Reconhecimento de Entidades Nomeadas

Processamento de Linguagem Natural

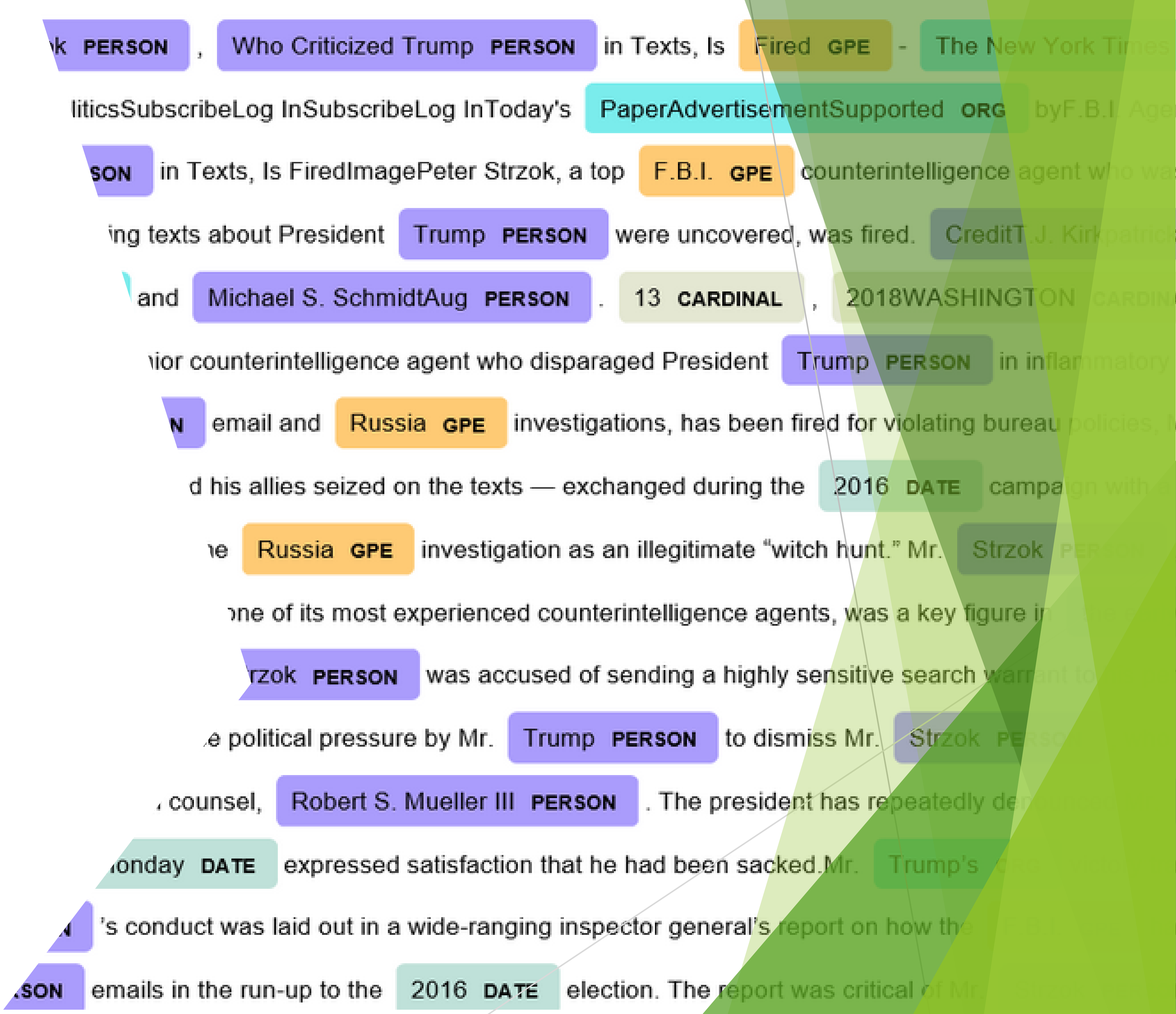
Prof. Leandro Alvim, D. Sc.

Agenda

- Definição
- Congressos
- Corpora
- Estado da Arte
- Trabalhos em HD
- Bibliotecas
- Ferramentas de Anotação
- Implementações

Reconhecimento de Entidades Nomeadas

- Subtarefa da extração da informação que consiste de encontrar e classificar menções de entidades nomeadas em um texto não estruturado a partir de categorias pre definidas como: pessoa, lugar, organização, ...



k PERSON , Who Criticized Trump PERSON in Texts, Is Fired GPE - The New York Times
liticsSubscribeLog InSubscribeLog InToday's PaperAdvertisementSupported ORG byF.B.I. Age
SON in Texts, Is FiredImagePeter Strzok, a top F.B.I. GPE counterintelligence agent who wa
ing texts about President Trump PERSON were uncovered, was fired. CreditT.J. Kirkpatrick
and Michael S. SchmidtAug PERSON . 13 CARDINAL , 2018WASHINGTON CARDIN
ior counterintelligence agent who disparaged President Trump PERSON in inflammatory
N email and Russia GPE investigations, has been fired for violating bureau policies,
d his allies seized on the texts — exchanged during the 2016 DATE campaign with
ne Russia GPE investigation as an illegitimate "witch hunt." Mr. Strzok PER
ne of its most experienced counterintelligence agents, was a key figure in
rzok PERSON was accused of sending a highly sensitive search warrant to
e political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON
counsel, Robert S. Mueller III PERSON . The president has repeatedly de
onday DATE expressed satisfaction that he had been sacked.Mr. Trump's
N 's conduct was laid out in a wide-ranging inspector general's report on how the
SON emails in the run-up to the 2016 DATE election. The report was critical of Mr.

