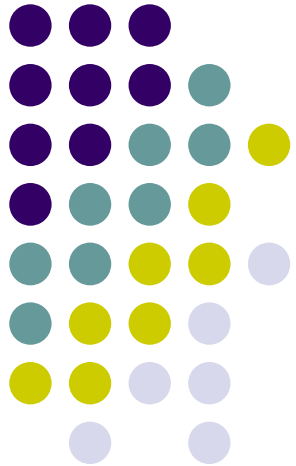
Natural Language Processing

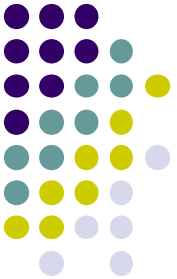


Related Efforts

- cTAKES,
- MetaMap,
- QuickUMLS

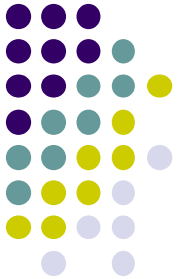
- BioBert
- ClinicalBert

1. Spacy



spaCy


Industrial-Strength
Natural Language
Processing
IN PYTHON

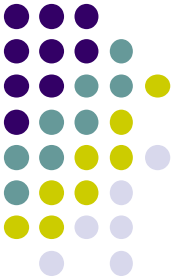


Get things done

spaCy is designed to help you do real work — to build real products, or gather real insights. The library respects your time, and tries to avoid wasting it. It's easy to install, and its API is simple and productive. We like to think of spaCy as the Ruby on Rails of Natural Language Processing.

GET STARTED

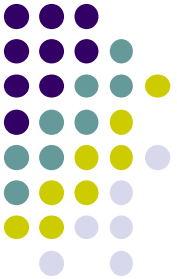
https://spacy.io/



Blazing fast

spaCy excels at large-scale information extraction tasks. It's written from the ground up in carefully memory-managed Cython. Independent research in 2015 found spaCy to be the fastest in the world. If your application needs to process entire web dumps, spaCy is the library you want to be using.

FACTS & FIGURES



Deep learning

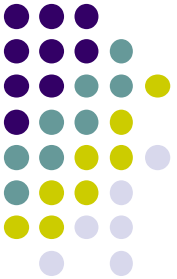
spaCy is the best way to prepare text for deep learning. It interoperates seamlessly with TensorFlow, PyTorch, scikit-learn, Gensim and the rest of Python's awesome AI ecosystem. With spaCy, you can easily construct linguistically sophisticated statistical models for a variety of NLP problems.

READ MORE

Features

- Non-destructive **tokenization**
- **Named entity** recognition
- Support for **53+ languages**
- **17 statistical models** for 11 languages
- pretrained **word vectors**
- State-of-the-art speed
- Easy **deep learning** integration
- Part-of-speech tagging
- Labelled dependency parsing
- Syntax-driven sentence segmentation
- Built in **visualizers** for syntax and NER
- Convenient string-to-hash mapping
- Export to numpy data arrays
- Efficient binary serialization
- Easy **model packaging** and deployment
- Robust, rigorously evaluated accuracy

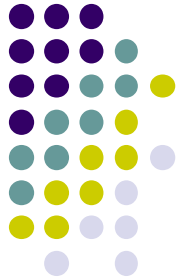
	SPACY	NLTK	CORENLP
Programming language	Python	Python	Java / Python
Neural network models	✓	✗	✓
Integrated word vectors	✓	✗	✗
Multi-language support	✓	✓	✓
Tokenization	✓	✓	✓
Part-of-speech tagging	✓	✓	✓
Sentence segmentation	✓	✓	✓
Dependency parsing	✓	✗	✓
Entity recognition	✓	✓	✓
Entity linking	✓	✗	✗
Coreference resolution	✗	✗	✓



	SPACY	NLTK	ALLEN-NLP	STANFORD-NLP	TENSOR-FLOW
I'm a beginner and just getting started with NLP.	✓	✓	✗	✓	✗
I want to build an end-to-end production application.	✓	✗	✗	✗	✓
I want to try out different neural network architectures for NLP.	✗	✗	✓	✗	✓
I want to try the latest models with state-of-the-art accuracy.	✗	✗	✓	✓	✓
I want to train models from my own data.	✓	✓	✓	✓	✓
I want my application to be efficient on CPU.	✓	✓	✗	✗	✗

NAME	DESCRIPTION
Tokenization	Segmenting text into words, punctuations marks etc.
Part-of-speech (POS) **Tagging**	Assigning word types to tokens, like verb or noun.
Dependency Parsing	Assigning syntactic dependency labels, describing the relations between individual tokens, like subject or object.
Lemmatization	Assigning the base forms of words. For example, the lemma of "was" is "be", and the lemma of "rats" is "rat".
Sentence Boundary Detection (SBD)	Finding and segmenting individual sentences.
Named Entity Recognition (NER)	Labelling named "real-world" objects, like persons, companies or locations.
Entity Linking (EL)	Disambiguating textual entities to unique identifiers in a Knowledge Base.
Similarity	Comparing words, text spans and documents and how similar they are to each other.
Text Classification	Assigning categories or labels to a whole document, or parts of a document.
Rule-based Matching	Finding sequences of tokens based on their texts and linguistic annotations, similar to regular expressions.
Training	Updating and improving a statistical model's predictions.
Serialization	Saving objects to files or byte strings.

Start with SpaCy



```
import spacy
nlp = spacy.load('en')
doc = nlp("Apple is looking at buying U.K. startup for $1 billion")
for token in doc:
  print(token.text)
```

```
Apple
is
looking
at
buying
U.K.
startup
for
$
1
billion
```

Tokenization

Tokenization

During processing, spaCy first **tokenizes** the text, i.e. segments it into words, punctuation and so on. This is done by applying rules specific to each language. For example, punctuation at the end of a sentence should be split off – whereas "U.K." should remain one token. Each `Doc` consists of individual tokens, and we can iterate over them:

```
tokens = doc.text.split()
print(tokens)
```

```
['Apple', 'is', 'looking', 'at', 'buying', 'U.K.', 'startup', 'for', '$1', 'billion']
```

