

Name: Harsh Zanwar Roll No.: 73 Branch:CSE(AIML)

Aim: Named Entity Recognition and Dependency Parsing for Information Extration using spacy. Consider any text file (research article technical blog , any unstructured corpus used before) Perform NER to extract entities from individual sentences using spacy. use Dependency parsing, POS tagging to extract relationships between the entities. Create a tuple for information extraction T1(Entity1, Entity2, Relation label) Display no of such tuples extracted from considered corpus as extracted information

```
In [1]: import spacy
import nltk
nltk.download('genesis')

# Load the Genesis corpus
from nltk.corpus import genesis
genesis_text = genesis.raw()

# Load the spaCy English model
nlp = spacy.load("en_core_web_sm")

# Increase the maximum input length limit
nlp.max_length = 2000000

# Define a function to extract relationships between entities using dependency parsing
def extract_relations(doc):
    relations = []
    for token in doc:
        if token.dep_ == "nsubj": # Subject of the verb
            subject = token.text
            for child in token.children:
                if child.dep_ == "relcl": # Relative clause modifier
                    verb = child.text
                    for grandchild in child.children:
                        if grandchild.dep_ == "dobj": # Direct object of the verb
                            object = grandchild.text
                            relations.append((subject, object, verb))
    return relations

# Process the Genesis text with spaCy
doc = nlp(genesis_text[:10000]) # Limit to first 1 million characters
sentences = [sent for sent in doc.sents][:100] # Limit to first 100 sentences

# Extract named entities from the individual sentences
entities = set()
for sent in sentences:
    for ent in sent.ents:
        entities.add(ent.text)

# Print the named entities
print("Named Entities:", ", ".join(entities))

# Extract relationships between the named entities
relations = []
for sent in sentences:
    relations += extract_relations(sent)
```

```
# Print the extracted relationships as tuples for information extraction
print("Tuples for Information Extraction:")
for relation in relations:
    print(relation)

# Print the number of tuples extracted
print(f"Number of Tuples Extracted: {len(relations)}")
```

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

```
Named Entities: first, second, the sixth day, Euphrates, Thou, the seventh day, Ethio
pia, the fifth day, the second day, the fourth day, Havilah, Ye, Earth, Eden, one, Sp
irit, Se, Adam, the
day, the seventh day God, Gihon, two, thou, Woman, Night, the night, Unto Adam, Eve,
Behold, four, the first day, fourth, the day, Pison, the third day, earth, Assyria, t
he light Day, the
morning, compasseth, third, days, seasons, thou shalt
Tuples for Information Extraction:
('moveth', 'which', 'brought')
('rib', 'which', 'taken')
Number of Tuples Extracted: 2
```

In []:

In [2]: `nlp=spacy.load('en_core_web_sm')`

In [3]: `kalam = """A. P. J. Abdul Kalam was an Indian aerospace scientist and politician who s`
`kalam`

Out[3]: "A. P. J. Abdul Kalam was an Indian aerospace scientist and politician who served as the 11th President of India from 2002 to 2007. He was born and raised in Rameswaram, Tamil Nadu and studied physics and aerospace engineering. He spent the next four decades as a scientist and science administrator, mainly at the Defence Research and Development Organisation (DRDO) and Indian Space Research Organisation (ISRO) and was intimately involved in India's civilian space programme and military missile development efforts."

