Named Entity Recognition (NER)

- What is Named Entity Recognition?

It is a subtask of information extraction that seeks to locate and classify named entities in text pro-defined categories such as the names of people, organizations, locations, expressions of time, quantities, monetary values, percentages, etc.

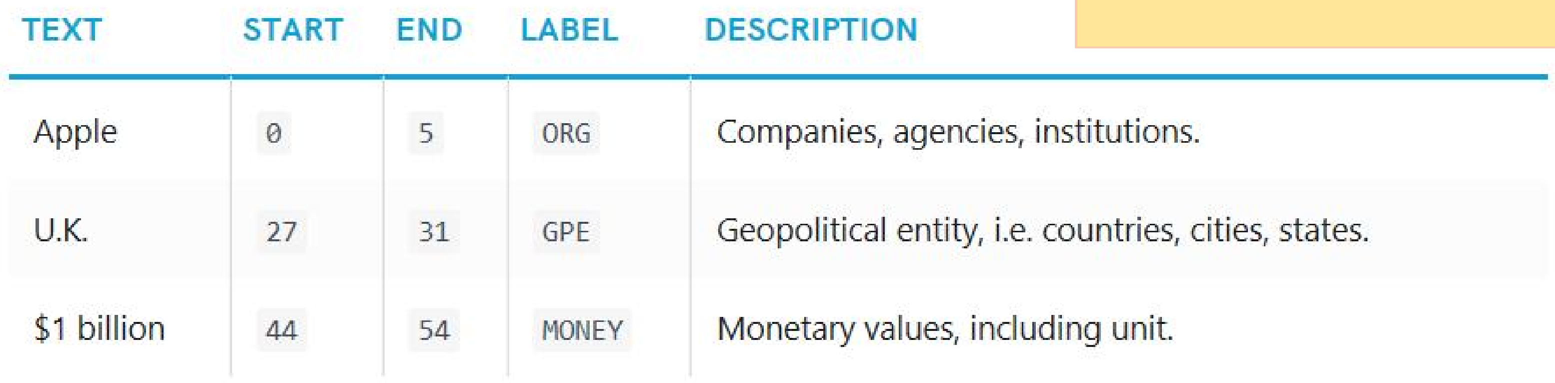
e.g.

$$\begin{tikzcd}
\textbf{Jim} \arrow[d] & \textbf{bought 300 shares of} & \textbf{Acme Corp.} \arrow[d] & \textbf{in} & \textbf{2006} \arrow[d] \\
\text{Person} & & \text{Organization} & & \text{Time}
\end{tikzcd}$$

- ✓ SpaCy features an extremely fast statistical entity recognition system
- ✓ This system assigns labels to spans of tokens
- ✓ The default model identifies a variety of named and numeric entities
- ✓ SpaCy can recognise various types of named entities in a document, by asking the model for a prediction
- ✓Named entities are available as the *ents* property of a *Doc*

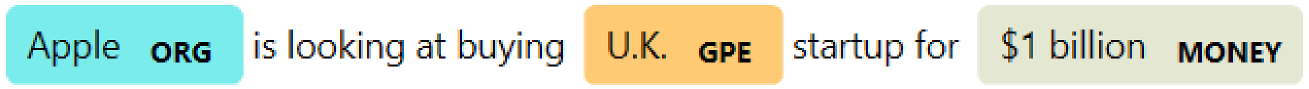
```
 doc = nlp(u ‘Apple is looking at buying U.K. startup for $1 billion‘)

for ent in doc.ents:
        print (ent.text, ent.start_char, ent.end_char, ent.label_)
```

Text: The original entity text
Start: Index of start of entity in the *Doc*
End: Index of end of entity in the *Doc*
Label: Entity label, i.e. type



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.