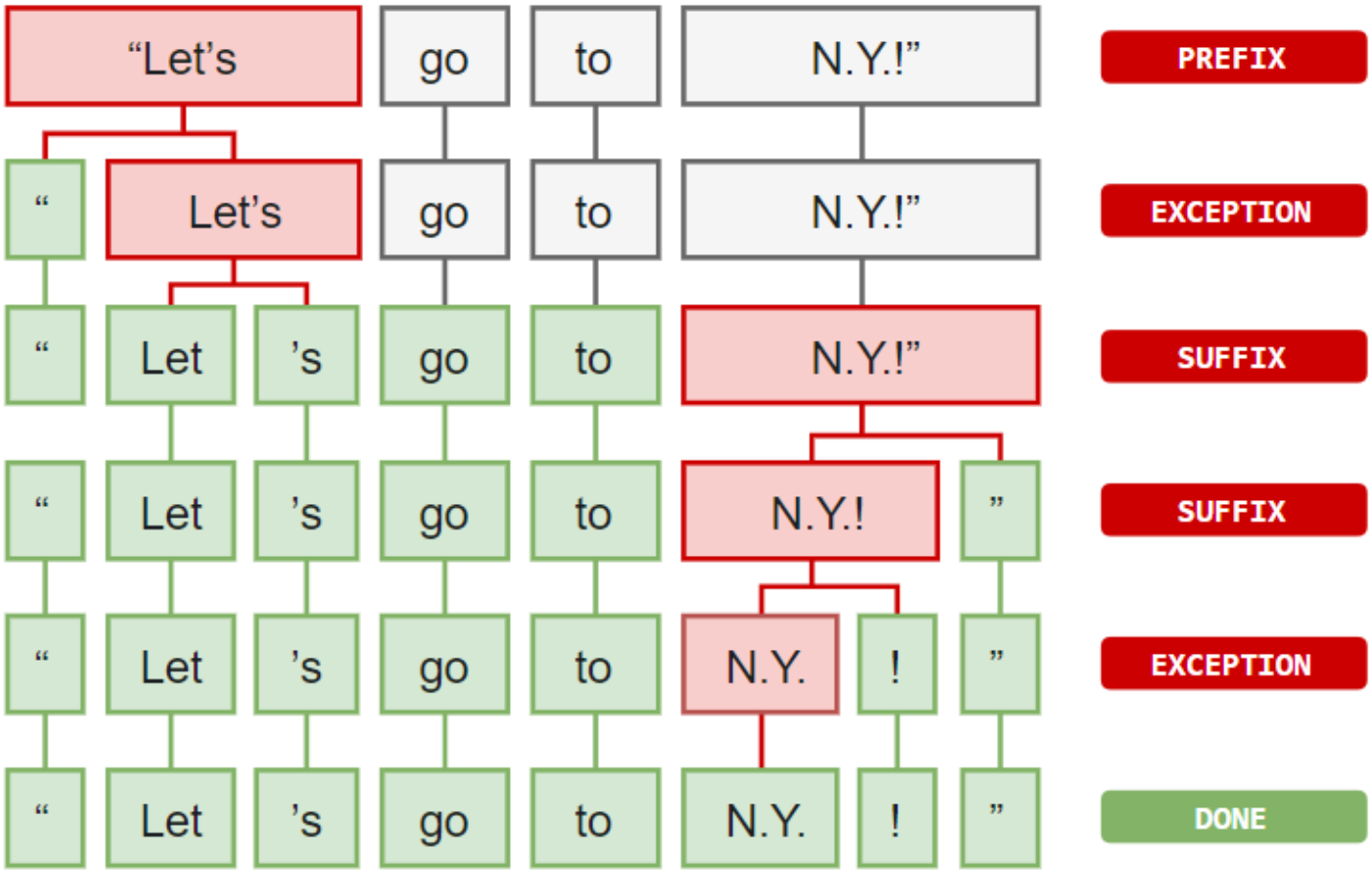
0	1	2	3	4	5	6	7	8	9	10
Apple	is	looking	at	buying	U.K.	startup	for	$	1	billion

First, the raw text is split on whitespace characters, similar to `text.split(' ')`. Then, the tokenizer processes the text from left to right. On each substring, it performs two checks:

1. **Does the substring match a tokenizer exception rule?** For example, "don't" does not contain whitespace, but should be split into two tokens, "do" and "n't", while "U.K." should always remain one token.

2. **Can a prefix, suffix or infix be split off?** For example punctuation like commas, periods, hyphens or quotes.

Tokenization



Get tokens without punctations or white space

https://spacy.io/api/

```
token_without_punct = [token.orth_ for token in doc if not token.is_punct | token.is_space]
print(token_without_punct)
```

```
['Apple', 'is', 'looking', 'at', 'buying', 'U.K.', 'startup', 'for', '$', '1', 'billion']
```

Load headache notes

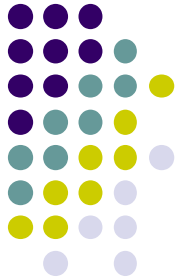
```
from google.colab import files
uploaded = files.upload()
```

Browse... notes_headache.txt

notes_headache.txt(text/plain) - 9433 bytes, last modified: n/a - 100% done

```
Saving notes_headache.txt to notes_headache.txt
```

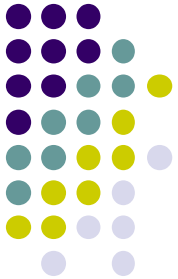
Read headache notes



```
notes = []
with open('notes_headache.txt', 'r') as fin:
  lines = fin.readlines()
  for line in lines:
    notes.append(line)
print(notes)
print(len(notes))
```

```
['50 year old female presents after having fallen
11
```

Print all the headache notes



```
for i in range(len(notes)):
  print(notes[i])
  print("**********************************************************************************")
```

```
50 year old female presents after having fallen in the bathtub 4 days ago and hitting the back of her head. Since then she has had a massive headache" which did not resolve with Tyle

**********
23 F with h/o Lupus, ESRD not on HD (planned PD), labile hypertension, RUE VTE on anticoagulation, recent facial swelling who presents with hypertensive emergency. Patient developed

**********
62yo RH M h/o HTN, Afib on coumadin, CAD s/p PTCA x 2 in [**2178**] who was in USOH today when he began to have a mild right-sided headache around noon, a/w some photophobia and naus

**********
49 year old with hx of lupus, pulmonary hypertension, RV enlargement and failure and an ASD who is being admitted for management of CHF/pulm HTN. The patient has had progressively wo

**********
83 year old female with apparently sudden-onset aphasia this morning at 7:45 PM. She apparently sustained a fall in the early part of the day on [**2171-9-25**] but has been ambulato

**********
The patient is a 72 year old woman with a history of CAD s/p CABG x4, hypertension, DM2, hyperlipidemia, and CKD stage IV who presents feeling shaky and diaphoretic at home, and in t

**********
46 yo female with no significant medical problems initially presented to PCP [**Name Initial (PRE) 151**] 2 weeks of dyspnea. She reports two weeks of incresing shortness of breath.

**********
17 year old RH boy with history of PANDAS s/p strep infection, cognitive decline over several years, junvenile rheumatoid arthritis, multiple seizures and autoimmune diseases in the

**********
82 year old female with HTN, CHF, AAA s/p repair [**43**] presenting with severe headache and substernal chest pressure. She states that headache is similar to past hypertensive head
```

