NAME :- MOHAN SAI KOTHAPALLI

ID :- 2112951

CODE :- GN22CDBDS001

ASSIGNMENT

CREATE AN NLP PIPELINE

IMPORTING LIBRARIES

In [1]:
```
import spacy
from spacy import displacy
import pandas as pd
from urllib.parse import quote
```

In [2]:
```
nlp = spacy.load("en_core_web_sm")
file=open('NLPPipeline.txt',encoding="mbcs").read()
of=nlp(file)
```

STEP 1:- Sentence Segmentation

In [3]:
```
sent=of.sents
print(sent)
```

```
<generator object at 0x00000239353A8CC8>
```

In [4]:
```
for se in sent:
    print(se)
```

```
London is the capital and most populous city of England and the United Kingdom.
Standing on the River Thames in the south east of the island of Great Britain, London has been a majo
r settlement for two millennia.
It was founded by the Romans, who named it Londinium.
```

In [5]:
```
sent1=list(of.sents)
print(sent1)
print(len(sent1))
```

```
[London is the capital and most populous city of England and the United Kingdom., Standing on the Riv
er Thames in the south east of the island of Great Britain, London has been a major settlement for tw
o millennia., It was founded by the Romans, who named it Londinium.]
3
```

STEP 2 :- Word Tokenization

In [6]:
```
print([token.text for token in of])
```

```
['London', 'is', 'the', 'capital', 'and', 'most', 'populous', 'city', 'of', 'England', 'and', 'the',
'United', 'Kingdom', '.', 'Standing', 'on', 'the', 'River', 'Thames', 'in', 'the', 'south', 'east',
'of', 'the', 'island', 'of', 'Great', 'Britain', ',', 'London', 'has', 'been', 'a', 'major', 'settlem
ent', 'for', 'two', 'millennia', '.', 'It', 'was', 'founded', 'by', 'the', 'Romans', ',', 'who', 'nam
ed', 'it', 'Londinium', '.']
```

STEP 3:- Predicting Parts Of Speech for each token

In [7]:
```
l=[]
for token in of:
    r=token.text,token.pos_,token.tag_,spacy.explain(token.pos_),spacy.explain(token.tag_)
    l.append(r)
df=pd.DataFrame(l)
print(df.head(4))
```

```
         0      1    2            3                                4
0   London  PROPN  NNP  proper noun                 noun, proper singular
1       is    AUX  VBZ    auxiliary  verb, 3rd person singular present
2      the    DET   DT   determiner                       determiner
3  capital   NOUN   NN         noun                 noun, singular or mass
```

STEP 4:- Text Lemmatization

In [8]:
```
ll=[]
sen=[]
for token in of:
    r=token.text,token.lemma_
    ll.append(r)
    sen.append(r[1])
sen=' '.join(sen)
df=pd.DataFrame(ll,columns=['Token','Lemma'])
print(df.head(5))
print()
print(sen)
```

```
     Token    Lemma
0   London   London
1       is       be
2      the      the
3  capital  capital
4      and      and

London be the capital and most populous city of England and the United Kingdom . stand on the River T
hames in the south east of the island of Great Britain , London have be a major settlement for two mi
llennia . it be found by the Romans , who name it Londinium .
```

STEP :-5 Remove Stop Words

In [9]:
```
twsw=[]
stopwords = nlp.Defaults.stop_words
#print(stopwords)

for token in nlp(sen):
    if token.text not in stopwords:
        twsw.append(token.text)
print(twsw)
```

```
['London', 'capital', 'populous', 'city', 'England', 'United', 'Kingdom', '.', 'stand', 'River', 'Tha
mes', 'south', 'east', 'island', 'Great', 'Britain', ',', 'London', 'major', 'settlement', 'millenni
a', '.', 'found', 'Romans', ',', 'Londinium', '.']
```

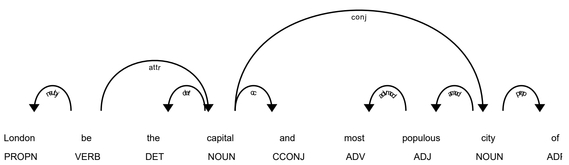
In [10]:
```
sentence=" ".join(twsw)
print(sentence)
```

```
London capital populous city England United Kingdom . stand River Thames south east island Great Brit
ain , London major settlement millennia . found Romans , Londinium .
```

STEP :-6a Dependency Parsing

In [11]:
```
tokens = nlp(sen)
```

In [12]:
```
displacy.render(tokens, style='dep', jupyter=True, options={'distance':100})
```



STEP :-6b Finding Noun Phrases

In [13]:
```
for np in tokens.noun_chunks:
    print(np.text)
```

```
London
the capital
most populous city
England
the United Kingdom
the River Thames
the south east
the island
Great Britain
London
a major settlement
two millennia
it
the Romans
who
it
Londinium
```

STEP :-7 Named Entity Recognition (NER)

In [14]:
```
displacy.render(tokens, style="ent")
```

London GPE be the capital and most populous city of England GPE and the United Kingdom GPE . stand on the River Thames in the south east of the island of Great Britain GPE , London GPE have be a major settlement for two CARDINAL millennia . it be found by the Romans NORP , who name it Londinium ORG .

STEP :-8 Coreference Resolution

In [15]:
```
text ='London be the capital and most populous city of England and the United Kingdom. stand on the Riv
er Thames in the south east of the island of Great Britain, London have be a major settlement for two m
illennia. it be found by the Romans, who name it Londinium.'
url = "https://huggingface.co/coref/?text=" + quote(text)
print("Check the LINK for visualizing Coreference Resolution:\n", url)
```

```
Check the LINK for visualizing Coreference Resolution:
 https://huggingface.co/coref/?text=London%20be%20the%20capital%20and%20most%20populous%20city%20of%2
0England%20and%20the%20United%20Kingdom.%20stand%20on%20the%20River%20Thames%20in%20the%20south%20eas
t%20of%20the%20island%20of%20Great%20Britain%2C%20London%20have%20be%20a%20major%20settlement%20for%2
0two%20millennia.%20it%20be%20found%20by%20the%20Romans%2C%20who%20name%20it%20Londinium.
```