Congressos

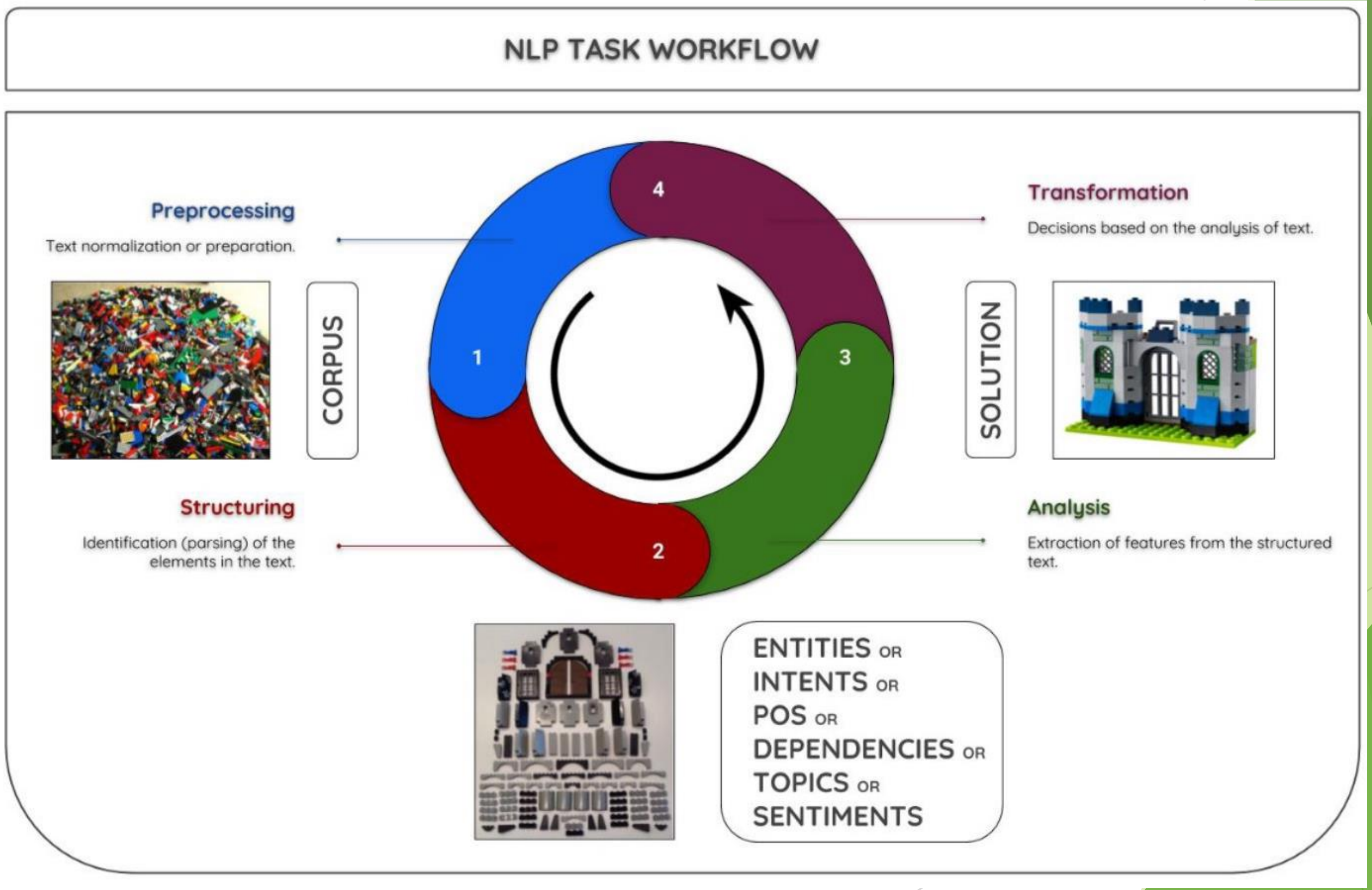
- MUC-7
- CoNLL 2003 NER task
- WNUT 2017 Emerging Entities task
- IberLEF 2019

Corpora

- Inglês
  - CoNLL 2003
    - PER, LOC, ORG, MISC
    - Reuters RCV1 corpus
  - OntoNotes Corpus v5
    - 18 tags
      - 12 tipos: PER, LOC, ORG, ...
      - 6 valores: data, percentual, dinheiro, ...
  - W-NUT2017
    - 6 tags: PER, LOC, Creative Work, group, ...
    - Palavras marcadas com baixa repetição

Corpora

- Português
  - HAREM
  - WikiNER
  - Paramopama
  - leNER-br
  - Peres-2017



NLP TASK WORKFLOW

Preprocessing
Text normalization or preparation.

Structuring
Identification (parsing) of the elements in the text.

Analysis
Extraction of features from the structured text.

Transformation
Decisions based on the analysis of text.