Get tokens without punctations or white space – for all the notes

50 year old female presents after having fallen in the bathtub 4 days ago and hitting the back of her head. Since then she has had a massive headache" which did not resolve with Tylenol. She states that she has a high threshold for pain and did not realize how bad it was during the day while at work but then when she got home at night she noticed it. The patient noticed ""silvery specks"" in her vision and she had trouble with some simple tasks like finding the tags on the back of her clothing in the morning. She reported that she had to check several times to make sure she did not put her clothes on backwards. She has had some dizziness, but no nausea or vomiting. Her speech has not been affected.

```
# Token
token_without_punct = []
for i in range(len(doc)):
  token_without_punct.append([token.orth_ for token in doc[i] if not token.is_punct | token.is_space])
  print(token_without_punct[-1])
  print('**********************************************************************************')
```

```
['50', 'year', 'old', 'female', 'presents', 'after', 'having', 'fallen', 'in', 'the', 'bathtub', '4', 'days', 'ago', 'and', 'hitting', 'the', 'back', 'of', 'her', 'head', 'Since', 'the
**********
['23', 'F', 'with', 'h', 'o', 'Lupus', 'ESRD', 'not', 'on', 'HD', 'planned', 'PD', 'labile', 'hypertension', 'RUE', 'VTE', 'on', 'anticoagulation', 'recent', 'facial', 'swelling', 'whc
**********
['62yo', 'RH', 'M', 'h', 'o', 'HTN', 'Afib', 'on', 'coumadin', 'CAD', 's', 'p', 'PTCA', 'x', '2', 'in', '2178', 'who', 'was', 'in', 'USOH', 'today', 'when', 'he', 'began', 'to', 'have'
**********
['49', 'year', 'old', 'with', 'hx', 'of', 'lupus', 'pulmonary', 'hypertension', 'RV', 'enlargement', 'and', 'failure', 'and', 'an', 'ASD', 'who', 'is', 'being', 'admitted', 'for', 'mar
**********
['83', 'year', 'old', 'female', 'with', 'apparently', 'sudden', 'onset', 'aphasia', 'this', 'morning', 'at', '7:45', 'PM', 'She', 'apparently', 'sustained', 'a', 'fall', 'in', 'the', '
**********
['The', 'patient', 'is', 'a', '72', 'year', 'old', 'woman', 'with', 'a', 'history', 'of', 'CAD', 's', 'p', 'CABG', 'x4', 'hypertension', 'DM2', 'hyperlipidemia', 'and', 'CKD', 'stage',
**********
['46', 'yo', 'female', 'with', 'no', 'significant', 'medical', 'problems', 'initially', 'presented', 'to', 'PCP', 'Name', 'Initial', 'PRE', '151', '2', 'weeks', 'of', 'dyspnea', 'She',
**********
['17', 'year', 'old', 'RH', 'boy', 'with', 'history', 'of', 'PANDAS', 's', 'p', 'strep', 'infection', 'cognitive', 'decline', 'over', 'several', 'years', 'junvenile', 'rheumatoid', 'ar
**********
['82', 'year', 'old', 'female', 'with', 'HTN', 'CHF', 'AAA', 's', 'p', 'repair', '43', 'presenting', 'with', 'severe', 'headache', 'and', 'substernal', 'chest', 'pressure', 'She', 'sta
**********
['27', 'M', 'last', 'seen', 'normal', '3', '29', 'PM', 'was', 'found', 'by', 'mother', '1800', 'with', 'EMS', 'in', 'bed', 'per', 'Location', 'un', 'Found', 'c', '7', 'fentanyl', 'patc
**********
['This', 'is', 'a', '51', 'y.o', 'Spanish', 'speaking', 'female', 'with', 'COPD', 'OSA', 'diabetes', 'insipidus', 'Down', "'s", 'Syndrome', 'hypothyroid', 'pan', 'hypopituitary', 'who'
**********
```

Lemmatization

EXAMPLE

```
from spacy.lemmatizer import Lemmatizer
from spacy.lookups import Lookups
lookups = Lookups()
lookups.add_table("lemma_rules", {"noun": [["s", ""]]})
lemmatizer = Lemmatizer(lookups)
lemmas = lemmatizer("ducks", "NOUN")
assert lemmas == ["duck"]
```

EXAMPLE

```
lookups = Lookups()
lookups.add_table("lemma_lookup", {"going": "go"})
assert lemmatizer.lookup("going") == "go"
```

Lemmatization

```
doc = nlp("Apples and oranges are similar. Boots and hippos aren't.")
for token in doc:
  print(token, token.lemma, token.lemma_)
```

```
Apples 8566208034543834098 apple
and 2283656566040971221 and
oranges 2208928596161743350 orange
are 10382539506755952630 be
similar 18166476740537071113 similar
. 12646065887601541794 .
Boots 9918665227421442029 boot
and 2283656566040971221 and
hippos 41336932911458790 83 hippos
are 10382539506755952630 be
n't 447765159362469301 not
. 12646065887601541794 .
```

POS Tagging

Part-of-speech tags and dependencies

NEEDS MODEL

After tokenization, spaCy can **parse** and **tag** a given `Doc`. This is where the statistical model comes in, which enables spaCy to **make a prediction** of which tag or label most likely applies in this context. A model consists of binary data and is produced by showing a system enough examples for it to make predictions that generalize across the language – for example, a word following "the" in English is most likely a noun.

Linguistic annotations are available as `Token` attributes . Like many NLP libraries, spaCy **encodes all strings to hash values** to reduce memory usage and improve efficiency. So to get the readable string representation of an attribute, we need to add an underscore `_` to its name:

POS Tagging

```
doc = nlp("Apple is looking for buying U.K. startup for $1 billion.")
for token in doc:
  print(token.text, token.lemma_, token.pos_, token.tag_, token.dep_, token.shape_, token.is_alpha, token.is_stop)
```

```
Apple Apple PROPN NNP nsubj Xxxxx True False
is be AUX VBZ aux xx True True
looking look VERB VBG ROOT xxxx True False
for for ADP IN prep xxx True True
buying buy VERB VBG pcomp xxxx True False
U.K. U.K. PROPN NNP compound X.X. False False
startup startup NOUN NN dobj xxxx True False
for for ADP IN prep xxx True True
$ $ SYM $ quantmod $ False False
1 1 NUM CD compound d False False
billion billion NUM CD pobj xxxx True False
. . PUNCT . punct . False False
```

TEXT	LEMMA	POS	TAG	DEP	SHAPE	ALPHA	STOP
Apple	apple	PROPN	NNP	nsubj	Xxxxx	True	False
is	be	VERB	VBZ	aux	xx	True	True
looking	look	VERB	VBG	ROOT	xxxx	True	False
at	at	ADP	IN	prep	xx	True	True
buying	buy	VERB	VBG	pcomp	xxxx	True	False
U.K.	u.k.	PROPN	NNP	compound	X.X.	False	False
startup	startup	NOUN	NN	dobj	xxxx	True	False
for	for	ADP	IN	prep	xxx	True	True
$	$	SYM	$	quantmod	$	False	False
1	1	NUM	CD	compound	d	False	False
billion	billion	NUM	CD	pobj	xxxx	True	False

https://universaldependencies.org/u/pos/all.html

ADP: adposition

Definition

Adposition is a cover term for prepositions and postpositions.

Examples

- *in*
- *to*
- *during*

Text: The original word text.
Lemma: The base form of the word.
POS: The simple part-of-speech tag.
Tag: The detailed part-of-speech tag.
Dep: Syntactic dependency, i.e. the relation between tokens.
Shape: The word shape – capitalization, punctuation, digits.
is alpha: Is the token an alpha character?
is stop: Is the token part of a stop list, i.e. the most common words of the language?

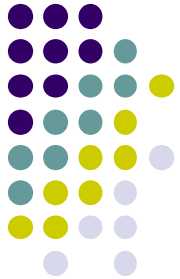
POS Tagging

```
for i in range(len(doc)):
  for token in doc[i]:
    print(token.text, token.pos_)
  print('************************************************')
```

```
50 NUM
year NOUN
old ADJ
female ADJ
presents NOUN
after ADP
having AUX
fallen VERB
in ADP
the DET
bathtub NOUN
4 NUM
days NOUN
ago ADV
and CCONJ
hitting VERB
the DET
back NOUN
of ADP
her PRON
head NOUN
. PUNCT
Since SCONJ
then ADV
she PRON
has AUX
had VERB
a DET
```

- ADJ: adjective
- ADP: adposition
- ADV: adverb
- AUX: auxiliary
- CCONJ: coordinating conjunction
- DET: determiner
- INTJ: interjection
- NOUN: noun
- NUM: numeral
- PART: particle
- PRON: pronoun
- PROPN: proper noun
- PUNCT: punctuation
- SCONJ: subordinating conjunction
- SYM: symbol
- VERB: verb
- X: other

Named Entities



Named Entities NEEDS MODEL

A named entity is a "real-world object" that's assigned a name – for example, a person, a country, a product or a book title. spaCy can **recognize** various **types** of named entities in a document, by asking the model for a **prediction**. Because models are statistical and strongly depend on the examples they were trained on, this doesn't always work *perfectly* and might need some tuning later, depending on your use case.

