Name Entity Recognition

- sentiment analysis
- formulate a relationship that can be used to find the semantic of the sentence (meaning)

1. Develop a Named Entity Recognition model to identify locations

2. Evaluate the model accuracy on a standard NER Datasets

In [17]:

```
import nltk
```

In [18]:

```
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.tag import pos_tag
from nltk.chunk import conllstr2tree, tree2conlltags
```

In [19]:

```
nltk.download('maxent_ne_chunker')
nltk.download('words')
```

```
[nltk_data] Downloading package maxent_ne_chunker to
[nltk_data]     C:\Users\ahana\AppData\Roaming\nltk_data...
[nltk_data]   Package maxent_ne_chunker is already up-to-date!
[nltk_data] Downloading package words to
[nltk_data]     C:\Users\ahana\AppData\Roaming\nltk_data...
[nltk_data]   Package words is already up-to-date!
```

Out[19]:

```
True
```

In [20]:

```
def preprocess(text):
    #tokenization
    tokens = word_tokenize(text)

    #remove stop words and punctuation
    stop_words = set(stopwords.words('english'))
    filtered_tokens = [token for token in tokens if token.lower() not in stop_words and token.isalpha()]
    return filtered_tokens

def extract_entities(text):
    #Tokenize and preprocess text
    tokens = preprocess(text)

    #Perform POS Tagging
    tagged_tokens = pos_tag(tokens)

    #Perform NER
    ne_tree = nltk.ne_chunk(tagged_tokens)

    #Convert the tree to IOB (Inside Outside Beginning) tags
    iob_tags = tree2conlltags(ne_tree)

    return iob_tags

#Sample Text
text = "Narendra Modi was born in India. He is the prime minister of India"

#Example Usage
entities = extract_entities(text)
print("Named Entites: ")
for word, pos_tag, entity_tag in entities:
    if entity_tag != 0:
        print(f"Word: {word}, POS Tag: {pos_tag}, Entity Tag: {entity_tag}")
```

```
Named Entites:
Word: Narendra, POS Tag: NNP, Entity Tag: B-PERSON
Word: Modi, POS Tag: NNP, Entity Tag: B-ORGANIZATION
Word: born, POS Tag: IN, Entity Tag: O
Word: India, POS Tag: NNP, Entity Tag: B-GPE
Word: prime, POS Tag: JJ, Entity Tag: O
Word: minister, POS Tag: NN, Entity Tag: O
Word: India, POS Tag: NNP, Entity Tag: B-GPE
```

In [21]:
```
!python -m spacy download en_core_web_sm
```

```
Collecting en-core-web-sm==3.6.0
  Downloading https://github.com/explosion/spacy-models/releases/download/en_core_web_sm-3.6.0/en_core_web_sm-3.6.0-py3-none-any.whl (h
ttps://github.com/explosion/spacy-models/releases/download/en_core_web_sm-3.6.0/en_core_web_sm-3.6.0-py3-none-any.whl) (12.8 MB)
     ---------------------------------------- 12.8/12.8 MB 2.2 MB/s eta 0:00:00
Requirement already satisfied: spacy<3.7.0,>=3.6.0 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from en
-core-web-sm==3.6.0) (3.6.1)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages
(from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (3.0.12)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages
(from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (1.0.5)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (f
rom spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (1.0.10)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from sp
acy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.0.8)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from
spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (3.0.9)
Requirement already satisfied: thinc<8.2.0,>=8.1.8 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from sp
acy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (8.1.12)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from s
pacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (1.1.2)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from sp
acy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.4.8)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (fro
m spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.0.10)
Requirement already satisfied: typer<0.10.0,>=0.3.0 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from s
pacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (0.9.0)
Requirement already satisfied: pathy>=0.10.0 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from spacy<3.
7.0,>=3.6.0->en-core-web-sm==3.6.0) (0.10.2)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (fr
om spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (6.4.0)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from sp
acy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (4.65.0)
Requirement already satisfied: numpy>=1.15.0 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from spacy<3.
7.0,>=3.6.0->en-core-web-sm==3.6.0) (1.24.1)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (fro
m spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.28.2)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-
packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.4.1)
Requirement already satisfied: jinja2 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from spacy<3.7.0,>=
3.6.0->en-core-web-sm==3.6.0) (3.1.2)
Requirement already satisfied: setuptools in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from spacy<3.7.
0,>=3.6.0->en-core-web-sm==3.6.0) (65.5.0)
Requirement already satisfied: packaging>=20.0 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from spacy<
3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (23.0)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (fro
m spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (3.3.0)
Requirement already satisfied: annotated-types>=0.4.0 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from
pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (0.5.0)
Requirement already satisfied: pydantic-core==2.10.1 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from
pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.10.1)
Requirement already satisfied: typing-extensions>=4.6.1 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (fr
om pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (4.8.0)
Requirement already satisfied: charset-normalizer<4,>=2 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (fr
om requests<3.0.0,>=2.13.0->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (3.0.1)
Requirement already satisfied: idna<4,>=2.5 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from requests<
3.0.0,>=2.13.0->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (3.4)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from
requests<3.0.0,>=2.13.0->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (1.26.14)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from req
uests<3.0.0,>=2.13.0->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2022.12.7)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from thi
nc<8.2.0,>=8.1.8->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (0.7.11)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (fr
om thinc<8.2.0,>=8.1.8->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (0.1.3)
Requirement already satisfied: colorama in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from tqdm<5.0.0,>=
4.38.0->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (0.4.6)
Requirement already satisfied: click<9.0.0,>=7.1.1 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from ty
per<0.10.0,>=0.3.0->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (8.1.7)
Requirement already satisfied: MarkupSafe>=2.0 in c:\users\ahana\appdata\local\programs\python\python311\lib\site-packages (from jinja2
->spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (2.1.2)
[+] Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')


[notice] A new release of pip available: 22.3.1 -> 24.0
[notice] To update, run: python.exe -m pip install --upgrade pip
```