ENTITIES OR
INTENTS OR
POS OR
DEPENDENCIES OR
TOPICS OR
SENTIMENTS

Estado da Arte

- CoNLL 2003 NER task

Modelo	F1	Artigo	Código
CNN Large + fine-tune (Baevski et al., 2019)	93.5	Cloze-driven Pretraining of Self-attention Networks	
Flair embeddings (Akbik et al., 2018)	93.09	Contextual String Embeddings for Sequence Labeling	Flair framework
BERT Large (Devlin et al., 2018)	92.8	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	
CVT + Multi-Task (Clark et al., 2018)	92.61	Semi-Supervised Sequence Modeling with Cross-View Training	Official
BERT Base (Devlin et al., 2018)	92.4	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	
BiLSTM-CRF+ELMo (Peters et al., 2018)	92.22	Deep contextualized word representations	AllenNLP Project AllenNLP GitHub
Peters et al. (2017)	91.93	Semi-supervised sequence tagging with bidirectional language models	

http://nlpprogress.com/english/named_entity_recognition.html

Estado da Arte

- OntoNotes corpus v5

Modelo	F1	Artigo	Código
Flair embeddings (Akbik et al., 2018)	89.71	Contextual String Embeddings for Sequence Labeling	Official
CVT + Multi-Task (Clark et al., 2018)	88.81	Semi-Supervised Sequence Modeling with Cross-View Training	Official
Bi-LSTM-CRF + Lexical Features (Ghaddar and Langlais 2018)	87.95	Robust Lexical Features for Improved Neural Network Named-Entity Recognition	Official
BiLSTM-CRF (Strubell et al, 2017)	86.99	Fast and Accurate Entity Recognition with Iterated Dilated Convolutions	Official
Iterated Dilated CNN (Strubell et al, 2017)	86.84	Fast and Accurate Entity Recognition with Iterated Dilated Convolutions	Official
Chiu and Nichols (2016)	86.28	Named entity recognition with bidirectional LSTM-CNNs	
Joint Model (Durrett and Klein 2014)	84.04	A Joint Model for Entity Analysis: Coreference, Typing, and Linking	
Averaged Perceptron (Ratinov and Roth 2009)	83.45	Design Challenges and Misconceptions in Named Entity Recognition	Official

http://nlpprogress.com/english/named_entity_recognition.html

Estado da Arte

- W-NUT2017

Modelo	F1	Artigo	Código
Flair embeddings (Akbik et al., 2018)	49.59	Pooled Contextualized Embeddings for Named Entity Recognition / Flair framework	-
Aguilar et al. (2018)	45.55	Modeling Noisiness to Recognize Named Entities using Multitask Neural Networks on Social Media	-
SpinningBytes	40.78	Transfer Learning and Sentence Level Features for Named Entity Recognition on Tweets	-
Flair embeddings (Akbik et al., 2018)	49.59	Pooled Contextualized Embeddings for Named Entity Recognition / Flair framework	-

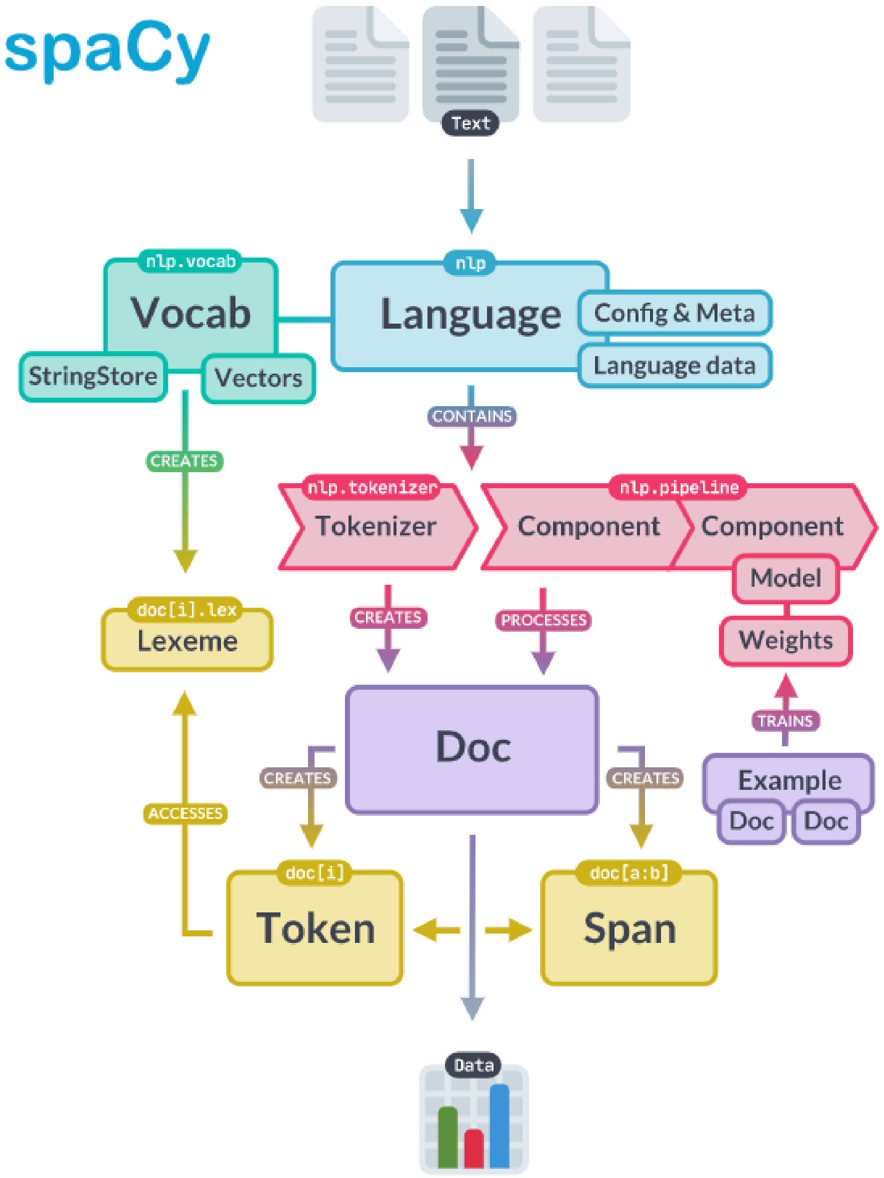
http://nlpprogress.com/english/named_entity_recognition.html

