Named Entity Recognition

Katharine Jarmul
Founder, kjamistan
What is Named Entity Recognition?

- NLP task to identify important named entities in the text
  - People, places, organizations
  - Dates, states, works of art
  - ... and other categories!
- Can be used alongside topic identification
  - ... or on its own!
- Who? What? When? Where?
Example of NER

In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. He was visiting the United States when Adolf Hitler came to power in 1933 and did not go back to Germany, where he had been a professor at the Berlin Academy of Sciences. He settled in the US, becoming an American citizen in 1940. On the eve of World War II, he endorsed a letter to President Franklin D. Roosevelt, alerting him to the potential development of "extremely powerful bombs of a new type" and recommending that the US begin similar research. This eventually led to what would become the Manhattan Project. Einstein supported defending the Allied forces, but largely denounced using the new discovery of nuclear fission as a weapon. Later, with the British philosopher Bertrand Russell, Einstein signed the Russell–Einstein Manifesto, which highlighted the danger of nuclear weapons. Einstein was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.

(Source: Europeana Newspapers (http://www.europeana-newspapers.eu))
nltk and the Stanford CoreNLP Library

- The Stanford CoreNLP library:
  - Integrated into Python via nltk
  - Java based
  - Support for NER as well as coreference and dependency trees
Using `nltk` for Named Entity Recognition

```
In [1]: import nltk

In [2]: sentence = '''In New York, I like to ride the Metro to visit MOMA and some restaurants rated well by Ruth Reichl.'''

In [3]: tokenized_sent = nltk.word_tokenize(sentence)

In [4]: tagged_sent = nltk.pos_tag(tokenized_sent)

In [5]: tagged_sent[:3]
Out[5]: [('In', 'IN'), ('New', 'NNP'), ('York', 'NNP')]
```
nltk's `ne_chunk()`

```python
In [6]: print(nltk.ne_chunk(tagged_sent))
(S
   In/IN
   (GPE New/NNP York/NNP)
   ,/,
   I/PRP
   like/VBP
to/TO
   ride/VB
   the/DT
   (ORGANIZATION Metro/NNP)
to/TO
   visit/VB
   (ORGANIZATION MOMA/NNP)
   and/CC
   some/DT
   restaurants/NNS
   rated/VBN
   well/RB
   by/IN
   (PERSON Ruth/NNP Reichl/NNP)
./)
```
Let's practice!
Introduction to SpaCy

Katharine Jarmul
Founder, kjamistan
What is SpaCy?

- NLP library similar to gensim, with different implementations
- Focus on creating NLP pipelines to generate models and corpora
- Open-source, with extra libraries and tools
  - Displacy
Displacy entity recognition visualizer

In New York, I like to ride the Metro to visit MOMA and some restaurants rated well by Ruth Reichl.

(source: https://demos.explosion.ai/displacy-ent/)
SpaCy NER

```python
In [1]: import spacy

In [2]: nlp = spacy.load('en')

In [3]: nlp.entity
Out[3]: <spacy.pipeline.EntityRecognizer at 0x7f76b75e68b8>

In [4]: doc = nlp("""Berlin is the capital of Germany;
     and the residence of Chancellor Angela Merkel."""")

In [5]: doc.ents
Out[5]: (Berlin, Germany, Angela Merkel)

In [6]: print(doc.ents[0], doc.ents[0].label_)
Berlin GPE
```
Why use SpaCy for NER?

- Easy pipeline creation
- Different entity types compared to nltk
- Informal language corpora
  - Easily find entities in Tweets and chat messages
- Quickly growing!
Let's practice!
Multilingual NER with polyglot

Katharine Jarmul
Founder, kjamistan
What is polyglot?

- NLP library which uses word vectors
- Why polyglot?
  - Vectors for many different languages
  - More than 130!
In [1]: from polyglot.text import Text

In [2]: text = "El presidente de la Generalitat de Cataluña, Carles Puigdemont, ha afirmado hoy a la alcaldesa de Madrid, Manuela Carmena, que en su etapa de alcalde de Girona (de julio de 2011 a enero de 2016) hizo una gran promoción de Madrid."

In [3]: ptext = Text(text)

In [4]: ptext.entities
Out[4]:
[I-ORG(['Generalitat', 'de']),
 I-LOC(['Generalitat', 'de', 'Cataluña']),
 I-PER(['Carles', 'Puigdemont']),
 I-LOC(['Madrid']),
 I-PER(['Manuela', 'Carmena']),
 I-LOC(['Girona']),
 I-LOC(['Madrid'])]
Let's practice!