# Spacy Tutorial

In [2]:

```python
import spacy
```

In [3]:

```python
! pipenv install spacy
```

```
Installing spacy...
Adding spacy to Pipfile's [packages]...
✔ Installation Succeeded
Installing dependencies from Pipfile.lock (110078)...
/home/christangrant/.local/lib/python3.9/site-packages/pipenv/patched/notpip/_internal/op
erations/prepare.py:218: PipDeprecationWarning: DEPRECATION: A future pip version will ch
ange local packages to be built in-place without first copying to a temporary directory.
We recommend you use --use-feature=in-tree-build to test your packages with this new beha
vior before it becomes the default.
 pip 21.3 will remove support for this functionality. You can find discussion regarding t
his at https://github.com/pypa/pip/issues/7555.
  deprecated(
  ☐  ██████████████████████████████████████ 2/2 — 00:00:00:00:00:00:00
```

In [4]:

```python
spacy.info()
```

Out[4]:

```
{'spacy_version': '3.2.2',
 'location': '/home/christangrant/.local/share/virtualenvs/cs5293sp22-WqRDOD2N/lib/python
3.10/site-packages/spacy',
 'platform': 'Linux-5.13.0-1015-gcp-x86_64-with-glibc2.34',
 'python_version': '3.10.2',
 'pipelines': {'en_core_web_sm': '3.2.0', 'en_core_web_lg': '3.2.0'}}
```

In [5]:

```python
nlp = spacy.blank("en")
```

In [6]:

```python
nlp?
```

```
Signature:
nlp(
    text: Union[str, spacy.tokens.doc.Doc],
    *,
    disable: Iterable[str] = [],
    component_cfg: Optional[Dict[str, Dict[str, Any]]] = None,
) -> spacy.tokens.doc.Doc
Type:           English
String form:    <spacy.lang.en.English object at 0x7f3ab39df8b0>
File:           ~/.local/share/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packa
ges/spacy/lang/en/__init__.py
Docstring:      <no docstring>
Class docstring:
A text-processing pipeline. Usually you'll load this once per process,
and pass the instance around your application.

Defaults (class): Settings, data and factory methods for creating the `nlp`
    object and processing pipeline.
lang (str): IETF language code, such as 'en'.

DOCS: https://spacy.io/api/language
Init docstring:
```

Initialise a Language object.

vocab (Vocab): A `Vocab` object. If `True`, a vocab is created.
meta (dict): Custom meta data for the Language class. Is written to by
    models to add model meta data.
max_length (int): Maximum number of characters in a single text. The
    current models may run out memory on extremely long texts, due to
    large internal allocations. You should segment these texts into
    meaningful units, e.g. paragraphs, subsections etc, before passing
    them to spaCy. Default maximum length is 1,000,000 charas (1mb). As
    a rule of thumb, if all pipeline components are enabled, spaCy's
    default models currently requires roughly 1GB of temporary memory per
    100,000 characters in one text.
create_tokenizer (Callable): Function that takes the nlp object and
    returns a tokenizer.
batch_size (int): Default batch size for pipe and evaluate.

DOCS: https://spacy.io/api/language#init
Call docstring:
Apply the pipeline to some text. The text can span multiple sentences,
and can contain arbitrary whitespace. Alignment into the original string
is preserved.

text (Union[str, Doc]): If `str`, the text to be processed. If `Doc`,
    the doc will be passed directly to the pipeline, skipping
    `Language.make_doc`.
disable (List[str]): Names of the pipeline components to disable.
component_cfg (Dict[str, dict]): An optional dictionary with extra
    keyword arguments for specific components.
RETURNS (Doc): A container for accessing the annotations.