Here is how our example sentence and its named entities looks like in the *displaCy visualizer*


Apple ORG is looking at buying U.K. GPE startup for $1 billion MONEY

Accessing entity annotations

```
doc = nlp (u ‚San Francisco considers banning sidewalk delivery robots‘)
```

1. By the *doc.ents* property, which produces a sequence *Span* objects

```
ents = [(e.text. E.start_char, e.end_char, e.label_) for e in doc.ents ]
assert ents == [(u ‚San Francisco‘, 0, 13, u ‚GPE‘)]
```

Accessing entity annotations

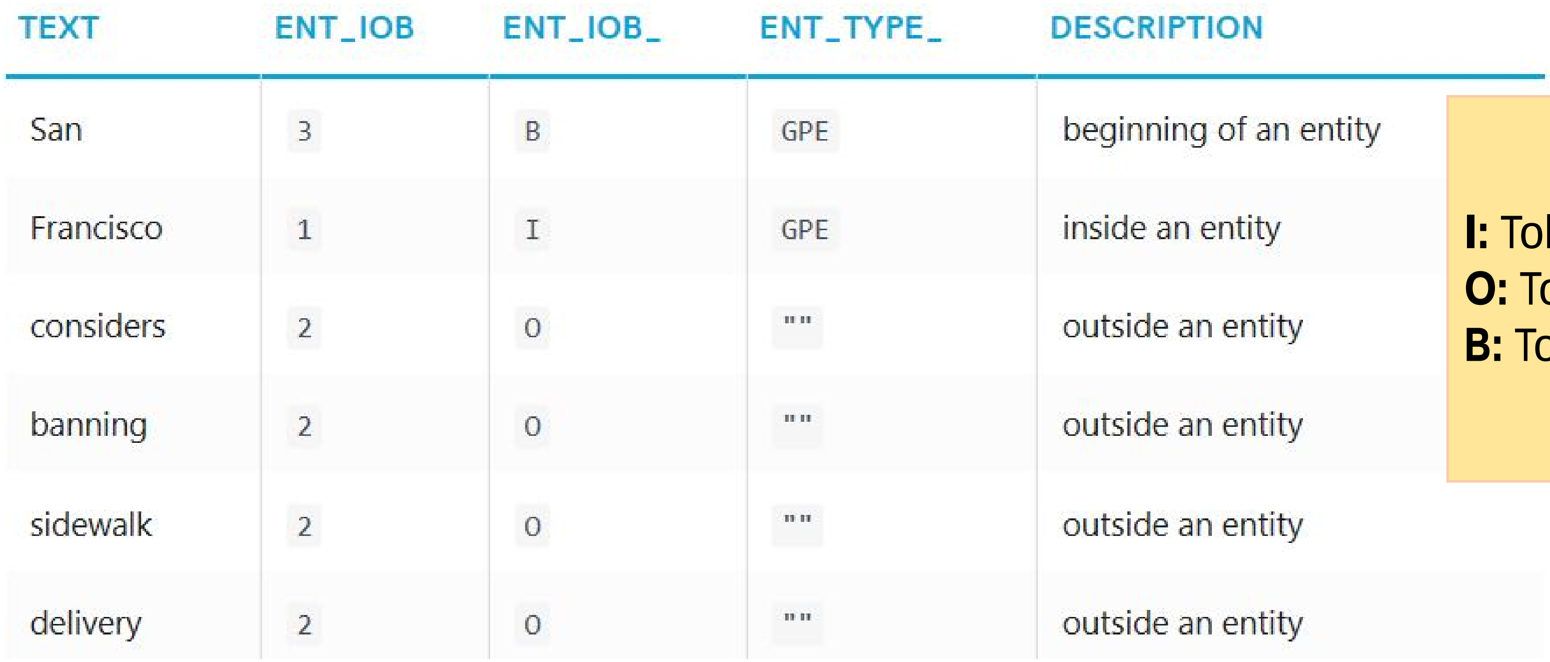
```
doc = nlp (u ‚San Francisco considers banning sidewalk delivery robots‘)
```

1. By the *token.ent_iob* and the *token.ent_type*

- *token.ent_iob* indicates whether an entity starts, continuous or ends on the tag.

```
ent_san = [doc [0].text, doc [0].ent_iob_, doc [0].ent_type_]
ent_francisco = [doc [1].text, doc [1].ent_iob_, doc [1].ent_type_]
assert ent_san == [u ‘San‘, u ‘B‘, u ‘GPE‘]
assert ent_francisco == [u ‘Francisco ‘, u ‘I‘, u ‘GPE‘]
```

```
ent_san = [doc [0].text, doc [0].ent_iob_, doc [0].ent_type_]
ent_francisco = [doc [1].text, doc [1].ent_iob_, doc [1].ent_type_]
assert ent_san == [u ‘San‘, u ‘B‘, u ‘GPE‘]
assert ent_francisco == [u ‘Francisco ‘, u ‘I‘, u ‘GPE‘]
```