Named Entities

```
doc = nlp("Apple is looking at buying U.K. startup for #1 billion.")
for ent in doc.ents:
  print(ent.text, ent.start_char, ent.end_char, ent.label_)
```

```
Apple 0 5 ORG
U.K. 27 31 GPE
#1 billion 44 54 MONEY
```

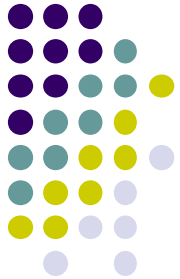
TEXT	START	END	LABEL	DESCRIPTION
Apple	0	5	ORG	Companies, agencies, institutions.
U.K.	27	31	GPE	Geopolitical entity, i.e. countries, cities, states.
$1 billion	44	54	MONEY	Monetary values, including unit.

Entity recognition

```
doc = []
for i in range(len(notes)):
  doc.append(nlp(notes[i]))
  for ent in doc[-1].ents:
    print(ent.text, ent.start_char, ent.end_char, ent.label_)
  print("************************************************************")
```

```
50 year old 0 11 DATE
4 days ago 63 73 DATE
Tylenol 178 185 ORG
the day 279 286 DATE
night 331 336 TIME
************************************************************
23 0 2 CARDINAL
Lupus 14 19 GPE
ESRD 21 25 ORG
RUE 71 74 ORG
HA 197 199 ORG
BP 228 230 ORG
un 252 254 ORG
1131 256 260 DATE
BP 272 274 ORG
10-20 312 317 CARDINAL
BP 369 371 ORG
BP 402 404 ORG
this past Friday 435 451 DATE
BP 130/70 457 466 ORG
today 560 565 DATE
GU/GI 727 732 ORG
ED 759 761 GPE
************************************************************
62yo 0 4 CARDINAL
Afib 19 23 PERSON
```

Entity Visualizer



```
# Entity Visualizer
from spacy import displacy
for i in range(len(doc)):
  displacy.render(doc[i], style="ent", jupyter=True)
  print('**********************************************************************************************************************************************************************')
```

50 year old DATE female presents after having fallen in the bathtub 4 days ago DATE and hitting the back of her head. Since then she has had a massive headache" which did not resolve with Tylenol ORG . She states that she has a high threshold for pain and did not realize how bad it was during the day DATE while at work but then when she got home at night TIME she noticed it. The patient noticed ""silvery specks"" in her vision and she had trouble with some simple tasks like finding the tags on the back of her clothing in the morning. She reported that she had to check several times to make sure she did not put her clothes on backwards. She has had some dizziness, but no nausea or vomiting. Her speech has not been affected.

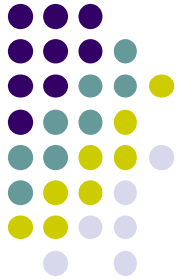
****

23 CARDINAL F with h/o Lupus GPE , ESRD ORG not on HD (planned PD), labile hypertension, RUE ORG VTE on anticoagulation, recent facial swelling who presents with hypertensive emergency. Patient developed severe frontal HA ORG last evenening, a/w nausea. BP ORG was not [**Location (un ORG) 1131 DATE **] on home BP ORG cuff. In the AM her HA was severe [** 10-20 CARDINAL **] a/w nausea and vomiting yellow/green liquid and BP ORG cuff again not able to obtain BP ORG . Patient was last seen by VNA this past Friday DATE with BP 130/70 ORG . Patient denies any CP, shortness of breath, abd pain. Her facial swelling is slightly worse today DATE . She denies any weakness, dizziness, difficulty with speach, no numbness or tingling. She says that she is compliant with all of her medications. She denies any GU/GI ORG complaints despite +UA in ED GPE .

****

62yo CARDINAL RH M h/o HTN, Afib PERSON on coumadin, CAD ORG s/p PTCA x ORG 2 CARDINAL in [** 2178 DATE **] who was in USOH ORG today DATE when he began to have a mild right-sided headache around noon, a/w some photophobia and nausea. He was sitting at the computer half an hour later TIME when he got up to go to [**Company 7546**] and noticed that his L foot was numb and "wobbly". The foot felt weak, "like it was asleep". He walked but was tripping and went to get a banana, thinking that he needed to eat something. He took his pulse which was regular and went to drive to [**Company 7546**]. While driving though, he felt confused and turned around. When he got home, he called his daughter and asked her if his speech was slurred, thinking he may be having a stroke. It was not and he had no difficulty speaking or

Sentence identifier



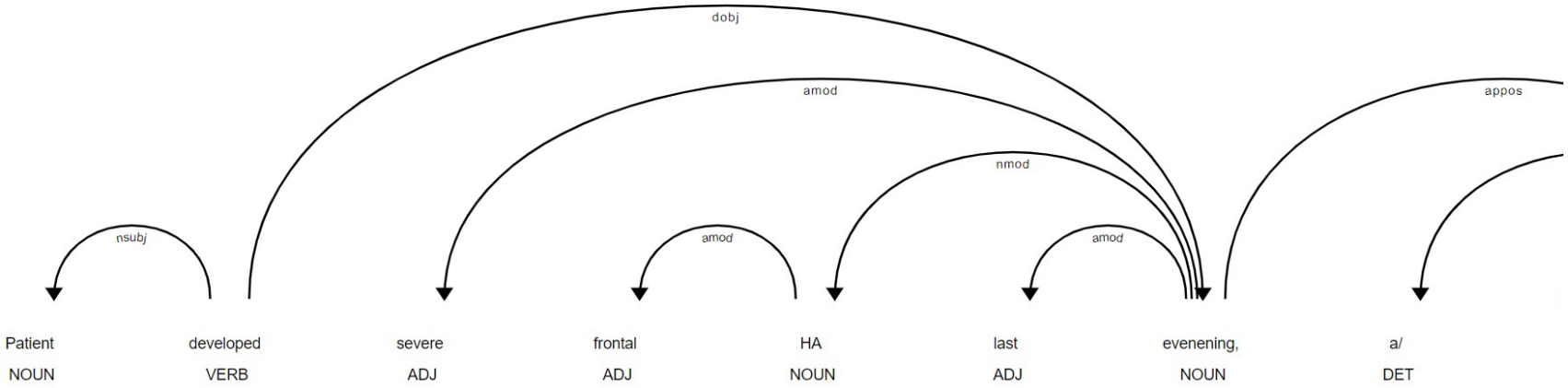
```
#sentence identifier
for i in range(len(doc)):
  for ix, sent in enumerate(doc[i].sents, 1):
    print("Sentence number {}:{}".format(ix, sent))
  print('********************************************************************************************')
```

```
Sentence number 1:50 year old female presents after having fallen in the bathtub 4 days ago and hitting the back of her head.
Sentence number 2:Since then she has had a massive headache" which did not resolve with Tylenol.
Sentence number 3:She states that she has a high threshold for pain and did not realize how bad it was during the day while at work
Sentence number 4:but then when she got home at night she noticed it.
Sentence number 5:The patient noticed ""silvery specks"" in her vision and she had trouble with some simple tasks like finding the tags on the back of her clothing in the morning.
Sentence number 6:She reported that she had to check several times to make sure she did not put her clothes on backwards.
Sentence number 7:She has had some dizziness, but no nausea or vomiting.
Sentence number 8:Her speech has not been affected.

********************************************************************************
Sentence number 1:23 F with h/o Lupus, ESRD not on HD (planned PD), labile hypertension, RUE VTE on anticoagulation, recent facial swelling who presents with hypertensive emergency.
Sentence number 2:Patient developed severe frontal HA last evenening, a/w nausea.
Sentence number 3:BP was not [**Location (un) 1131**] on home BP cuff.
Sentence number 4:In the AM her HA was severe [**10-20**] a/w nausea and vomiting yellow/green liquid and BP cuff again not able to obtain BP.
Sentence number 5:Patient was last seen by VNA this past Friday with BP 130/70.
Sentence number 6:Patient denies any CP, shortness of breath, abd pain.
Sentence number 7:Her facial swelling is slightly worse today.
Sentence number 8:She denies any weakness, dizziness, difficulty with speach, no numbness or tingling.
Sentence number 9:She says that she is compliant with all of her medications.
Sentence number 10:She denies any GU/GI complaints despite +UA in ED.

********************************************************************************
Sentence number 1:62yo RH M h/o HTN, Afib on coumadin, CAD s/p PTCA x 2 in [**2178**] who was in USOH today when he began to have a mild right-sided headache around noon, a/w some pl
Sentence number 2:He was sitting at the computer half an hour later when he got up to go to [**Company 7546**] and noticed that his L foot was numb and "wobbly".
Sentence number 3:The foot felt weak, "like it was asleep".
Sentence number 4:He walked but was tripping and went to get a banana, thinking that he needed to eat something.
Sentence number 5:He took his pulse which was regular and went to drive to [**Company 7546**].
```

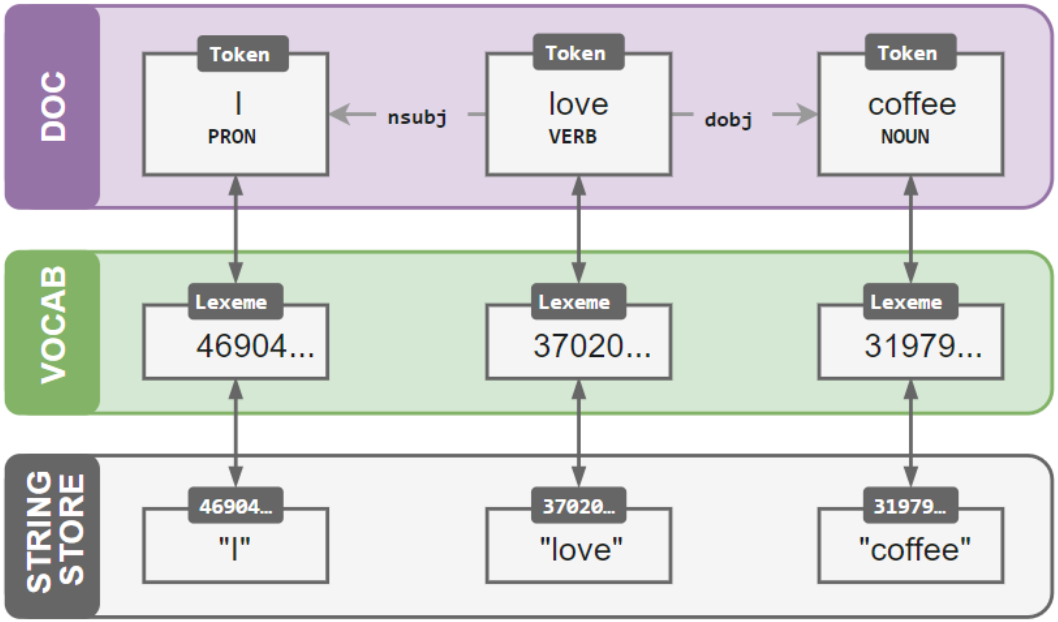
Visualizing dependencies

```
from spacy import displacy
```

```
# dependence tree
for i in range(len(doc)):
  sentence_spans = list(doc[i].sents)
  displacy.render(sentence_spans, style="dep", jupyter=True)
```



Whenever possible, spaCy tries to store data in a vocabulary, the `Vocab` , that will be **shared by multiple documents**. To save memory, spaCy also encodes all strings to **hash values** – in this case for example, "coffee" has the hash `3197928453018144401`. Entity labels like "ORG" and part-of-speech tags like "VERB" are also encoded. Internally, spaCy only "speaks" in hash values.



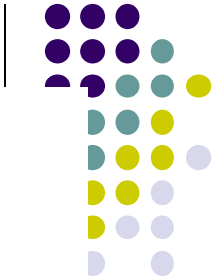
Token: A word, punctuation mark etc. *in context*, including its attributes, tags and dependencies.

Lexeme: A "word type" with no context. Includes the word shape and flags, e.g. if it's lowercase, a digit or punctuation.

Doc: A processed container of tokens in context.

Vocab: The collection of lexemes.

StringStore: The dictionary mapping hash values to strings, for example `3197928453018144401` → "coffee".

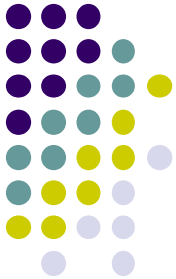


If you process lots of documents containing the word “coffee” in all kinds of different contexts, storing the exact string “coffee” every time would take up way too much space. So instead, spaCy hashes the string and stores it in the `StringStore` . You can think of the `StringStore` as a **lookup table that works in both directions** – you can look up a string to get its hash, or a hash to get its string:

Editable Code spaCy v2.2.0 · Python 3 · via Binder

```
import spacy

nlp = spacy.load("en_core_web_sm")
doc = nlp("I love coffee")
print(doc.vocab.strings["coffee"])  # 3197928453018144401
print(doc.vocab.strings[3197928453018144401])  # 'coffee'
```

RUN



Editable Code spaCy v2.2.0 · Python 3 · via Binder

```
import spacy

nlp = spacy.load("en_core_web_sm")
doc = nlp("I love coffee")
for word in doc:
    lexeme = doc.vocab[word.text]
    print(lexeme.text, lexeme.orth, lexeme.shape_, lexeme.prefix_, lexeme.suffix_,
            lexeme.is_alpha, lexeme.is_digit, lexeme.is_title, lexeme.lang_)
```

RUN

TEXT	ORTH	SHAPE	PREFIX	SUFFIX	IS_ALPHA	IS_DIGIT
I	4690420944186131903	X	I	I	True	False
love	3702023516439754181	xxxx	l	ove	True	False
coffee	3197928453018144401	xxxx	c	fee	True	False

```
doc = nlp("I love apple")
for word in doc:
  lexeme = doc.vocab[word.text]
  print(lexeme.text, lexeme.orth, lexeme.shape_, lexeme.prefix_, lexeme.suffix_, lexeme.is_alpha, lexeme.is_digit, lexeme.is_title,
```

```
I 4690420944186131903 X I I True False True en
love 3702023516439754181 xxxx l ove True False False en
apple 8566208034543834098 xxxx a ple True False False en
```

https://towardsdatascience.com/using-scispacy-for-named-entity-recognition-785389e7918d

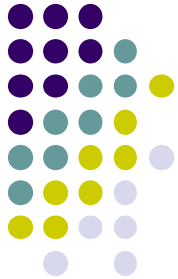
SCISPACY

SciSpacy

- Spacy is not good at extracting entities in biomedical domain.
- SciSpacy is specialized for biomedical text processing

https://allenai.github.io/scispacy/

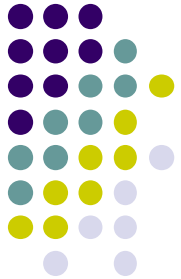
Installing SciSpacy



```
#install Scispacy
!pip install -U spacy
!pip install scispacy
```

```
!pip install https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.2.4/en_core_sci_sm-0.2.4.tar.gz
!pip install https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.2.4/en_ner_craft_md-0.2.4.tar.gz
!pip install https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.2.4/en_ner_jnlpba_md-0.2.4.tar.gz
!pip install https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.2.4/en_ner_bc5cdr_md-0.2.4.tar.gz
!pip install https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.2.4/en_ner_bionlp13cg_md-0.2.4.tar.gz
!pip install https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.2.4/en_core_sci_lg-0.2.4.tar.gz
```

Pick up a pretrained model



- After installing scispaCy, you next need to install one of their pre-trained models.
- scispaCy models come in two flavors: Core and NER.
  - The Core models come in three sizes (small, medium, large) based on the amount of vocabulary stored, and they identify entities but do not classify them.
  - The NER models, on the other hand, identify and classify entities. There are 4 different NER models built on different entity categories.

Pre-trained Model

Model	Description	Install URL
en_core_sci_sm	A full spaCy pipeline for biomedical data.	Download
en_core_sci_md	A full spaCy pipeline for biomedical data with a larger vocabulary and 50k word vectors.	Download
en_core_sci_lg	A full spaCy pipeline for biomedical data with a larger vocabulary and 600k word vectors.	Download
en_ner_craft_md	A spaCy NER model trained on the CRAFT corpus.	Download
en_ner_jnlpba_md	A spaCy NER model trained on the JNLPBA corpus.	Download
en_ner_bc5cdr_md	A spaCy NER model trained on the BC5CDR corpus.	Download
en_ner_bionlp13cg_md	A spaCy NER model trained on the BIONLP13CG corpus.	Download

```
!pip install https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.2.4/en_core_sci_sm-0.2.4.tar.gz
!pip install https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.2.4/en_ner_craft_md-0.2.4.tar.gz
!pip install https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.2.4/en_ner_jnlpba_md-0.2.4.tar.gz
!pip install https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.2.4/en_ner_bc5cdr_md-0.2.4.tar.gz
!pip install https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.2.4/en_ner_bionlp13cg_md-0.2.4.tar.gz
!pip install https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.2.4/en_core_sci_lg-0.2.4.tar.gz
```

Pre-trained NER model

model	F1	Entity Types
en_ner_craft_md	77.03	GGP, SO, TAXON, CHEBI, GO, CL
en_ner_jnlpba_md	73.45	DNA, CELL_TYPE, CELL_LINE, RNA, PROTEIN
en_ner_bc5cdr_md	84.12	DISEASE, CHEMICAL
en_ner_bionlp13cg_md	79.33	AMINO_ACID, ANATOMICAL_SYSTEM, CANCER, CELL, CELLULAR_COMPONENT, DEVELOPING_ANATOMICAL_STRUCTURE, GENE_OR_GENE_PRODUCT, IMMATERIAL_ANATOMICAL_ENTITY, MULTI-TISSUE_STRUCTURE, ORGAN, ORGANISM, ORGANISM_SUBDIVISION, ORGANISM_SUBSTANCE, PATHOLOGICAL_FORMATION, SIMPLE_CHEMICAL, TISSUE

Spacy vs. scispacy

```
from google.colab import files
uploaded = files.upload()
```

Browse... notes_headache.txt

notes_headache.txt(text/plain) - 9433 bytes, last modified: n/a - 100% done

```
Saving notes_headache.txt to notes_headache (1).txt
```

```
notes = []
with open('notes_headache.txt', 'r') as fin:
  lines = fin.readlines()
  for line in lines:
    notes.append(line)
print(notes)
print(len(notes))
```

```
[ ]  # Entity Visualizer
     from spacy import displacy
     for i in range(len(doc)):
       displacy.render(doc[i], style="ent", jupyter=True)
       print('**********************************************
```

```
# Entity Visualizer
from spacy import displacy
displacy.render(doc, style="ent", jupyter=True)
```

What Spacy can do

50 year old DATE female presents after having fallen in the bathtub 4 days ago DATE and hitting the back of her head. Since then she ha not resolve with Tylenol GPE . She states that she has a high threshold for pain and did not realize how bad it was during the day DATE v home at night TIME she noticed it. The patient noticed ""silvery specks"" in her vision and she had trouble with some simple tasks like finding t the morning TIME . She reported that she had to check several times to make sure she did not put her clothes on backwards. She has had so vomiting. Her speech has not been affected.

Spacy vs. scispacy

What SciSpacy can do

```
import scispacy
import spacy
import en_core_sci_sm
import en_ner_bc5cdr_md
import en_core_sci_lg
nlp = en_core_sci_sm.load()
```

```
[ ]  doc = []
     for i in range(len(notes)):
       doc.append(nlp(notes[i]))
       for ent in doc[-1].ents:
         print(ent.text, ent.start_char, ent.end_char, ent.label_)
       print("**********************************************************************
```

```
finding 454 461 ENTITY
tags 466 470 ENTITY
clothing 490 498 ENTITY
morning 506 513 ENTITY
clothes 597 604 ENTITY
backwards 608 617 ENTITY
dizziness 636 645 ENTITY
nausea 654 660 ENTITY
vomiting 664 672 ENTITY
speech 678 684 ENTITY
affected 698 706 ENTITY
**********************************************************************************
F 3 4 ENTITY
h/o Lupus 10 19 ENTITY
ESRD 21 25 ENTITY
```

```
# Entity Visualizer
from spacy import displacy
for i in range(len(doc)):
  displacy.render(doc[i], style="ent", jupyter=True)
  print("**********************************************************************************************************************************************")
```

50 year ENTITY old female ENTITY presents after having fallen in the bathtub 4 ENTITY days ENTITY ago and hitting the back of her head ENTITY . Since then she has had a massive headache ENTITY " which did not resolve with Tylenol ENTITY . She states ENTITY that she has a high threshold ENTITY for pain ENTITY and did not realize how bad it was during the day ENTITY while at work but then wher she got home ENTITY at night ENTITY she noticed it. The patient ENTITY noticed ""silvery spects"" in her vision ENTITY and she had trouble with some simple tasks like finding ENTITY the tags ENTITY on the back of her clothing ENTITY in the morning ENTITY . She reported that she had to check several times to make sure she did not put her clothes ENTITY on backwards ENTITY . She has had some dizziness ENTITY , but no nausea ENTITY or vomiting ENTITY . Her speech ENTITY has not been affected ENTITY .

****

23 F ENTITY with h/o Lupus ENTITY , ESRD ENTITY not on HD ENTITY (planned PD ENTITY), labile hypertension ENTITY , RUE ENTITY VTE ENTITY on anticoagulation ENTITY , recent facial swelling ENTITY who presents with hypertensive emergency ENTITY . Patient ENTITY developed severe ENTITY frontal HA last evenening ENTITY , a/w nausea ENTITY . BP ENTITY was not [** Location ENTITY (un) 1131**] on home BP ENTITY cuff ENTITY . In the AM ENTITY her HA ENTITY was severe ENTITY [**10-20**] a/w nausea and vomiting yellow/green liquid ENTITY and BP ENTITY cuff ENTITY again not able to obtain BP ENTITY . Patient ENTITY was last seen by VNA ENTITY this past Friday ENTITY with BP ENTITY 130/70. Patient ENTITY denies any CP ENTITY , shortness ENTITY of breath, abd pain ENTITY . Her facial swelling ENTITY is slightly worse today ENTITY . She denies any weakness ENTITY , dizziness ENTITY , difficulty ENTITY with speach ENTITY , no numbness ENTITY or tingling ENTITY . She says that she is compliant ENTITY with all of her medications ENTITY . She denies any GU/GI complaints ENTITY despite + UA ENTITY in E ENTITY .

****

Scispacy

```
# Entity Visualizer
from spacy import displacy
for i in range(len(doc)):
  displacy.render(doc[i], style="ent", jupyter=True)
  print("**********")
```

50 year ENTITY old female ENTITY presents after having fallen in the bathtub 4 ENTITY days ENTITY ago and hitting the back of her head ENTITY . Since then she has had a massive headache ENTITY " which did not resolve with Tylenol ENTITY . She states ENTITY that she has a high threshold ENTITY for pain ENTITY and did not realize how bad it was during the day ENTITY while at work but then when she got home ENTITY at night ENTITY she noticed it. The patient ENTITY noticed ""silvery specks"" in her vision ENTITY and she had trouble with some simple tasks like finding ENTITY the tags ENTITY on the back of her clothing ENTITY in the morning ENTITY . She reported that she had to check several times to make sure she did not put her clothes ENTITY on backwards ENTITY . She has had some dizziness ENTITY , but no nausea ENTITY or vomiting ENTITY . Her speech ENTITY has not been affected ENTITY .

****

23 F ENTITY with h/o Lupus ENTITY , ESRD ENTITY not on HD ENTITY (planned PD ENTITY), labile hypertension ENTITY , RUE ENTITY VTE ENTITY on anticoagulation ENTITY , recent facial swelling ENTITY who presents with hypertensive emergency ENTITY . Patient ENTITY developed severe ENTITY frontal HA last evenening ENTITY , a/w nausea ENTITY . BP ENTITY was not [** Location ENTITY (un) 1131**] on home BP ENTITY cuff ENTITY . In the AM ENTITY her HA ENTITY was severe ENTITY [**10-20**] a/w nausea and vomiting yellow/green liquid ENTITY and BP ENTITY cuff ENTITY again not able to obtain BP ENTITY . Patient ENTITY was last seen by VNA ENTITY this past Friday ENTITY with BP ENTITY 130/70. Patient ENTITY denies any CP ENTITY , shortness ENTITY of breath, abd pain ENTITY . Her facial swelling ENTITY is slightly worse today ENTITY . She denies any weakness ENTITY , dizziness ENTITY , difficulty ENTITY with speach ENTITY , no numbness ENTITY or tingling ENTITY . She says that she is compliant ENTITY with all of her medications ENTITY . She denies any GU/GI complaints ENTITY despite + UA ENTITY in E ENTITY .

****

```
import en_core_sci_lg
nlp = en_core_sci_lg.load()
doc = []
for i in range(len(notes)):
  doc.append(nlp(notes[i]))
  displacy.render(doc, style="ent", jupyter=True)
  print("**********")
```

```
/usr/local/lib/python3.10/dist-packages/spacy/util.py:910: UserWarning: [W095] Model 'en_core_sci_lg' (0.4.0) was trained with spaCy v3.0.1 and may not be 100% compatible with the
```

50 year ENTITY old female ENTITY presents after having fallen ENTITY in the bathtub 4 ENTITY days ENTITY ago and hitting ENTITY the back of her head ENTITY . Since then she has had a massive headache ENTITY " which did not resolve with Tylenol ENTITY . She states ENTITY that she has a high threshold ENTITY for pain ENTITY and did not realize how bad it was during the day ENTITY while at work but then when she got home ENTITY at night ENTITY she noticed it. The patient ENTITY noticed ""silvery specks"" in her vision ENTITY and she had trouble with some simple tasks like finding ENTITY the tags ENTITY on the back of her clothing ENTITY in the morning ENTITY . She reported that she had to check several times to make sure she did not put her clothes ENTITY on backwards ENTITY . She has had some dizziness ENTITY , but no nausea ENTITY or vomiting ENTITY . Her speech ENTITY has not been affected ENTITY .

****

50 year ENTITY old female ENTITY presents after having fallen ENTITY in the bathtub 4 ENTITY days ENTITY ago and hitting ENTITY the back of her head ENTITY . Since then she has had a massive headache ENTITY " which did not resolve with Tylenol ENTITY . She states ENTITY that she has a high threshold ENTITY for pain ENTITY and did not realize how bad it was during the day ENTITY while at work but then when she got home ENTITY at night ENTITY she noticed it. The patient ENTITY noticed ""silvery specks"" in her vision ENTITY and she had trouble with some simple tasks like finding ENTITY the tags ENTITY on the back of her clothing ENTITY in the morning ENTITY . She reported that she had to check several times to make sure she did not put her clothes ENTITY on backwards ENTITY . She has had some dizziness ENTITY , but no nausea ENTITY or vomiting ENTITY . Her speech ENTITY has not been affected ENTITY .

23 F with h/o Lupus, ESRD ENTITY not on HD ENTITY (planned PD ENTITY), labile hypertension ENTITY , RUE ENTITY VTE ENTITY on anticoagulation ENTITY , recent facial swelling ENTITY

Different NERs

```
import en_ner_craft_md
nlp = en_ner_craft_md.load()
doc = []
for i in range(len(notes)):
  doc.append(nlp(notes[i]))
  displacy.render(doc, style="ent", jupyter=True)
  print("**********************************************************************************************************************************************************")
```

total of 2 minutes. Afterwards she was sleepy. The code stroke was cancelled, but emergent neurology consult was then called.

46 yo female with no significant medical problems initially presented to PCP [**Name Initial (PRE) 151**] 2 weeks of dyspnea. She reports two weeks of incresing shortness of breath. SOB is worse with lying flat or bending over and better when she lies on her stomach on two pillows or sits up. She has had minimal dyspnea on exertion, but otherwise is asymptomatic including no headache, nausea, vomiting, abdominal pain, weight loss, fevers or pain. Her PCP at [** Name9 GGP (PRE) **] [** Name9 GGP (PRE) 1459**] ordered a CT that showed a 13x8 cm medistinal mass with invasion of the pericardium, ? compression of the SVC and small left pleural effusion. She was sent to [** Hospital1 GGP 18**] for further workup. Family history includes only father with [** Name2 GGP (NI) 499**] cancer.

17 year old RH boy with history of PANDAS s/p strep infection, cognitive decline over several years, junvenile rheumatoid arthritis, multiple seizures and autoimmune diseases in the family who is being transferred from the MICU to neurology service to rule out epilepsy as a cause of hallucinations and episodes of pain/emesis.Patient was born full term, normal spontaneous vaginal delivery. Mom had UTI and was treated with bactrim x 2 weeks a the beginning of her pregnancy. He had Rh incompatibility, hyperbilirubin, treated with phototherapy CHEBI . Developmental milestones met on time, although he never learned to tie shoes or ride bicycle. Did well all through childhood into high school-intelligent, popular, athletic.

****

50 year old female presents after having fallen in the bathtub 4 days ago and hitting the back of her head. Since then she has had a massive headache" which did not resolve with Tylenol. She states that she has a high threshold for pain and did not realize how bad it was during the day while at work but then when she got home at night she noticed it. The patient noticed ""silvery specks"" in her vision and she had trouble with some simple tasks like finding the tags SO on the back of her clothing in the morning. She reported that she had to check several times to make sure she did not put her clothes on backwards. She has had some dizziness, but no

```
import en_ner_bc5cdr_md
nlp = en_ner_bc5cdr_md.load()
doc = []
for i in range(len(notes)):
  doc.append(nlp(notes[i]))
  displacy.render(doc, style="ent", jupyter=True)
  print("**********************************************************************************************************************************************************")
```

```
/usr/local/lib/python3.10/dist-packages/spacy/util.py:910: UserWarning: [W095] Model 'en_ner_bc5cdr_md' (0.4.0) was trained with spaCy v3.0.1 and may not be 100% compatible with the
```

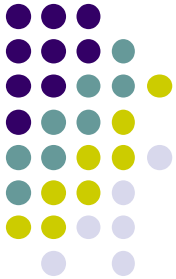
50 year old female presents after having fallen in the bathtub 4 days ago and hitting the back of her head. Since then she has had a massive headache DISEASE " which did not resolve with Tylenol CHEMICAL . She states that she has a high threshold for pain DISEASE and did not realize how bad it was during the day while at work but then when she got home at night she noticed it. The patient noticed ""silvery specks"" in her vision and she had trouble with some simple tasks like finding the tags on the back of her clothing in the morning. She reported that she had to check several times to make sure she did not put her clothes on backwards. She has had some dizziness DISEASE , but no nausea DISEASE or vomiting DISEASE . Her speech has not been affected.