DOCS: https://spacy.io/api/language#call

In [7]:

```
nlp_es = spacy.blank("es")
nlp_de = spacy.blank("de") # xx
```

In [8]:

```
doc = nlp("Brandon's the quick brown fox jumped over the lazy dog! Thirty")
```

In [9]:

```
for token in doc:
    print(f"{token.text}")
```

Brandon
's
the
quick
brown
fox
jumped
over
the
lazy
dog
!
Thirty

In [10]:

```
tok = doc[5]
type(tok)
```

Out[10]:

```
spacy.tokens.token.Token
```

In [11]:

```
span1 = doc[3:6]
type(span1)
```

Out[11]:

```
spacy.tokens.span.Span
```

In [12]:

```
tok.i
```

Out[12]:

```
5
```

In [13]:

```
tok.is_alpha
```

Out[13]:

```
True
```

In [14]:

```
tok.is_punct
```

Out[14]:

```
False
```

In [15]:

```
tok.like_num
```

Out[15]:

```
False
```

In [16]:

```
doc[-1].like_num # Works for numbers and textul representations of numbers
```

Out[16]:

```
True
```

In [17]:

```
! pipenv run python -m spacy download en_core_web_sm
```

```
Collecting en-core-web-sm==3.2.0
  Downloading https://github.com/explosion/spacy-models/releases/download/en_core_web_sm-
3.2.0/en_core_web_sm-3.2.0-py3-none-any.whl (13.9 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 13.9/13.9 MB 31.5 MB/s eta 0:00:0000:010:01
Requirement already satisfied: spacy<3.3.0,>=3.2.0 in /home/christangrant/.local/share/vi
rtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from en-core-web-sm==3.2.0) (
3.2.2)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.8 in /home/christangrant/.local/s
hare/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2
.0->en-core-web-sm==3.2.0) (3.0.9)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /home/christangrant/.local/
share/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.
2.0->en-core-web-sm==3.2.0) (1.0.1)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /home/christangrant/.local/share/
virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0->e
n-core-web-sm==3.2.0) (3.0.6)
Requirement already satisfied: packaging>=20.0 in /home/christangrant/.local/share/virtua
lenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0->en-core
-web-sm==3.2.0) (21.3)
Requirement already satisfied: srsly<3.0.0,>=2.4.1 in /home/christangrant/.local/share/vi
rtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0->en-
core-web-sm==3.2.0) (2.4.2)
Requirement already satisfied: thinc<8.1.0,>=8.0.12 in /home/christangrant/.local/share/v
```

rtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0->en
-core-web-sm==3.2.0) (8.0.13)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /home/christangrant/.local/share/vi
rtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0->en-
core-web-sm==3.2.0) (4.62.3)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /home/christangrant/.local/shar
e/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0-
>en-core-web-sm==3.2.0) (2.27.1)
Requirement already satisfied: pathy>=0.3.5 in /home/christangrant/.local/share/virtualen
vs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0->en-core-we
b-sm==3.2.0) (0.6.1)
Requirement already satisfied: numpy>=1.15.0 in /home/christangrant/.local/share/virtuale
nvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0->en-core-w
eb-sm==3.2.0) (1.22.2)
Requirement already satisfied: setuptools in /home/christangrant/.local/share/virtualenvs
/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0->en-core-web-
sm==3.2.0) (60.9.3)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /home/christangrant/.local/share/vi
rtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0->en-
core-web-sm==3.2.0) (2.0.6)
Requirement already satisfied: wasabi<1.1.0,>=0.8.1 in /home/christangrant/.local/share/v
irtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0->en
-core-web-sm==3.2.0) (0.9.0)
Requirement already satisfied: typer<0.5.0,>=0.3.0 in /home/christangrant/.local/share/vi
rtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0->en-
core-web-sm==3.2.0) (0.4.0)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in /home/christangrant/.local/shar
e/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0-
>en-core-web-sm==3.2.0) (3.3.0)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<1.9.0,>=1.7.4 in /home/christangran
t/.local/share/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3
.3.0,>=3.2.0->en-core-web-sm==3.2.0) (1.8.2)
Requirement already satisfied: jinja2 in /home/christangrant/.local/share/virtualenvs/cs5
293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0->en-core-web-sm==
3.2.0) (3.0.3)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /home/christangrant/.local/sh
are/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.
0->en-core-web-sm==3.2.0) (1.0.6)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /home/christangrant/.local/shar
e/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0-
>en-core-web-sm==3.2.0) (2.0.6)
Requirement already satisfied: blis<0.8.0,>=0.4.0 in /home/christangrant/.local/share/vir
tualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0->en-c
ore-web-sm==3.2.0) (0.7.5)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /home/christangrant/.local/sha
re/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from packaging>=20.0->sp
acy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (3.0.7)
Requirement already satisfied: smart-open<6.0.0,>=5.0.0 in /home/christangrant/.local/sha
re/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from pathy>=0.3.5->spacy
<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (5.2.1)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /home/christangrant/.local/s
hare/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from pydantic!=1.8,!=1
.8.1,<1.9.0,>=1.7.4->spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (4.1.1)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /home/christangrant/.local/share/
virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from requests<3.0.0,>=2.13.
0->spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (1.26.8)
Requirement already satisfied: charset-normalizer~=2.0.0 in /home/christangrant/.local/sh
are/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from requests<3.0.0,>=2
.13.0->spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (2.0.12)
Requirement already satisfied: idna<4,>=2.5 in /home/christangrant/.local/share/virtualen
vs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from requests<3.0.0,>=2.13.0->spacy<
3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (3.3)
Requirement already satisfied: certifi>=2017.4.17 in /home/christangrant/.local/share/vir
tualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from requests<3.0.0,>=2.13.0->
spacy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (2021.10.8)
Requirement already satisfied: click<9.0.0,>=7.1.1 in /home/christangrant/.local/share/vi
rtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from typer<0.5.0,>=0.3.0->spa
cy<3.3.0,>=3.2.0->en-core-web-sm==3.2.0) (8.0.4)
Requirement already satisfied: MarkupSafe>=2.0 in /home/christangrant/.local/share/virtua
lenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from jinja2->spacy<3.3.0,>=3.2.0-
>en-core-web-sm==3.2.0) (2.1.0)
✔ Download and installation successful