Trabalhos em HD

- *Information Extraction from Historical Handwritten Document Images with a Context-aware Neural Model*. J. Ignacio Toledo, Manuel Carbonell, Alicia Fornés, Josep Lladós. Pattern Recognition, Vol. 86, Feb 2019, Pags. 27-36

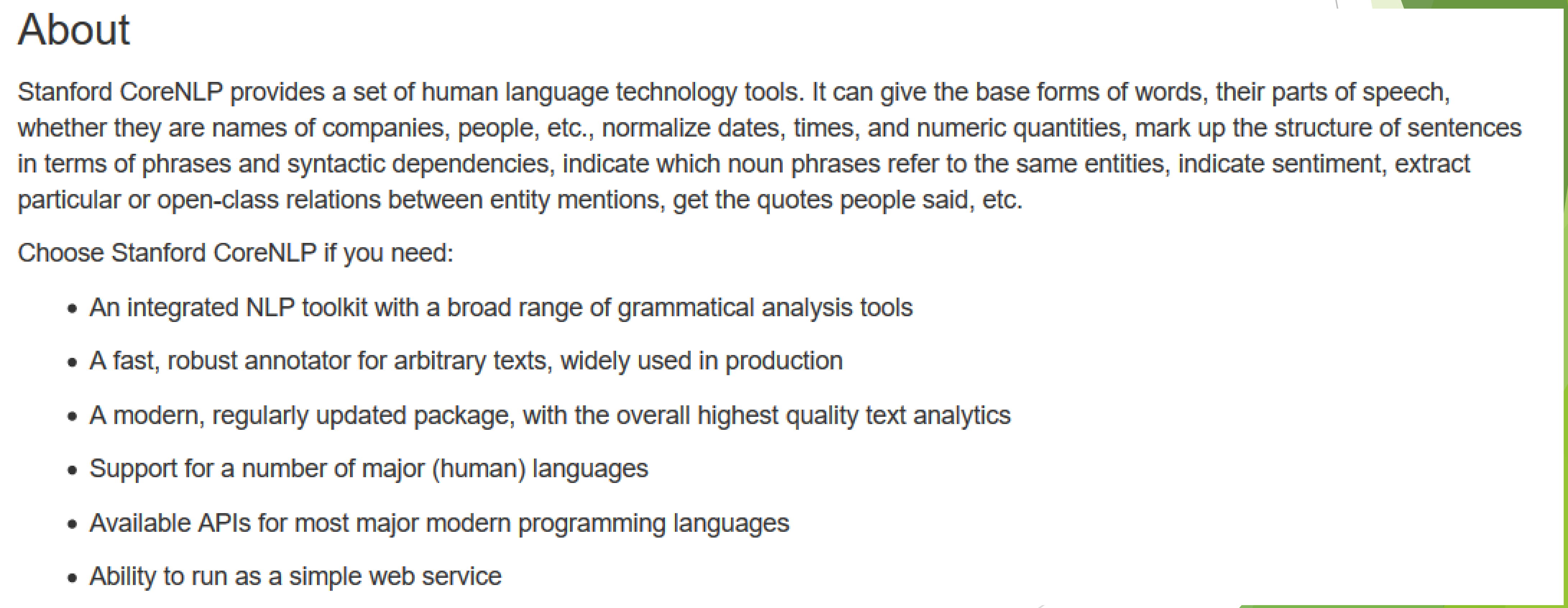
Bibliotecas

- Spacy
- NLTK
- TextBLOB
- UDPIPE



Ferramentas de Anotação

- Stanford CoreNLP



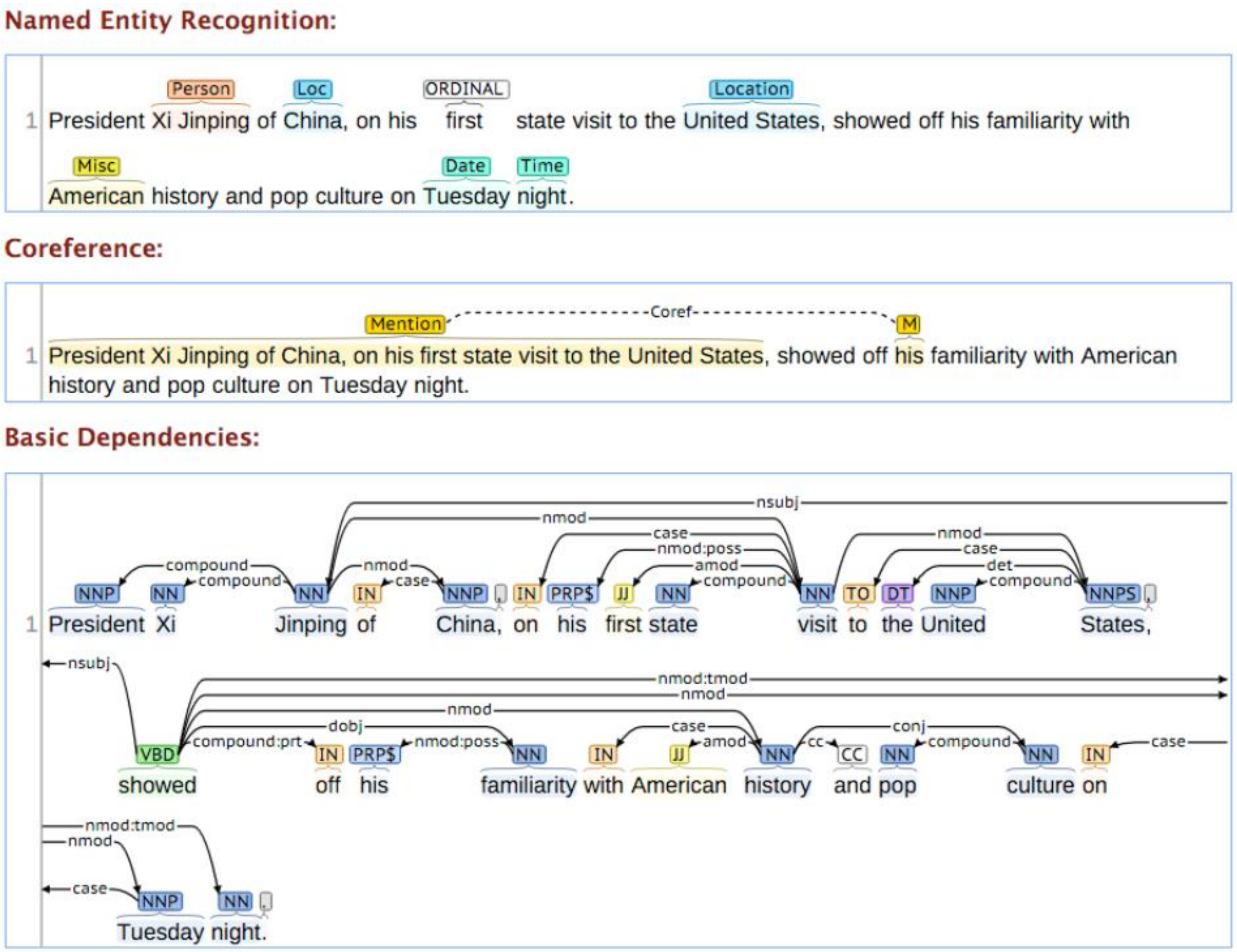
About

Stanford CoreNLP provides a set of human language technology tools. It can give the base forms of words, their parts of speech, whether they are names of companies, people, etc., normalize dates, times, and numeric quantities, mark up the structure of sentences in terms of phrases and syntactic dependencies, indicate which noun phrases refer to the same entities, indicate sentiment, extract particular or open-class relations between entity mentions, get the quotes people said, etc.

Choose Stanford CoreNLP if you need:

- An integrated NLP toolkit with a broad range of grammatical analysis tools
- A fast, robust annotator for arbitrary texts, widely used in production
- A modern, regularly updated package, with the overall highest quality text analytics
- Support for a number of major (human) languages
- Available APIs for most major modern programming languages
- Ability to run as a simple web service

Ferramentas de Anotação



Named Entity Recognition:

1 President Xi Jinping of China, on his first state visit to the United States, showed off his familiarity with American history and pop culture on Tuesday night.

Coreference:

1 President Xi Jinping of China, on his first state visit to the United States, showed off his familiarity with American history and pop culture on Tuesday night.

Basic Dependencies:

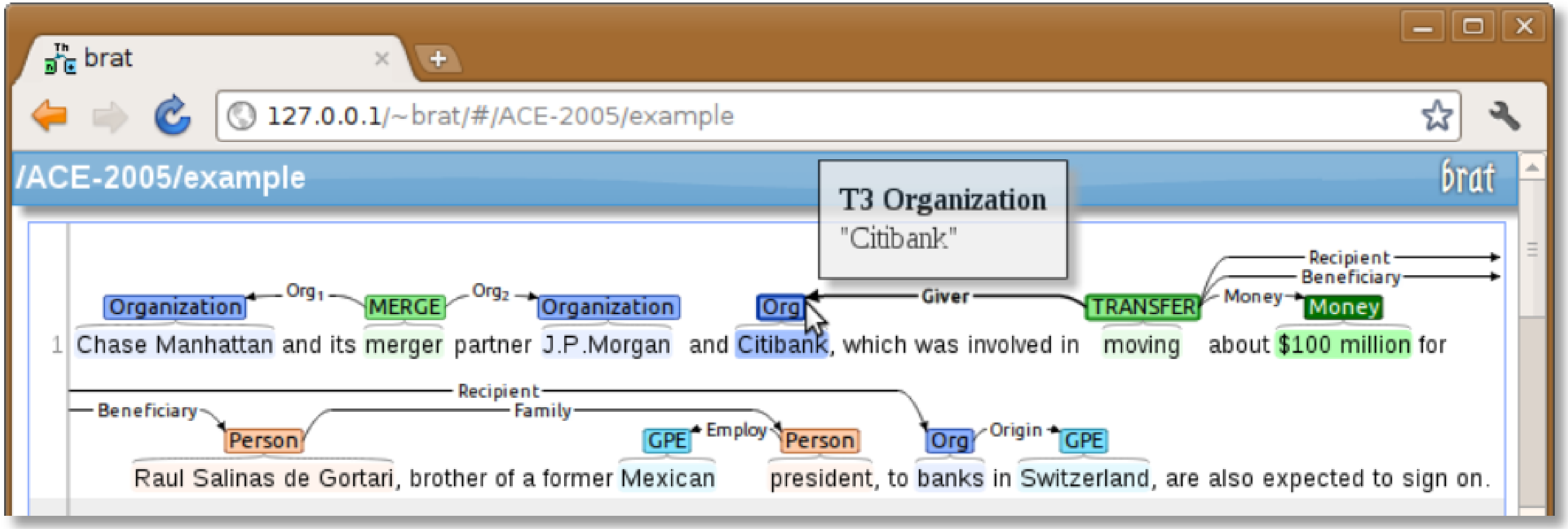
1 President Xi Jinping of China, on his first state visit to the United States, showed off his familiarity with American history and pop culture on Tuesday night.