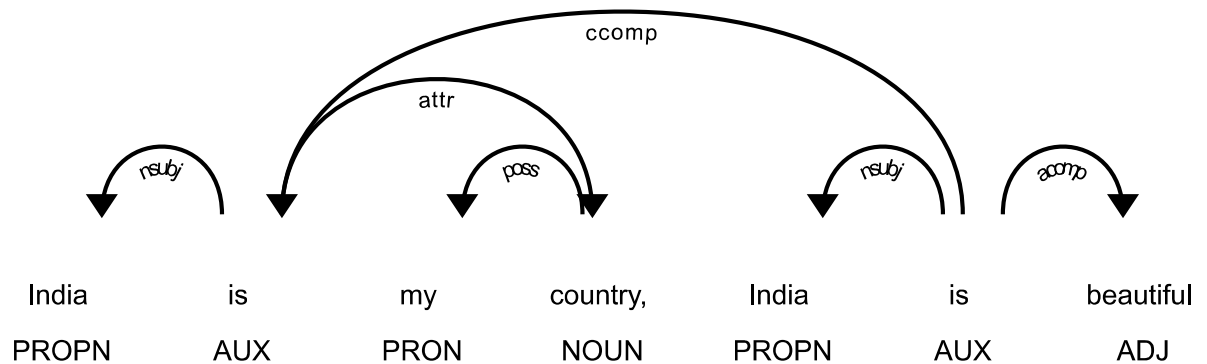
In [5]: `kalam_nlp_model = nlp(kalam)`

In [7]: `print(f"{'Entity'.ljust(50)} {'Label'}" + '\n' + '-' * 50)`
`for entity in kalam_nlp_model.ents:`
 `print(entity.text.ljust(50), entity.label_)`

Entity	Label
A. P. J. Abdul Kalam	PERSON
Indian	NORP
11th	ORDINAL
India	GPE
2002	DATE
2007	DATE
Rameswaram	GPE
Tamil Nadu	PERSON
the next four decades	DATE
the Defence Research and Development Organisation	ORG
Indian Space Research Organisation	ORG
India	GPE

```
In [8]: dep_nlp = nlp('India is my country, India is beautiful')
```

```
In [9]: spacy.displacy.render(dep_nlp, style='dep', jupyter=True, options={'distance': 90})
```



```
In [10]: import spacy
from spacy import displacy

text = "When Sebastian Thrun started working on self-driving cars at Google in 2007, f

nlp = spacy.load("en_core_web_sm")
doc = nlp(text)
displacy.render(doc, style="ent")
```

When Sebastian Thrun **PERSON** started working on self-driving cars at Google **ORG** in 2007 **DATE**, few people outside of the company took him seriously.

```
In [11]: text = """
Shivaji Bhosle was born on February 19, 1630 to Shahaji Bhosle and Jijabai in the fort

Shivaji turned out to be a born leader from a very young age. An active outdoorsman, h
"""
```

```
In [12]: text_nlp = nlp(text)
```

```
In [13]: print(f"{'Entity'.ljust(50)} {'Label'}" + '\n' + '-' * 50)
         for entity in text_nlp.ents:
             print(entity.text.ljust(50), entity.label_)
```

Entity	Label
Shivaji Bhosle	PERSON
February 19, 1630	DATE
Shahaji Bhosle	GPE
Jijabai	GPE
Shivneri	ORG
Junnar	GPE
Pune	ORG
Shahaji	PERSON
Bijapuri Sultanate	NORP
Bijapur	PERSON
Ahmednagar	ORG
Golconda	ORG
Jaigirdari	GPE
Pune	GPE
Jijabai	GPE
Sindkhed	ORG
Lakhujirao Jadhav	PERSON
Shivaji	PERSON
Shahaji	NORP
Pune	ORG
Shivaji	PERSON
Balkrishna Pant	PERSON
Raghunath Ballal	PERSON
Dabir	PERSON
Dadoji Konddeo	PERSON
Kanhoji Jedhe	PERSON
Baji Pasalkar	PERSON
Shivaji	PERSON
Shivaji	PERSON
Saibai Nimbalkar	PERSON
1640	DATE
Shivneri	NORP
15	DATE
Maval	GPE

```
In [14]: entity_relations = [(i, i.label_, i.label) for i in text_nlp.ents]

         entity_relations
```

```
Out[14]: [(Shivaji Bhosle, 'PERSON', 380),
          (February 19, 1630, 'DATE', 391),
          (Shahaji Bhosle, 'GPE', 384),
          (Jijabai, 'GPE', 384),
          (Shivneri, 'ORG', 383),
          (Junnar, 'GPE', 384),
          (Pune, 'ORG', 383),
          (Shahaji, 'PERSON', 380),
          (Bijapuri Sultanate, 'NORP', 381),
          (Bijapur, 'PERSON', 380),
          (Ahmednagar, 'ORG', 383),
          (Golconda, 'ORG', 383),
          (Jaigirdari, 'GPE', 384),
          (Pune, 'GPE', 384),
          (Jijabai, 'GPE', 384),
          (Sindkhed, 'ORG', 383),
          (Lakhujirao Jadhav, 'PERSON', 380),
          (Shivaji, 'PERSON', 380),
          (Shahaji, 'NORP', 381),
          (Pune, 'ORG', 383),
          (Shivaji, 'PERSON', 380),
          (Balkrishna Pant, 'PERSON', 380),
          (Raghunath Ballal, 'PERSON', 380),
          (Dabir, 'PERSON', 380),
          (Dadoji Konddeo, 'PERSON', 380),
          (Kanhoji Jedhe, 'PERSON', 380),
          (Baji Pasalkar, 'PERSON', 380),
          (Shivaji, 'PERSON', 380),
          (Shivaji, 'PERSON', 380),
          (Saibai Nimbalkar, 'PERSON', 380),
          (1640, 'DATE', 391),
          (Shivneri, 'NORP', 381),
          (15, 'DATE', 391),
          (Maval, 'GPE', 384)]
```

```
In [15]: from spacy import displacy

options = {"compact": True, "bg": "#09a3d5", "color": "white", "font": "Source Sans Pr

displacy.render(text_nlp, style="ent", jupyter=True, options=options)
```