You can now load the package via spacy.load('en_core_web_sm')

In [18]:

```python
nlp = spacy.load('en_core_web_sm')
```

In [19]:

```python
doc = nlp("How much wood could a Woodchuck chuck if a wood chuck could chuck wood.")
```

In [20]:

```python
for token in doc:
    print(f"{token.text}--{token.pos_}")
```

```
How--SCONJ
much--ADJ
wood--NOUN
could--AUX
a--DET
Woodchuck--PROPN
chuck--NOUN
if--SCONJ
a--DET
wood--NOUN
chuck--NOUN
could--AUX
chuck--VERB
wood--NOUN
.--PUNCT
```
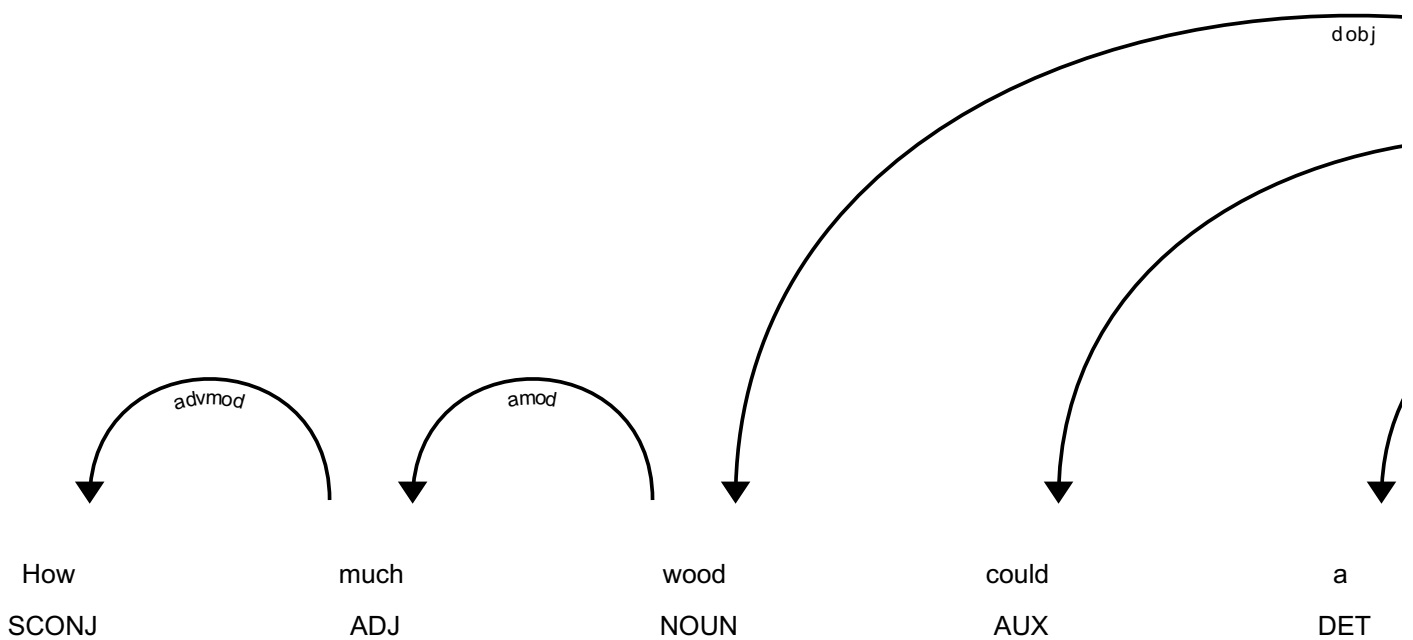
In [21]:

```python
for token in doc:
    print(f"{token.text}--{token.pos_}--{token.dep_}--{token.head.text}")
```

```
How--SCONJ--advmod--much
much--ADJ--amod--wood
wood--NOUN--dobj--chuck
could--AUX--aux--chuck
a--DET--det--chuck
Woodchuck--PROPN--compound--chuck
chuck--NOUN--ROOT--chuck
if--SCONJ--mark--chuck
a--DET--det--chuck
wood--NOUN--compound--chuck
chuck--NOUN--nsubj--chuck
could--AUX--aux--chuck
chuck--VERB--advcl--chuck
wood--NOUN--dobj--chuck
.--PUNCT--punct--chuck
```