TEXT	ENT_IOB	ENT_IOB_	ENT_TYPE_	DESCRIPTION
San	3	B	GPE	beginning of an entity
Francisco	1	I	GPE	inside an entity
considers	2	O	""	outside an entity
banning	2	O	""	outside an entity
sidewalk	2	O	""	outside an entity
delivery	2	O	""	outside an entity

I: Token is inside an entity
O: Token is outside an entity
B: Token is the beginning of an entity

Setting entity annotation

To ensure that the sequence of token annotations remains consistent, you have to set entity annotations at the document level:

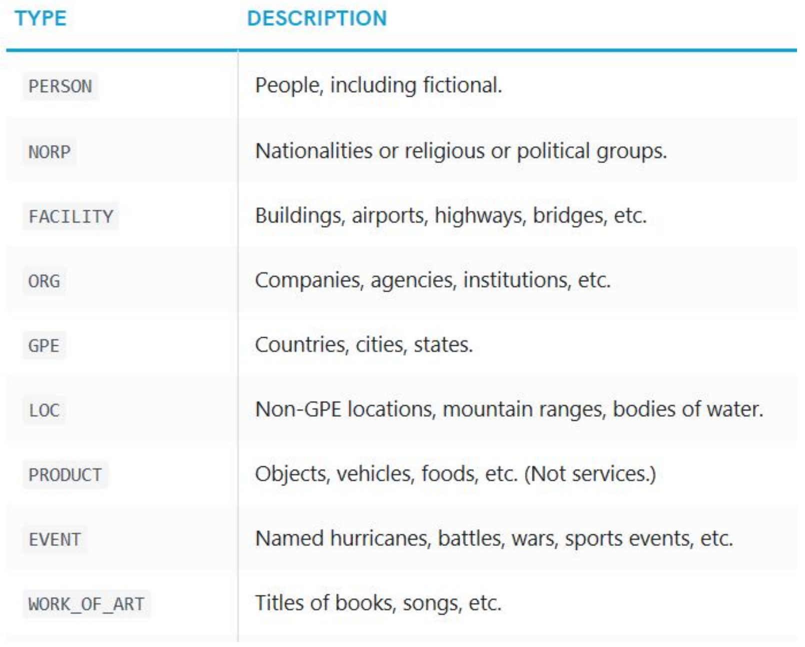
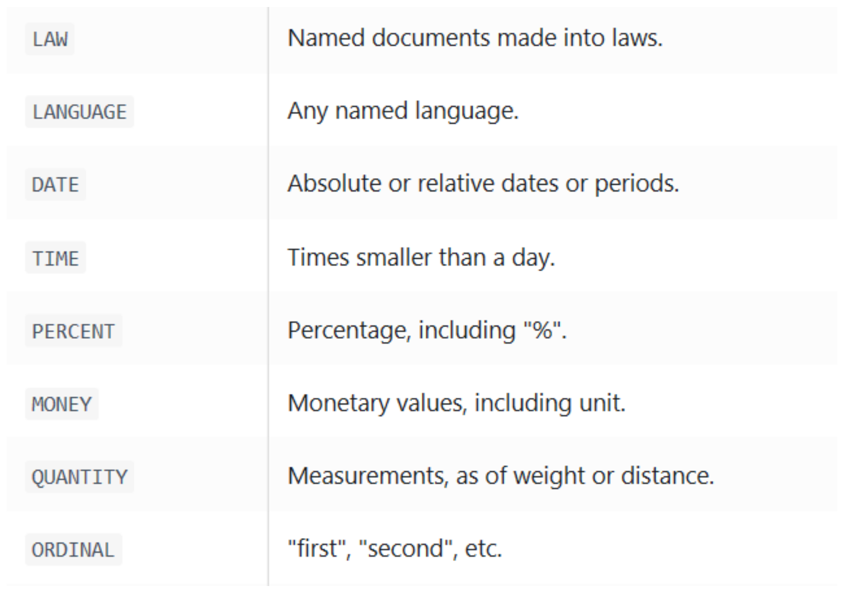
1. By assigning to the *doc.ents* attribute and creating the new entity as a *Span*

2. By using the *doc.from_array()* method (You sould include here both *ENT_TYPE* and *ENT_IOB* attributes in the array you are importing from.