Shivaji Bhosle **PERSON** was born on February 19, 1630 **DATE** to Shahaji Bhosle **GPE** and Jijabai **GPE** in the fort of Shivneri **ORG**, near the city of Junnar **GPE** of the Pune **ORG** district. Shivaji's father Shahaji **PERSON** was in service of the Bijapuri Sultanate **NORP** - a tripartite association between Bijapur **PERSON**, Ahmednagar **ORG**, and Golconda **ORG**, as a general. He also owned a Jaigirdari **GPE** near Pune **GPE**. Shivaji's mother Jijabai **GPE** was the daughter of Sindkhed **ORG** leader Lakhujirao Jadhav **PERSON** and a deeply religious woman. Shivaji **PERSON** was especially close to his mother who instilled in him a strict sense of right and wrong. Since Shahaji **NORP** spent most of his time outside of Pune **ORG**, the responsibility of overseeing Shivaji **PERSON**'s education rested on the shoulders of a small council of ministers which included a Peshwa (Shamrao Nilkanth), a Mazumdar (Balkrishna Pant **PERSON**), a Sabnis (Raghunath Ballal **PERSON**), a Dabir **PERSON** (Sonopant) and a chief teacher (Dadoji Konddeo **PERSON**). Kanhoji Jedhe **PERSON** and Baji Pasalkar **PERSON** were appointed to train Shivaji **PERSON** in military and martial arts. Shivaji **PERSON** was married to Saibai Nimbalkar **PERSON** in 1640 **DATE**.

Shivaji turned out to be a born leader from a very young age. An active outdoorsman, he explored the Sahayadri Mountains surrounding the Shivneri **NORP** forts and came to know the area like the back of his hands. By the time he was 15 **DATE**, he had accumulated a band of faithful soldiers from the Maval **GPE** region who later aided in his early conquests.

```
In [16]: entity_types = ["PERSON", "ORG", "GPE", "PRODUCT", "DATE"]

entities = [(ent.text, ent.label_) for ent in text_nlp.ents if ent.label_ in entity_ty

# Extract the relationships between the entities using dependency parsing and POS tags
relations = []
for token in text_nlp:
    if token.dep_ == 'ROOT' and token.pos_ == 'VERB':
        for child in token.children:
            if child.ent_type_ in entity_types:
                relations.append((child.text, token.text, child.ent_type_))

# Create tuples for information extraction
tuples = []
```

```
for relation in relations:
    for entity in entities:
        if entity[0] == relation[0]:
            tuples.append((entity[0], entity[1], relation[2]))

# Display the extracted tuples
print("Extracted Tuples:")
for t in tuples:
    print("T1({}, {}, {})".format(t[0], t[1], t[2]))
```

Extracted Tuples:
T1(Jaigirdari, GPE, GPE)

In []: