codebasics/nlp-tutorials/blob/main/7_pos/news_story.txt) which is on whatsapp. You need to,

1. Extract all NOUN tokens from this story. You will have to read the file in python first to collect all the text and then extract NOUNs in a python list
2. Extract all numbers (NUM POS type) in a python list
3. Print a count of all POS tags in this story

### **Named Entity Recognition (NER): Exercises**

1.Extract all the Geographical (cities, Countries, states) names from a given text

text = """Kiran want to know the famous foods in each state of India. So, he opened Google and search for this question. Google showed that

in Delhi it is Chaat, in Gujarat it is Dal Dhokli, in Tamilnadu it is Pongal, in Andhrapradesh it is Biryani, in Assam it is Papaya Khar,

in Bihar it is Litti Chowkha and so on for all other states"""

**Expected Output:**

Geographical location Names: [India, Delhi, Gujarat, Tamilnadu, Andhrapradesh, Assam, Bihar]

#### 2.Excersie: 2

* Extract all the birth dates of cricketers in the given Text
* text = """Sachin Tendulkar was born on 24 April 1973, Virat Kholi was born on 5 November 1988, Dhoni was born on 7 July 1981
* and finally Ricky ponting was born on 19 December 1974."""
* **Expected Output:**
* All Birth Dates: [24 April 1973, 5 November 1988, 7 July 1981, 19 December 1974]

### Bag of words: Exercises

* In this Exercise, you are going to classify whether a given movie review is **positive or negative**.
* you are going to use Bag of words for pre-processing the text and apply different classification algorithms.
* Sklearn CountVectorizer has the inbuilt implementations for Bag of Words.
* #Import necessary libraries
* import pandas as pd
* import numpy as np
* from sklearn.model\_selection import train\_test\_split
* from sklearn.feature\_extraction.text import CountVectorizer
* from sklearn.naive\_bayes import MultinomialNB
* from sklearn.pipeline import Pipeline
* from sklearn.metrics import classification\_report