Understanding Entity Types

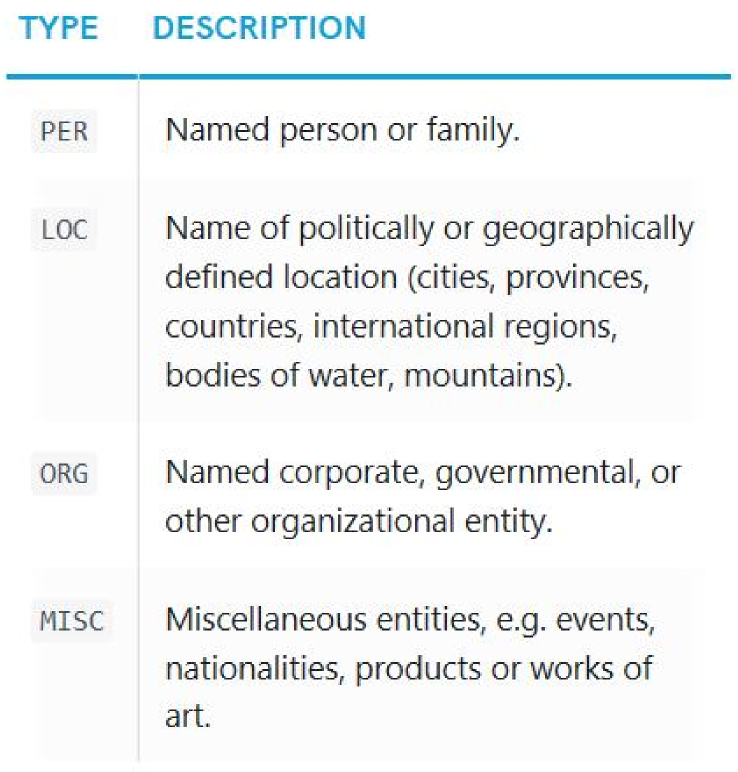
- To get the description for the string representation of an entity label you can use the *spacy.explain()* method

- For example if you use the OntoNotes 5 corpus, which supports following entity types, you would get that result:

TYPE	DESCRIPTION
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FACILITY	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
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.

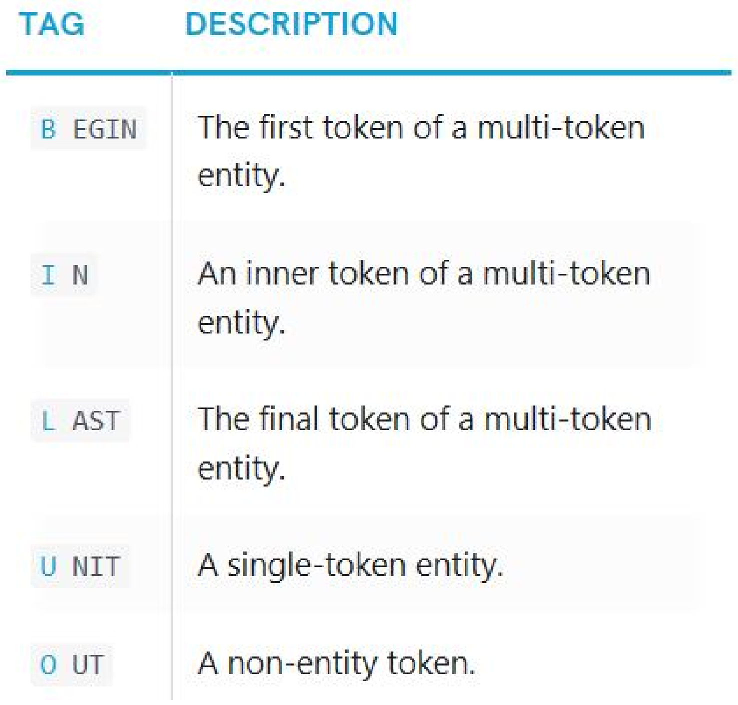
- Models trained on Wikipedia corpus use a less fine-grained NER annotation scheme and recognise that following entities:



TYPE	DESCRIPTION
PER	Named person or family.
LOC	Name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains).
ORG	Named corporate, governmental, or other organizational entity.
MISC	Miscellaneous entities, e.g. events, nationalities, products or works of art.

The BILUO Scheme

- There are several coding schemes for encoding entity annotations as token tags.