Ferramentas de Anotação

- Brat

Comprehensive visualization

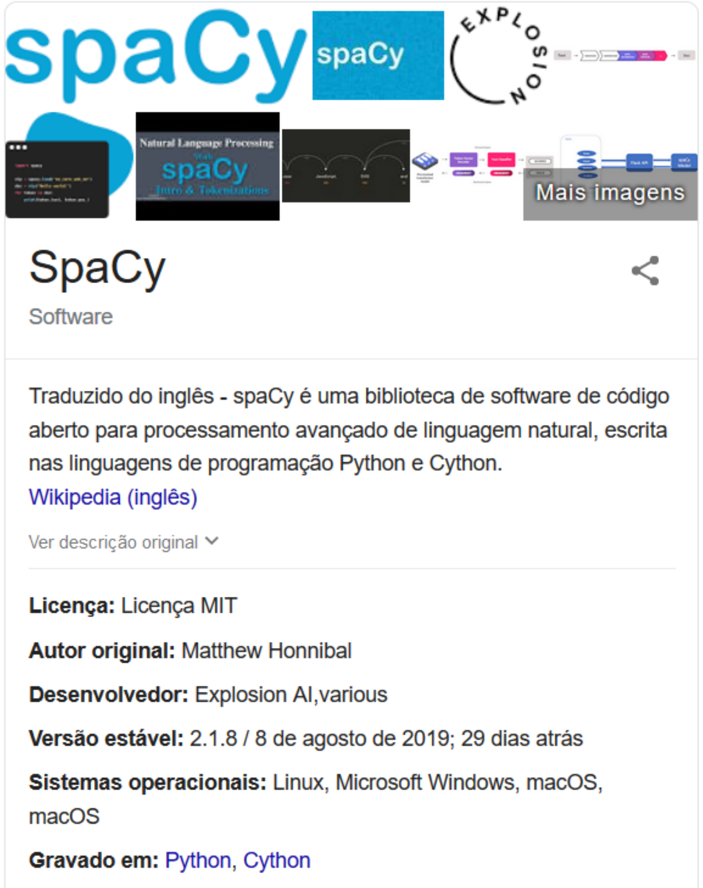
The brat annotation visualization is based on the concept of "what you see is what you get": all aspects of the underlying annotation are visually represented in an intuitive way.


brat
127.0.0.1/~brat/#/ACE-2005/example
/ACE-2005/example
brat
T3 Organization
"Citibank"
Organization
Org1
MERGE
Org2
Organization
Org
Giver
TRANSFER
Recipient
Beneficiary
Money
Money
1
Chase Manhattan and its merger partner J.P.Morgan and Citibank, which was involved in moving about $100 million for
Recipient
Beneficiary
Family
Person
GPE
Employ
Person
Org
Origin
GPE
Raul Salinas de Gortari, brother of a former Mexican president, to banks in Switzerland, are also expected to sign on.

Annotation visualization

Implementação

- Spacy



SpaCy

Software

Traduzido do inglês - spaCy é uma biblioteca de software de código aberto para processamento avançado de linguagem natural, escrita nas linguagens de programação Python e Cython.

Wikipedia (inglês)