In [22]:

```python
from spacy import displacy

displacy.render(doc, style="dep")
```

|  | How | much | wood | could | a |
|---|---|---|---|---|---|
|  | SCONJ | ADJ | NOUN | AUX | DET |

In [23]:

```
doc
```

Out[23]:

How much wood could a Woodchuck chuck if a wood chuck could chuck wood.

In [24]:

```
for ent in doc.ents:
    print(f"{ent.text}->{ent.label_}")
```

```
Woodchuck->PERSON
chuck wood->PERSON
```

In [25]:

```
doc = nlp("Apple shared tanked with the release of the buggy software written by Greg.")
```

In [26]:

```
for ent in doc.ents:
    print(f"{ent.text}->{ent.label_}")
```

```
Apple->ORG
Greg->PERSON
```

In [27]:

```
spacy.explain("ORG")
```

Out[27]:

'Companies, agencies, institutions, etc.'

In [28]:

```
spacy.explain("mark")
```

Out[28]:

'marker'

```python
doc = nlp("Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo")

for ent in doc.ents:
    print(f"{ent.text}->{ent.label_}")

for token in doc:
    print(f"{token.text}--{token.pos_}--{token.dep_}--{token.head.text}")
```

```
Buffalo--PROPN--compound--buffalo
buffalo--PROPN--compound--buffalo
Buffalo--PROPN--compound--buffalo
buffalo--PROPN--compound--buffalo
buffalo--PROPN--compound--buffalo
buffalo--PROPN--compound--buffalo
Buffalo--PROPN--compound--buffalo
buffalo--PROPN--ROOT--buffalo
```