TAG	DESCRIPTION
B EGIN	The first token of a multi-token entity.
I N	An inner token of a multi-token entity.
L AST	The final token of a multi-token entity.
U NIT	A single-token entity.
O UT	A non-entity token.

1. SpaCy translates the character offset into this scheme, to decide the cost of each action (given the current state of the entity recognizer)

2. The costs are then used to calculate the gradient of the loss, to train the model.

Why BILUO, not IOB?

- These coding schemes are equally expressive **but**
- **IOB** was more difficult to learn
- **BILUO** marks explicitly boundary tokens

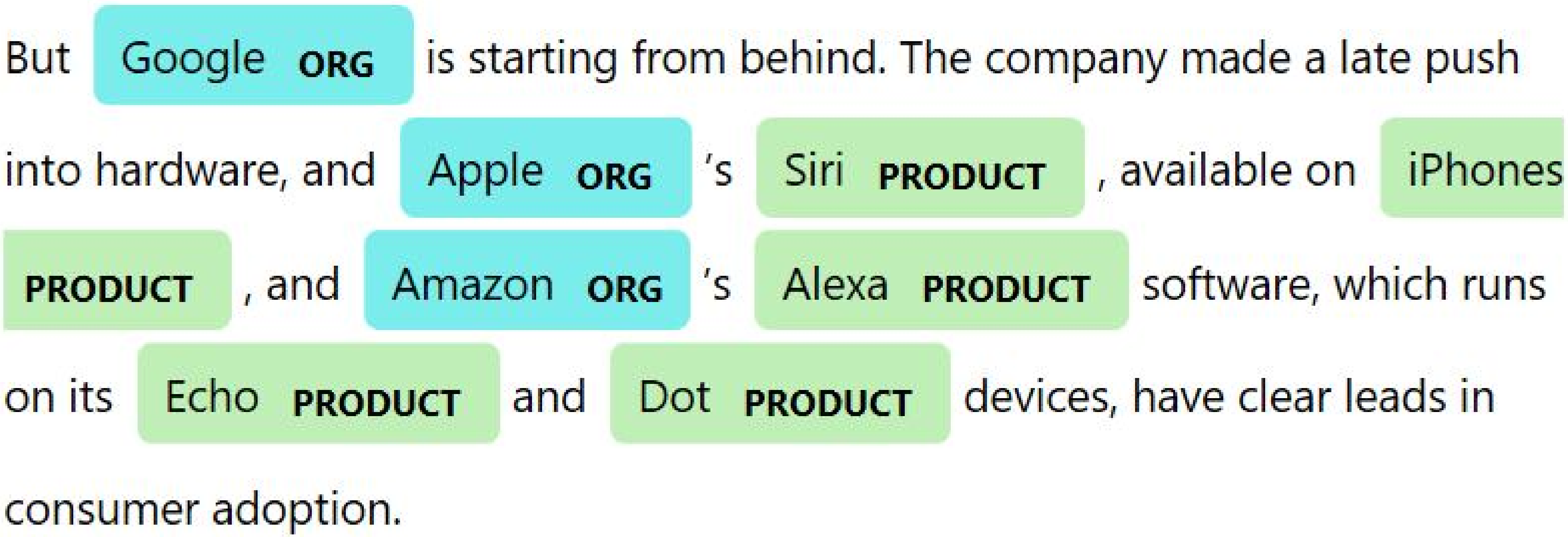
Visualizing named entities

- displaCy ENT visualizer lets you explore an entity recognition model's behavior interactively.

- spaCy v2.0+ comes with a visualizer module to run the visualization yourself

```
import spacy
from spacy import displacy

text = """But Google is starting from behind. The company made a late push into
hardware, and Apple's Siri, available on iPhones, and Amazon's Alexa software, which
runs on ist Echo and Dot devices, have clear leads in consumer adoption."""

nlp = spacy.load ('custom_ner_model')
doc = nlp(text)
Displacy.serve (doc, style = 'ent')
```

1. Pass a *Doc* or a list of *Doc* objects to displaCy and
2. Run *displacy.serve* to run the web server



But Google **ORG** is starting from behind. The company made a late push into hardware, and Apple **ORG** ’s Siri **PRODUCT** , available on iPhones **PRODUCT** , and Amazon **ORG** ’s Alexa **PRODUCT** software, which runs on its Echo **PRODUCT** and Dot **PRODUCT** devices, have clear leads in consumer adoption.