09-05-2024 NLP

Introduction to the spaCy library.

Why Spacy?

NLTK is a procedure based library which causes programs to be lengthy while Spacy is an Object Oriented Library

Installation of Spacy

For Anaconda

```
conda install -c conda-forge Spacy
python -m spacy download en_core_web_sm
```

For PIP

```
pip install -U setuptools wheel
pip install -U spacy
python -m spacy download en_core_web_sm
```

Example of Spacy Library

```
import spacy
nlp = spacy.load("en_core_web_sm")
doc = nlp("Apple is looking at buying U.K startup for $1 billion")

for token in doc:
    print(token.text, token.lemma_, token.pos_)
```

Output:

> *Apple Apple PROP*
> *is be AUX*
> *looking look VERB*
> *at at ADP*
> *$ $ SYM*
> *1 1 NUM*

Example of Fetching emails from text using python

```
import spacy
nlp = spacy.load("en_core_web_sm")
doc = nlp("Hi, my personal email address is hrushkoli1@gmail.com, and my university email
address is hrishikesh.koli.mscai2024@aurouniveristy.edu.in")
```

```
emails = []
for tokens in doc:
    if tokens.like_email:
        emails.append(tokens.text)

print(emails)
```

Output:

> *['hrushkoli1@gmail.com', 'hrishikesh.koli.mscai2024@aurouniveristy.edu.in']*

Example of Named Entity Recognition (NER)

```
import spacy
nlp = spacy.load("en_core_web_sm")
doc = nlp('Microsoft is generating $198.3 billion in USA')

for entity in doc.ents:
    print(entity.text,'|',entity.label_,'|',spacy.explain(entity.label_))
```

Output:

> *Microsoft | ORG | Companies, agencies, institutions, etc.*
> *$198.3 billion | MONEY | Monetary values, including unit*
> *USA | GPE | Countries, cities, states*

Example of Highligting NER in text using spaCy.displacy

```
import spacy
from spacy import displacy

nlp = spacy.load("en_core_web_sm")
doc = nlp('Microsoft is generating $198.3 billion in USA')

displacy.render(doc,style='ent') # It renders a HTML object highligting entities (ent) in
the doc
```

Output:

Microsoft **ORG** is generating $198.3 billion **MONEY** in USA **GPE**