In [30]:

```
! pipenv run python -m spacy download en_core_web_lg
```

```
Collecting en-core-web-lg==3.2.0
  Downloading https://github.com/explosion/spacy-models/releases/download/en_core_web_lg-
3.2.0/en_core_web_lg-3.2.0-py3-none-any.whl (777.4 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 777.4/777.4 MB 1.0 MB/s eta 0:00:0000:0100:0
1
Requirement already satisfied: spacy<3.3.0,>=3.2.0 in /home/christangrant/.local/share/vi
rtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from en-core-web-lg==3.2.0) (
3.2.2)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /home/christangrant/.local/
share/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.
2.0->en-core-web-lg==3.2.0) (1.0.1)
Requirement already satisfied: srsly<3.0.0,>=2.4.1 in /home/christangrant/.local/share/vi
rtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0->en-
core-web-lg==3.2.0) (2.4.2)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<1.9.0,>=1.7.4 in /home/christangran
t/.local/share/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3
.3.0,>=3.2.0->en-core-web-lg==3.2.0) (1.8.2)
Requirement already satisfied: packaging>=20.0 in /home/christangrant/.local/share/virtua
lenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0->en-core
-web-lg==3.2.0) (21.3)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /home/christangrant/.local/shar
e/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0-
>en-core-web-lg==3.2.0) (2.0.6)
Requirement already satisfied: numpy>=1.15.0 in /home/christangrant/.local/share/virtuale
nvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0->en-core-w
eb-lg==3.2.0) (1.22.2)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /home/christangrant/.local/share/vi
rtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0->en-
core-web-lg==3.2.0) (2.0.6)
Requirement already satisfied: wasabi<1.1.0,>=0.8.1 in /home/christangrant/.local/share/v
irtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0->en
-core-web-lg==3.2.0) (0.9.0)
Requirement already satisfied: thinc<8.1.0,>=8.0.12 in /home/christangrant/.local/share/v
irtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0->en
-core-web-lg==3.2.0) (8.0.13)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /home/christangrant/.local/shar
e/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0-
>en-core-web-lg==3.2.0) (2.27.1)
Requirement already satisfied: pathy>=0.3.5 in /home/christangrant/.local/share/virtualen
vs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0->en-core-we
b-lg==3.2.0) (0.6.1)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in /home/christangrant/.local/shar
e/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0-
>en-core-web-lg==3.2.0) (3.3.0)
Requirement already satisfied: setuptools in /home/christangrant/.local/share/virtualenvs
/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2.0->en-core-web-
lg==3.2.0) (60.9.3)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.8 in /home/christangrant/.local/s
hare/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages (from spacy<3.3.0,>=3.2
.0->en-core-web-lg==3.2.0) (3.0.9)
```

In [31]:

```python
nlp = spacy.load('en_core_web_lg')
```

In [32]:

```python
doc = nlp("Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo")
```

In [33]:

```python
for ent in doc.ents:
    print(f"{ent.text}->{ent.label_}")

for token in doc:
    print(f"{token.text}--{token.pos_}--{token.dep_}--{token.head.text}")
```

```
Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo->ORG
Buffalo--PROPN--compound--buffalo
buffalo--PROPN--compound--buffalo
Buffalo--PROPN--compound--buffalo
buffalo--PROPN--compound--buffalo
buffalo--PROPN--compound--buffalo
buffalo--PROPN--compound--buffalo
Buffalo--PROPN--compound--buffalo
```

buffalo--PROPN--ROOT--buffalo

# Matcher

In [34]:

```python
from spacy.matcher import Matcher
```

In [35]:

```python
matcher = Matcher(nlp.vocab)
```

In [36]:

```python
pattern = [
    {"TEXT": "Android"},
    {"TEXT": "X"}
]
matcher.add("ANDROID_PATTERN", [pattern])
```

In [37]:

```python
doc = nlp("The class had a farm with pigs, cows, and bombs and they managed it on their Android X phone.")
```

In [38]:

```python
for ent in doc.ents:
    print(f"{ent.text}->{ent.label_}")

for token in doc:
    print(f"{token.text}--{token.pos_}--{token.dep_}--{token.head.text}")
```

```
The--DET--det--class
class--NOUN--nsubj--had
had--VERB--ROOT--had
a--DET--det--farm
farm--NOUN--dobj--had
with--ADP--prep--farm
pigs--NOUN--pobj--with
,--PUNCT--punct--pigs
cows--NOUN--conj--pigs
,--PUNCT--punct--cows
and--CCONJ--cc--cows
bombs--NOUN--conj--cows
and--CCONJ--cc--had
they--PRON--nsubj--managed
managed--VERB--conj--had
it--PRON--dobj--managed
on--ADP--prep--managed
their--PRON--poss--phone
Android--PROPN--compound--X
X--PROPN--compound--phone
phone--NOUN--pobj--on
.--PUNCT--punct--managed
```

