# Analisi linguistica: spaCy

Rachele Sprugnoli

rachele.sprugnoli@unipr.it

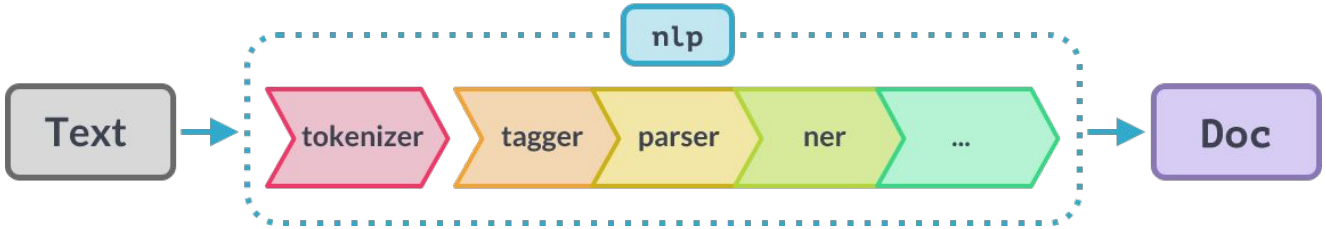

UNIVERSITÀ
DI PARMA

# spaCy

- spaCy: `import spacy`
  - libreria open source multilingue
  - documentazione: https://spacy.io/
  - lo useremo per elaborazioni automatiche più avanzate (NLP)
  - modelli NLP per molte lingue: modelli computazionali che consentono al computer di acquisire le competenze necessarie per elaborare/riprodurre le lingue naturali

# spaCy: pipeline

- Dettagli: https://spacy.io/usage/spacy-101




  - Tokenization
  - Sentence Boundary Detection (SBD)
  - Lemmatization
  - Part-of-speech (POS) Tagging
  - Dependency Parsing
  - Named Entity Recognition (NER)

# spaCy: modelli

- Lista: https://spacy.io/usage/models
- Modelli per l'italiano: https://spacy.io/models/it
  - Addestrati su testi giornalistici, legali, Wikipedia
  - Tokenizzazione, divisione in frasi, lemmatizzazione, analisi morfologica, attribuzione parti del discorso, analisi sintattica a dipendenze, classificazione delle entità nominate (NER, Named Entity Recognition)
  - Per analisi morfologica, parti del discorso e analisi sintattica: https://universaldependencies.org/
  - Per NER: PER, ORG, LOC, MISC ("Miscellaneous entities, e.g., events, nationalities, products, or works of art.")

# spaCy: caricare un modello

```
!python -m spacy download nome_modello
import spacy
nlp = spacy.load('nome_modello')
doc=nlp(testo)
```

---

```
!python -m spacy download it_core_news_sm
import spacy
nlp = spacy.load("it_core_news_sm")
doc=nlp("Tant'è amara che poco è più morte;")
```

# spaCy: tokenizzazione

- `nlp.tokenizer.explain("testo")` → spiega come è avvenuta la tokenizzazione

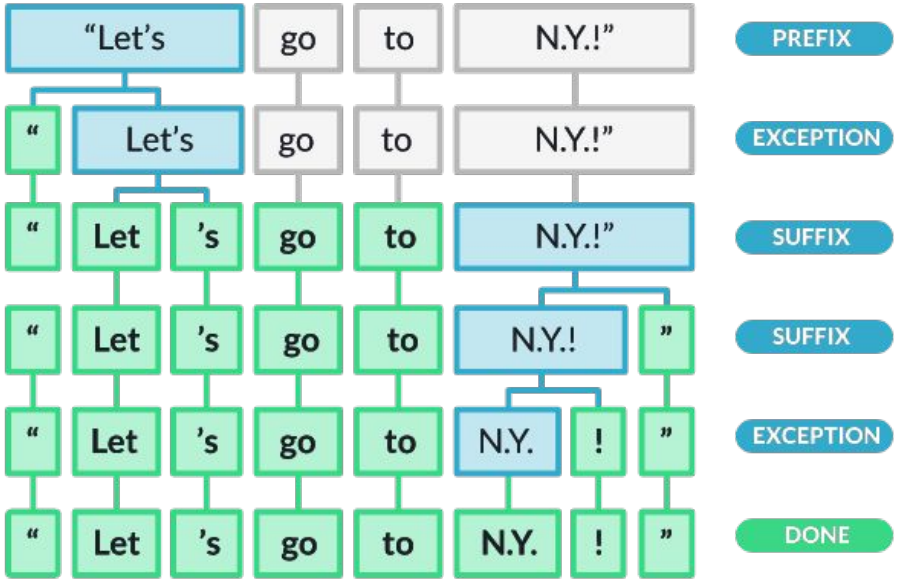



Immagine da: https://spacy.io/usage/spacy-101

# spaCy: tokenizzazione, esempi di attributi

Lista completa: https://spacy.io/api/attributes

- `token.text` → testo del token
- `token.sent` → frase in cui si trova il token
- `token.lemma_` → lemma del token
- `token.pos_` → parte del discorso del token
- `token.morph` → analisi morfologica
- `token.dep_` → analisi sintattica
- `token.ent_iob_` → il token fa parte o meno di un'entità nominata
- `token.ent_type_` → classificazione dell'entità nominata

# spaCy: tokenizzazione, attributi

- `token.lower_` → forma tutta in minuscolo del token
- `token.shape_` → forma ortografica del token, es. Xxxx / dd
- `token.is_alpha` → il token è formato da caratteri alfabetici?
- `token.is_punct` → il token è un segno di punteggiatura?
- `token.is_bracket` → il token è una parentesi?
- `token.is_stop` → il token è una stopword?
- `token.currency` → il token è un simbolo di valuta?
- `token.like_url_` → il token sembra una URL?
- `token.like_email_` → il token sembra un indirizzo mail?

# spaCy: NER

- NER = Named Entity Recognition

| 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 | 0 | "" | outside an entity |
| banning | 2 | 0 | "" | outside an entity |
| sidewalk | 2 | 0 | "" | outside an entity |
| delivery | 2 | 0 | "" | outside an entity |
| robots | 2 | 0 | "" | outside an entity |

Immagine da: https://spacy.io/usage/linguistic-features

# spaCy: visualizzazioni

- Modulo `displaCy`: https://demos.explosion.ai/displacy
- Deve essere importato

```
import spacy
from spacy import displacy
```

  - `displacy.render(testo, style="dep")` → visualizzazione degli alberi sintattici
  - `displacy.render(testo, style="ent")` → visualizzazione del NER
  - `displacy.render([testo], style="ent", page=True)` → visualizzazione in HTML (`style="dep"` per l'analisi sintattica)

# Un po’ di pratica

- Lezione10.ipynb

  https://colab.research.google.com/drive/1z-5GgFy0gTkBOGzypEK3Fqo5Qpt62Rf4?usp=sharing

# Esercizio 1

- Analizzare una frase a piacere (anche in una lingua diversa dall'italiano):
  - per ogni token identificare se è alfabetico, parentesi, stopword, attribuire la parte del discorso, classificare le entità
  - salvare l'output in un file

# Esercizio 2

- Visualizzare l'output del NER per il file News.txt in un file HTML

# Soluzione esercizi

- https://colab.research.google.com/drive/1RyCO-ZvBzPwxZGpSLfie8XfesCWIO2aA?usp=sharing