Named Entity Recognition

INTRODUCTION TO NATURAL LANGUAGE PROCESSING IN PYTHON

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 U.S., 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 U.S. 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.

Tag colours:
LOCATION   TIME   PERSON   ORGANIZATION   MONEY   PERCENT   DATE

(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

```python
import nltk
sentence = '''In New York, I like to ride the Metro to
               visit MOMA and some restaurants rated
               well by Ruth Reichl.'''
tokenized_sent = nltk.word_tokenize(sentence)
tagged_sent = nltk.pos_tag(tokenized_sent)
tagged_sent[:3]
```

```
[('In', 'IN'), ('New', 'NNP'), ('York', 'NNP')]
```

```
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!

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 `GPE` , I like to ride the Metro to visit MOMA `ORG` and some restaurants rated well by Ruth Reichl `PERSON`

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

```
import spacy
nlp = spacy.load('en')
nlp.entity
```

```
<spacy.pipeline.EntityRecognizer at 0x7f76b75e68b8>
```

```
doc = nlp("""Berlin is the capital of Germany;
                and the residence of Chancellor Angela Merkel.""")
doc.ents
```

```
(Berlin, Germany, Angela Merkel)
```

```
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

INTRODUCTION TO NATURAL LANGUAGE PROCESSING IN PYTHON

Katharine Jarmul
Founder, kjamistan

datacamp

What is polyglot?

- NLP library which uses word vectors

- Why `polyglot` ?
  - Vectors for many different languages
  - More than 130!

```
which          ويكه
India          ينديا
beat           بيت
Bermuda        بيرمودا
in             ين
Port           بورت
of             وف
Spain          سباين
in             ين
2007

,
which          ويكه
was            واس
equalled       يكالليد
five           فيفي
days           دايس
ago            اغو
by             يي
South          سووث
Africa         افريكا
in             ين
their          ثير
victory        فيكتوري
over           وفير
West           ويست
Indies         يندييس
in             ين
Sydney         سيدني
.
```

Spanish NER with polyglot

```python
from polyglot.text import Text
?ext = """"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."""
ptext = Text(text)
ptext.entities
```

```
[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!