In [22]:
```
import pandas as pd
import spacy
import requests
from bs4 import BeautifulSoup
nlp = spacy.load("en_core_web_sm")
pd.set_option("display.max_rows", 200)
```

In [23]:
```
content = "Senior Congress leader Rahul Gandhi on Thursday claimed that Prime Minister Narendra Modi was not born into an Other Backward
doc = nlp(content)
for ent in doc.ents:
    print(ent.text,ent.start_char,ent.end_char,ent.label_)
```

```
Congress 7 15 ORG
Rahul Gandhi 23 35 PERSON
Thursday 39 47 DATE
Narendra Modi 76 89 PERSON
```

In [24]:

```
from spacy import displacy
displacy.render(doc,style="ent")
```

Senior Congress **ORG** leader Rahul Gandhi **PERSON** on Thursday **DATE** claimed that Prime Minister Narendra Modi **PERSON** was not born into an Other Backward Class family, and he was “misleading” people by identifying himself as an OBC.

Question 2

In [29]:

```
import nltk
from nltk.chunk import ne_chunk
from nltk.chunk.util import tree2conlltags
from nltk.corpus import conll2002
from sklearn.metrics import accuracy_score
```

In [30]:

```
def conll2002_data():
    train_sents = list(conll2002.iob_sents('esp.train'))
    test_sents = list(conll2002.iob_sents('esp.testb'))
    return train_sents,test_sents
```

In [32]:

```
nltk.download('conll2002')

def evaluate_ner(train_sents,test_sents):
    chunking_rule = r'''
    NP: {<DT|JJ|NN.*>+}
    PP: {<IN><NP>}
    VP: {<VB.*><NP|PP|CLAUSE>+$}
    CLAUSE: {<NP><VP>}
    '''

    chunker = nltk.RegexpParser(chunking_rule)
    parsed_test_sents = [chunker.parse(sent) for sent in test_sents]
    predicted_labels = []
    true_labels = []

    for parsed_sent , test_sent in zip(parsed_test_sents,test_sents):
        iob_tags = tree2conlltags(parsed_sent)
        predicted_labels.extend([tag for word,pos,tag in iob_tags])
        true_labels.extend([tag for word,pos,tag in test_sent])

    accuracy = accuracy_score(true_labels,predicted_labels)

    return accuracy

train_data , test_data = conll2002_data()
accuracy = evaluate_ner(train_data,test_data)
print('NER MODEL ACCURACY = ',accuracy)
```

```
[nltk_data] Downloading package conll2002 to
[nltk_data]     C:\Users\ahana\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\conll2002.zip.
```

```
NER MODEL ACCURACY =  0.8800186288397726
```

Assignments:

1. Use CRF for NER
2. Modify te NER to recognise Hindi and Hinglish

In [ ]:

```
```