Natural Language Processing ¶

Natural language processing is the science of understanding text.

We will use spacy which is a library for NLP and the en_core_web_sm which is a model trained on English corpus.

In [1]:

```
# install spacy.
#%conda install -c conda-forge spacy
```

In [2]:

```
# Contains English tokenizer, tagger, parser, NER and word vectors.
#%conda install -c conda-forge spacy-model-en_core_web_sm
```

In [3]:

```
import spacy
import en_core_web_sm

English = en_core_web_sm.load()
```

In [4]:

```
doc = English('European regulators have fined Microsoft about $730 mi
llion '
              'for failing to honor an agreement to give users a choi
ce of Internet browser.')
[(ent.text, ent.label_) for ent in doc.ents]
```

Out[4]:

```
[('European', 'NORP'), ('Microsoft', 'ORG'), ('about $730
million', 'MONEY')]
```

```
TYPE          DESCRIPTION
PERSON        People, including fictional.
NORP          Nationalities or religious or political groups.
FAC           Buildings, airports, highways, bridges, etc.
ORG           Companies, agencies, institutions, etc.
GPE           Countries, cities, states.
LOC           Non-GPE locations, mountain ranges, bodies of wat
er.
PRODUCT       Objects, vehicles, foods, etc. (Not services.)
EVENT         Named hurricanes, battles, wars, sports events, e
tc.
WORK_OF_ART   Titles of books, songs, etc.
LAW           Named documents made into laws.
LANGUAGE      Any named language.
DATE          Absolute or relative dates or periods.
TIME          Times smaller than a day.
PERCENT       Percentage, including ”%“.
MONEY         Monetary values, including unit.
QUANTITY      Measurements, as of weight or distance.
ORDINAL       “first”, “second”, etc.
CARDINAL      Numerals that do not fall under another type.
```

In [5]:

```
spacy.displacy.render(doc, jupyter=True, style='ent')
```

European **NORP** regulators have fined Microsoft **ORG** about $730 million **MONEY** for failing to honor an agreement to give users a choice of Internet browser.

In [6]:

```
#%pip install newspaper3k
```

In [7]:

```
url = 'https://thenextweb.com/security/2019/09/10/us-court-says-scrap
ing-a-site-without-permission-isnt-illegal/'
```

In [8]:

```
from newspaper import Article
```

In [9]:

```
article = Article(url)
```

In [10]:

```
article.download()
```

In [11]:

```
article.parse()
```

In [12]:

```
article.title
```

Out[12]:

```
'US court says scraping a site without permission isn’t i
llegal'
```

In [13]:

```
article.authors
```

Out[13]:

```
['Ivan Mehta']
```

In [14]:

```
article.publish_date
```

Out[14]:

```
datetime.datetime(2019, 9, 10, 0, 0)
```

In [15]:

```
article.text
```

Out[15]:

```
"An appeals court situated in California, US, today said
it’s not illegal to scrape data from public websites with
out any prior approval. Web scraping refers to the proces
s of collecting large troves of data with the use of web
crawlers — scripts designed to lift information from web
pages.\n\nThe ruling comes after a legal dispute between
LinkedIn and data analytics firm HiQ. LinkedIn sent a cea
se-and-desist letter to HiQ, demanding it to stop scrapin
g the site. In response, the data analytics company count
er-sued in hopes of blocking LinkedIn from interfering.\n
\nThe company argued that it blocked HiQ from scraping th
e data to protect its users’ privacy. On the flip side, t
he data analytics company said LinkedIn started blocking
its scraping requests only after it launched its own anal
ytics tool.\n\nThe court banned the Microsoft-owned compa
ny from blocking HiQ’s attempts to scrape data from publi
cly available profiles on the platform.\n\nBIG NEWS: 9th
Circuit holds that scraping a public website likely does
not violate the CFAA, even after website owner prohibits
with a cease-and-desist letter; language strongly suggest
s CFAA only applies to bypassing authentication. Blog pos
t up soon. https://t.co/OiWWDSFsFA #N pic.twitter.com/A7h
jg0iife — Orin Kerr (@OrinKerr) September 9, 2019\n\nA Li
nkedIn spokesperson told the Register the company will co
ntinue to fight the case:\n\nWe’re disappointed in the co
urt’s decision, and we are evaluating our options followi
ng this appeal. LinkedIn will continue to fight to protec
t our members and the information they entrust to LinkedI
n.\n\nIn earlier cases, such as Facebook v Power.com and
Craiglist v 3Taps, courts have sided with companies whose
data was being scrapped. However, this case might set a n
ew precedent if the appeals court’s decision stands. Howe
ver, it might jeopardize the privacy and data of users wh
o has a public profile.\n\nRead next: Moog's retro Matria
rch analog synth is now shipping"
```

In [16]:

```
article.top_img
```

Out[16]:

```
'https://img-cdn.tnwcdn.com/image/tnw?filter_last=1&fit=1
280%2C640&url=https%3A%2F%2Fcdn0.tnwcdn.com%2Fwp-content%
2Fblogs.dir%2F1%2Ffiles%2F2019%2F09%2Fsocial-media-143298
5_1920-1.jpg&signature=18274de15a38e676eafd47aa6eb38a3a'
```

In [18]:

```
text = article.text.replace('\n', '')
doc = English(text)
```

In [20]:

```
spacy.displacy.render(doc, jupyter=True, style='ent')
```

An appeals court situated in California GPE , US GPE , today DATE said it’s not illegal to scrape data from public websites without any prior approval. Web scraping refers to the process of collecting large troves of data with the use of web crawlers – scripts designed to lift information from web pages.The ruling comes after a legal dispute between LinkedIn ORG and data analytics firm HiQ. LinkedIn ORG sent a cease-and-desist letter to HiQ ORG , demanding it to stop scraping the site. In response, the data analytics company counter-sued in hopes of blocking LinkedIn ORG from interfering.The company argued that it blocked HiQ from scraping the data to protect its users’ privacy. On the flip side, the data analytics company said LinkedIn ORG started blocking its scraping requests only after it launched its own analytics tool.The court banned the Microsoft ORG -owned company from blocking HiQ ORG ’s attempts to scrape data from publicly available profiles on the platform.BIG NEWS: 9th Circuit ORG holds that scraping a public website likely does not violate the CFAA ORG , even after website owner prohibits with a cease-and-desist letter; language strongly suggests CFAA ORG only applies to bypassing authentication. Blog post up soon. https://t.co/OiWWDSFsFA #N pic.twitter.com/A7hjg0iife — Orin Kerr PERSON (@OrinKerr) September 9 DATE , 2019A CARDINAL LinkedIn ORG spokesperson told the Register ORG the company will continue to fight the case:We’re disappointed in the court’s decision, and we are evaluating our options following this appeal. LinkedIn ORG will continue to fight to protect our members and the information they entrust to LinkedIn ORG .In earlier cases, such as Facebook v Power.com ORG and Craiglist ORG v 3Taps CARDINAL , courts have

sided with companies whose data was being scrapped. However, this case might set a new precedent if the appeals court’s decision stands. However, it might jeopardize the privacy and data of users who has a public profile.Read next: Moog **PERSON** 's retro Matriarch **FAC** analog synth **ORDINAL** is now shipping