In [39]:

```python
matches = matcher(doc)
matches
```

Out[39]:

```
[(10465695285070383609, 18, 20)]
```

In [40]:

```python
for match_id, start, end in matches:
```

```
    matched_span = doc[start:end]
    print(f"{matched_span.text}")
```

Android X

In [41]:

```
pattern1 = [
    {"TEXT": "Android"},
    {"TEXT": "X"},
]
pattern2 = [
    {"IS_DIGIT", True},
    {"LOWER", "android"},
    {"IS_PUNCT": True, "OP": "?"},
    {"POS": "NOUN"}
]
```

In [42]:

```
doc[1].orth
```

Out[42]:

12859657019542931961

In [44]:

```
nlp.vocab?
```

```
Type:           Vocab
String form:    <spacy.vocab.Vocab object at 0x7f3ae6d35f30>
Length:         787
File:           ~/.local/share/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packag
es/spacy/vocab.cpython-310-x86_64-linux-gnu.so
Docstring:
Vocab(lex_attr_getters=None, strings=tuple(), lookups=None, oov_prob=-20., vectors_name=N
one, writing_system={}, get_noun_chunks=None, **deprecated_kwargs)
A look-up table that allows you to access `Lexeme` objects. The `Vocab`
    instance also provides access to the `StringStore`, and owns underlying
    C-data that is shared between `Doc` objects.

    DOCS: https://spacy.io/api/vocab

Init docstring:
Create the vocabulary.

lex_attr_getters (dict): A dictionary mapping attribute IDs to
    functions to compute them. Defaults to `None`.
strings (StringStore): StringStore that maps strings to integers, and
    vice versa.
lookups (Lookups): Container for large lookup tables and dictionaries.
oov_prob (float): Default OOV probability.
vectors_name (str): Optional name to identify the vectors table.
get_noun_chunks (Optional[Callable[[Union[Doc, Span], Iterator[Tuple[int, int, int]]]]]):
    A function that yields base noun phrases used for Doc.noun_chunks.
```

In [46]:

```
nlp.vocab.strings["coffee"]
```

Out[46]:

3197928453018144401

In [50]:

```
nlp.vocab.strings["coffeeew"]
```

Out[50]:

1615321019953645392
```

```
In [48]:
```

```
nlp.vocab.strings[3197928453018144401]
```

```
Out[48]:
```

```
'coffee'
```

```
In [49]:
```

```
nlp.vocab.strings[12696802605381437503]
```

```
Out[49]:
```

```
'coffeee'
```

```
In [64]:
```

```
doc = nlp("I love coffee .com")
```

```
In [65]:
```

```
t = doc[2]
print(t)
```

```
coffee
```

```
In [66]:
```

```
t?
```

```
Type:        Token
String form: coffee
Length:      6
File:        ~/.local/share/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packages/
spacy/tokens/token.cpython-310-x86_64-linux-gnu.so
Docstring:
An individual token – i.e. a word, punctuation symbol, whitespace,
etc.

DOCS: https://spacy.io/api/token
```

```
In [67]:
```

```
t.orth
```

```
Out[67]:
```

```
3197928453018144401
```

```
In [68]:
```

```
lexeme = nlp.vocab["coffee"]
```

```
In [69]:
```

```
lexeme?
```

```
Type:           Lexeme
String form:    <spacy.lexeme.Lexeme object at 0x7f3ab1fb5a00>
File:           ~/.local/share/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packag
es/spacy/lexeme.cpython-310-x86_64-linux-gnu.so
Docstring:
Lexeme(Vocab vocab, attr_t orth)
An entry in the vocabulary. A `Lexeme` has no string context – it's a
    word-type, as opposed to a word token.  It therefore has no part-of-speech
    tag, dependency parse, or lemma (lemmatization depends on the
    part-of-speech tag).