Ver descrição original

Licença: Licença MIT

Autor original: Matthew Honnibal

Desenvolvedor: Explosion AI,various

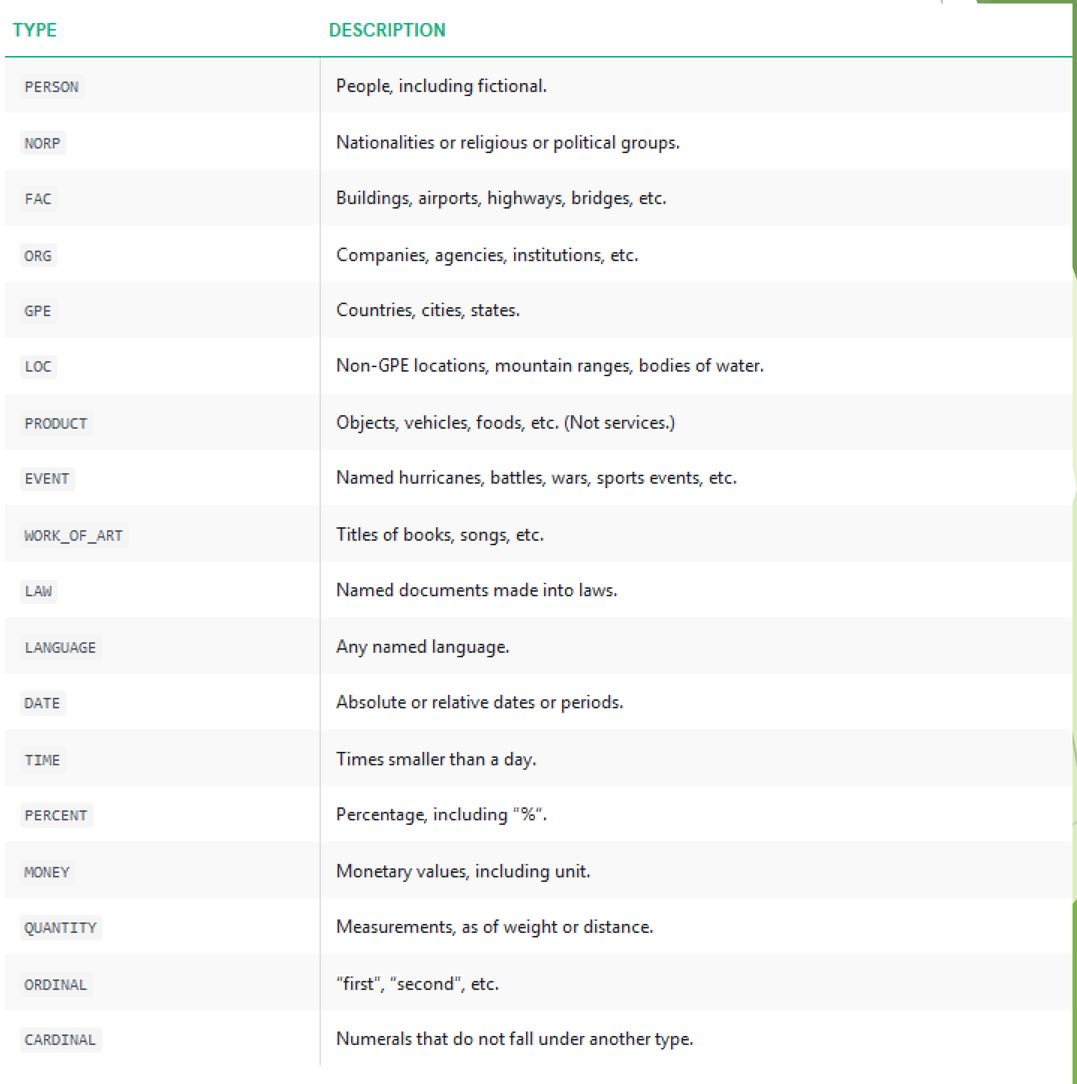
Versão estável: 2.1.8 / 8 de agosto de 2019; 29 dias atrás

Sistemas operacionais: Linux, Microsoft Windows, macOS, macOS

Gravado em: Python, Cython

Implementação

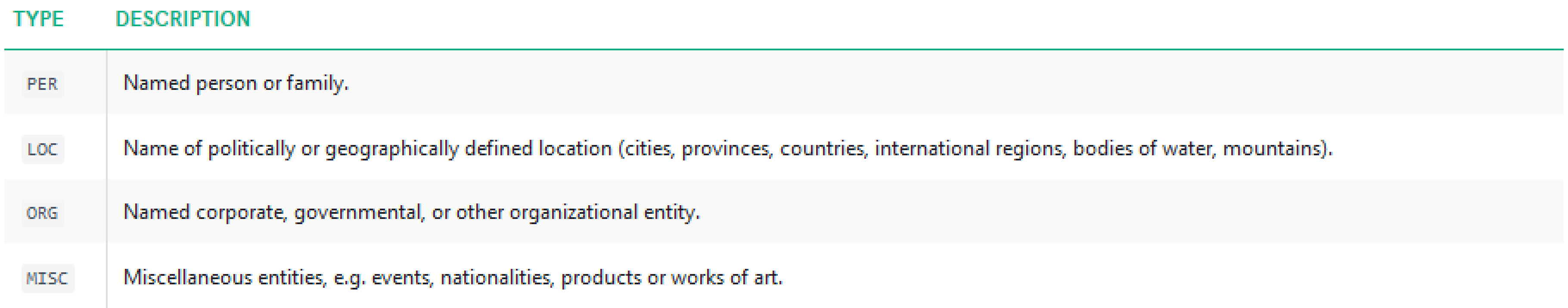
- Modelo treinado no Ontonotes v5



TYPE	DESCRIPTION
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	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.
CARDINAL	Numerals that do not fall under another type.