****

50 year old female presents after having fallen in the bathtub 4 days ago and hitting the back of her head. Since then she has had a massive headache DISEASE " which did not resolve with Tylenol CHEMICAL . She states that she has a high threshold for pain DISEASE and did not realize how bad it was during the day while at work but then when she got home at night she noticed it. The patient noticed ""silvery specks"" in her vision and she had trouble with some simple tasks like finding the tags on the back of her clothing in the morning. She reported that she had to check several times to make sure she did not put her clothes on backwards. She has had some dizziness DISEASE , but no nausea DISEASE or vomiting DISEASE . Her speech has not been affected.

23 F with h/o Lupus, ESRD DISEASE not on HD DISEASE (planned PD), labile hypertension DISEASE , RUE VTE DISEASE on anticoagulation, recent facial swelling who presents with hypertensive DISEASE emergency. Patient developed severe frontal HA last evenening, a/w nausea DISEASE . BP was not [**Location (un) 1131**] on home BP cuff. In the AM her HA was severe [**10-20**] a/w nausea DISEASE and vomiting

Different NERs



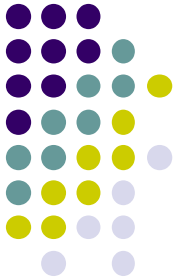
```
import en_ner_bionlp13cg_md
nlp = en_ner_bionlp13cg_md.load()
doc = []
for i in range(len(notes)):
  doc.append(nlp(notes[i]))
  displacy.render(doc, style="ent", jupyter=True)
  print("**********************************************************************************************************************************************************")
```

silvery IMMATERIAL_ANATOMICAL_ENTITY spects"" in her vision and she had trouble with some simple tasks like finding the tags on the back of her clothing in the morning. She reported that she had to check several times to make sure she did not put her clothes on backwards. She has had some dizziness, but no nausea or vomiting. Her speech has not been affected.

23 F with h/o Lupus, ESRD not on HD (planned PD), labile hypertension, RUE VTE on anticoagulation, recent facial swelling who presents with hypertensive emergency. Patient ORGANISM developed severe frontal HA last evenening, a/w nausea. BP SIMPLE_CHEMICAL was not [**Location (un) 1131**] on SIMPLE_CHEMICAL home BP SIMPLE_CHEMICAL cuff. In the AM her HA GENE_OR_GENE_PRODUCT was severe [**10-20**] a/w nausea and vomiting yellow/green liquid and BP SIMPLE_CHEMICAL cuff again not able to obtain BP SIMPLE_CHEMICAL . Patient ORGANISM was last seen by VNA ORGANISM this past Friday with BP SIMPLE_CHEMICAL 130/70. Patient ORGANISM denies any CP GENE_OR_GENE_PRODUCT , shortness of breath, abd ORGAN pain. Her facial swelling is slightly worse today. She denies any weakness, dizziness, difficulty with speach, no numbness or tingling. She says that she is compliant with all of her medications. She denies any GU/GI complaints despite +UA in ED.

62yo RH M h/o HTN, Afib on coumadin SIMPLE_CHEMICAL , CAD s/p PATHOLOGICAL_FORMATION PTCA x 2 in [**2178**] who was in USOH GENE_OR_GENE_PRODUCT today when he began to have a mild right-sided headache around noon, a/w some photophobia and nausea. He was sitting at the computer half an hour later when he got up to go to [**Company 7546**] and noticed that his L foot was SIMPLE_CHEMICAL numb and "wobbly". The foot ORGANISM_SUBDIVISION felt weak, "like it was asleep". He walked but was tripping and went to get a banana, thinking that he needed to eat something. He took his pulse which was regular and went to drive to [**Company 7546**]. While driving though, he felt confused and turned around. When he got home, he called his daughter and asked her if his speech was slurred, thinking he may be having a stroke. It was not and he had no difficulty speaking or comprehending what she was saying. She called 911 to get him checked out and he was brought to an OSH. There, the family noticed an increasing left facial droop. His ankle felt better. Head CT showed an ICH and INR GENE_OR_GENE_PRODUCT was 2.89 and the patient ORGANISM was given 3U FFP, vit K GENE_OR_GENE_PRODUCT 5mg IM, labetalol SIMPLE_CHEMICAL 10mg IV x 1 and dilantin 1g IV

Different NERs



```
[ ] import en_ner_bc5cdr_md
    nlp = en_ner_bc5cdr_md.load()
    doc = []
    for i in range(len(notes)):
      doc.append(nlp(notes[i]))
      displacy.render(doc, style="ent", jupyter=True)
      print("**********************************************************************************************************************************************************")
```

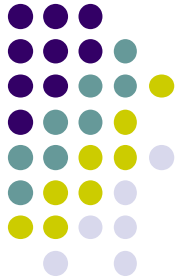
```
/usr/local/lib/python3.10/dist-packages/spacy/util.py:910: UserWarning: [W095] Model 'en_ner_bc5cdr_md' (0.4.0) was trained with spaCy v3.0.1 and may not be 100% compatible with the
```

50 year old female presents after having fallen in the bathtub 4 days ago and hitting the back of her head. Since then she has had a massive headache DISEASE " which did not resolve with Tylenol CHEMICAL . She states that she has a high threshold for pain DISEASE and did not realize how bad it was during the day while at work but then when she got home at night she noticed it. The patient noticed ""silvery spects"" in her vision and she had trouble with some simple tasks like finding the tags on the back of her clothing in the morning. She reported that she had to check several times to make sure she did not put her clothes on backwards. She has had some dizziness DISEASE , but no nausea DISEASE or vomiting DISEASE . Her speech has not been affected.

```
**********************************************************************************************************************************************************
```

50 year old female presents after having fallen in the bathtub 4 days ago and hitting the back of her head. Since then she has had a massive headache DISEASE " which did not resolve with Tylenol CHEMICAL . She states that she has a high threshold for pain DISEASE and did not realize how bad it was during the day while at work but then when she got home at night she noticed it. The patient noticed ""silvery spects"" in her vision and she had trouble with some simple tasks like finding the tags on the back of her clothing in the morning. She reported that she had to check several times to make sure she did not put her clothes on backwards. She has had some dizziness DISEASE , but no nausea DISEASE or vomiting DISEASE . Her speech has not been affected.

23 F with h/o Lupus, ESRD DISEASE not on HD DISEASE (planned PD), labile hypertension DISEASE , RUE VTE DISEASE on anticoagulation, recent facial swelling who presents with hypertensive DISEASE emergency. Patient developed severe frontal HA last evenening, a/w nausea DISEASE . BP was not [**Location (un) 1131**] on home BP cuff. In the AM her HA was severe [**10-20**] a/w nausea DISEASE and vomiting



Thank you!