    DOCS: https://spacy.io/api/lexeme

Init docstring:
Create a Lexeme object
```

Create a Lexeme object.

    vocab (Vocab): The parent vocabulary
    orth (uint64): The orth id of the lexeme.
    Returns (Lexeme): The newly constructd object.

```python
nlp = spacy.blank("en")

from spacy.tokens import Doc
```

```python
Doc?
```

Docstring:
Doc(Vocab vocab, words=None, spaces=None, user_data=None, *, tags=None, pos=None, morphs=
None, lemmas=None, heads=None, deps=None, sent_starts=None, ents=None)
A sequence of Token objects. Access sentences and named entities, export
    annotations to numpy arrays, losslessly serialize to compressed binary
    strings. The `Doc` object holds an array of `TokenC` structs. The
    Python-level `Token` and `Span` objects are views of this array, i.e.
    they don't own the data themselves.

    EXAMPLE:
        Construction 1
        >>> doc = nlp(u'Some text')

        Construction 2
        >>> from spacy.tokens import Doc
        >>> doc = Doc(nlp.vocab, words=["hello", "world", "!"], spaces=[True, False, Fals
e])

    DOCS: https://spacy.io/api/doc

Init docstring:
Create a Doc object.

vocab (Vocab): A vocabulary object, which must match any models you
    want to use (e.g. tokenizer, parser, entity recognizer).
words (Optional[List[Union[str, int]]]): A list of unicode strings or
    hash values to add to the document as words. If `None`, defaults to
    empty list.
spaces (Optional[List[bool]]): A list of boolean values, of the same
    length as `words`. `True` means that the word is followed by a space,
    `False` means it is not. If `None`, defaults to `[True]*len(words)`
user_data (dict or None): Optional extra data to attach to the Doc.
tags (Optional[List[str]]): A list of unicode strings, of the same
    length as words, to assign as token.tag. Defaults to None.
pos (Optional[List[str]]): A list of unicode strings, of the same
    length as words, to assign as token.pos. Defaults to None.
morphs (Optional[List[str]]): A list of unicode strings, of the same
    length as words, to assign as token.morph. Defaults to None.
lemmas (Optional[List[str]]): A list of unicode strings, of the same
    length as words, to assign as token.lemma. Defaults to None.
heads (Optional[List[int]]): A list of values, of the same length as
    words, to assign as heads. Head indices are the position of the
    head in the doc. Defaults to None.
deps (Optional[List[str]]): A list of unicode strings, of the same
    length as words, to assign as token.dep. Defaults to None.
sent_starts (Optional[List[Union[bool, None]]]): A list of values, of
    the same length as words, to assign as token.is_sent_start. Will be
    overridden by heads if heads is provided. Defaults to None.
ents (Optional[List[str]]): A list of unicode strings, of the same
    length as words, as IOB tags to assign as token.ent_iob and
    token.ent_type. Defaults to None.

DOCS: https://spacy.io/api/doc#init
File:           ~/.local/share/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packag
es/spacy/tokens/doc.cpython-310-x86_64-linux-gnu.so
Type:           type

In [73]:

```
words = ["Go", "Oklahoma", "Sooners", "!"]
spaces = [True, True, True, False]
```

In [74]:

```
doc = Doc(nlp.vocab, words=words, spaces=spaces)
```

In [75]:

```
for token in doc:
    print(token)
```

```
Go
Oklahoma
Sooners
!
```

In [76]:

```
doc[1:3]
```

Out[76]:

```
Oklahoma Sooners
```

In [77]:

```
from spacy.tokens import Span
```

In [78]:

```
Span?
```

```
Init signature: Span(self, /, *args, **kwargs)
Docstring:
A slice from a Doc object.

DOCS: https://spacy.io/api/span
File:           ~/.local/share/virtualenvs/cs5293sp22-WqRDOD2N/lib/python3.10/site-packag
es/spacy/tokens/span.cpython-310-x86_64-linux-gnu.so
Type:           type
Subclasses:
```

In [79]:

```
span = Span(doc, 1, 3, label="University Team")
```

In [83]:

```
doc.ents = [span]
for ent in doc.ents:
    print(ent.text, ent.label_)
```

```
Oklahoma Sooners University Team
```

In [ ]: