ROLL NO – 162010022

PRACTICAL – 5

NAMED ENTITY RECOGNITION (NER)

AIM:- To identifying people, organizations, news topics from the data set.

THEORY:-

NER is a subtask of information extraction that seeks to locate and classify named entity mentions in unstructured text into predefined categories such as the person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.

To work on the given dataset we made use of Stanford NER Tagger

CODE:-

```
# coding: utf-8

from nltk.tag import StanfordNERTagger
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
import os
import pandas as pd
import re

def removeStopWords(string):
    stopWords = set(stopwords.words("english"))
    tokens = word_tokenize(string)
    finalStr = []
    for word in tokens:
        if word.lower() not in stopWords:
            finalStr.append(word)
        #else:
        #    print word
    return ' '.join(finalStr)

def removeInvalidChars(string):
    if pd.isnull(string):
        return None
    if "₹" in string:
        string = string.replace("₹","Rs. ")
    newStr = re.sub("[^A-Za-z0-9-' ]+", " ", string)
    newStr = newStr.encode("utf-8")
    #print "\n\n\nYo\n\n\n"
    return newStr

os.environ['CLASSPATH'] = "/home/pearl/stanford/stanford-ner.jar"

os.environ['STANFORD_MODELS'] = "/home/pearl/stanford/stanford-classifier-2018-02-27"
```

```
os.environ['JAVA_HOME'] = "/usr/lib/jvm/jdk/"

st = StanfordNERTagger('english.all.3class.distsim.crf.ser.gz',encoding='utf-8')

dF =
pd.read_csv("/home/pearl/Desktop/hadoop/Java/Bigdata/data/newWD/rediff_realtime_news_201710_201712",
delimiter='\t',nrows=20)
print dF['summary']
dF.fillna('',inplace=True)
#print dF
dF['summary'] = dF.apply(lambda row: removeInvalidChars(row['summary']),axis = 1)
dF['summary'] = dF.apply(lambda row: removeStopWords(row['summary']),axis = 1)

loc_str = ""
person_str = ""
organization_str = ""

for row in dF.itertuples():
    tokenized_text = word_tokenize(row.summary)
    classified_text = st.tag(tokenized_text)
    for i in range(len(classified_text)):
        if classified_text[i][1]=="LOCATION":
            loc_str+=classified_text[i][0]+" , "
        elif classified_text[i][1]=="PERSON":
            person_str+=classified_text[i][0]+" , "
        elif classified_text[i][1]=="ORGANIZATION":
            organization_str+=classified_text[i][0]+" , "

print("\nLocations : "+loc_str)
print("\nPersons :" +person_str)
print("\nOrganizations : "+organization_str)
```

```
In [3]: %run -i 'NER.py'
0     BARCELONA (Reuters) - Thousands of demonstrato...
1     SAN JUAN, Puerto Rico (Reuters) - U.S. Preside...
2     ABOARD THE USS RONALD REAGAN, South China Sea ...
3     BARCELONA (Reuters) - Spanish police monitored...
4     BAGHDAD (Reuters) - Iraqi Prime Minister Haide...
5     A round-up of the top stories from the pharma ...
6     F1 2017: Lewis Hamilton claims pole position a...
7     MAYAGUEZ, Puerto Rico (Reuters) - Few people i...
8     DUBLIN (Reuters) - Tens of thousands of people...
9     Home Minister Rajnath Singh said on Saturday t...
10    The minister had acknowledged that Saeed, the ...
11    Xi met US Secretary of State Rex Tillerson her...
12    The PBOC has already started monitoring and an...
13    The Border Security Force (BSF) on Saturday un...
14    The issue was discussed during a meeting betwe...
15    "After inquiries in the recent incidences of v...
16    Tata Motors said its Chief Financial Officer (...
17    2017 INMRC: TVS Racing's Jagan Kumar moves int...
18    The People#39;s Bank of China (PBOC) said on i...
19    The European Commission has shied away from ra...
Name: summary, dtype: object

Locations : Spanish , Spanish , Catalonia , Asia , North , Korea , China , Spanish , Spain , BAGHDAD , Iraqi , Baghdad ,
MAYAGUEZ , Puerto , Rico , DUBLIN , Dublin , China , India , US , Beijing , Pyongyang , Jammu , Europe ,

Persons :Donald , Trump , Haider , al-Abadi , Lewis , Hamilton , Rajnath , Singh , Saeed , Haqqanis , Lashkar-e-Taiba ,
Rex , Tillerson , Suresh , Prabhu , Selangor , Dato , Seri , Mohamed , Azmin , Bin , Ali , Ramakrishnan , Jagan , Kumar
,

Organizations : BARCELONA , Reuters , Barcelona , SAN , JUAN , Puerto , Rico , Reuters , Hurricane , Maria , RONALD , RE
AGAN , South , China , Sea , Reuters , BARCELONA , Reuters , Reuters , Kurdistan , Puerto , Rico , Reuters , Reuters , I
TBP , PBOC , Xinhua , Border , Security , Force , BSF , Sangh , Tata , Motors , People , 39 , Bank , China , PBOC , Rese
rve , Requirement , Ratio , RRR , European , Commission ,
```