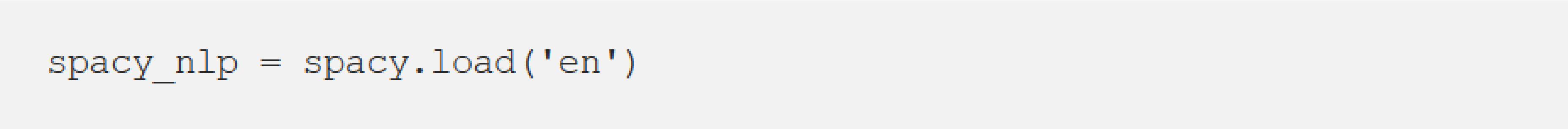
Implementação

- Modelo treinado no Wikipedia Corpus

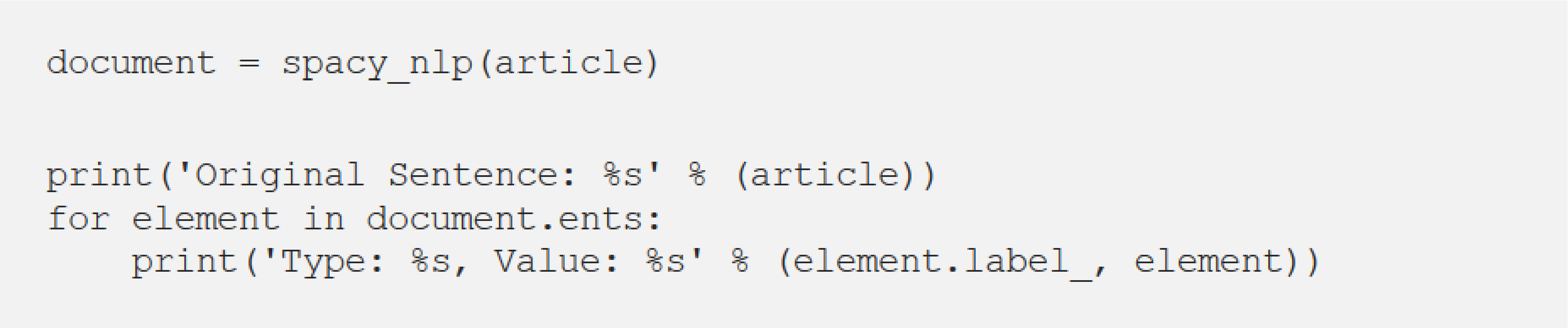


TYPE	DESCRIPTION
PER	Named person or family.
LOC	Name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains).
ORG	Named corporate, governmental, or other organizational entity.
MISC	Miscellaneous entities, e.g. events, nationalities,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

Implementação



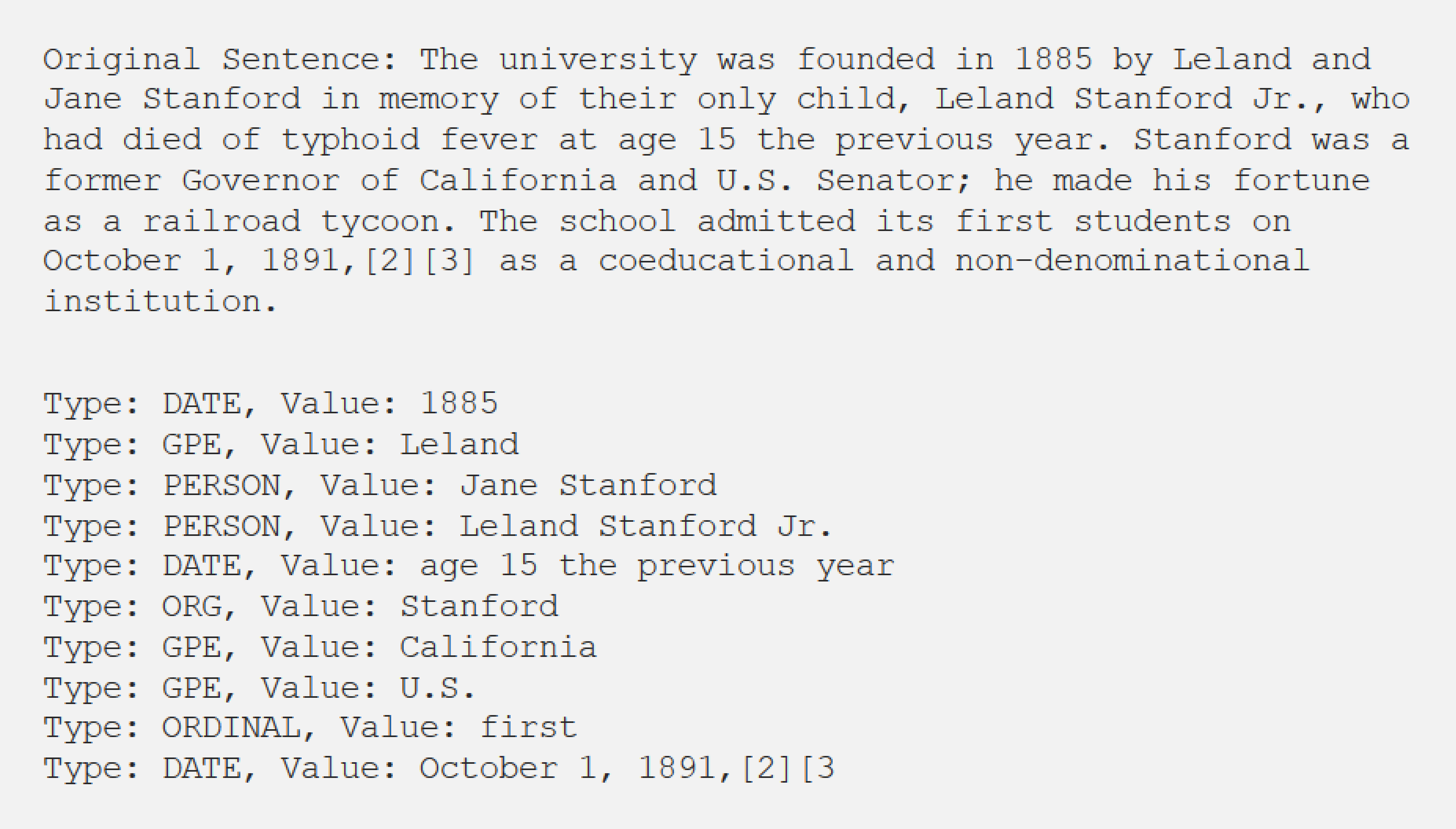
```
spacy_nlp = spacy.load('en')
```



```
document = spacy_nlp(article)


print('Original Sentence: %s' % (article))
for element in document.ents:
    print('Type: %s, Value: %s' % (element.label_, element))
```

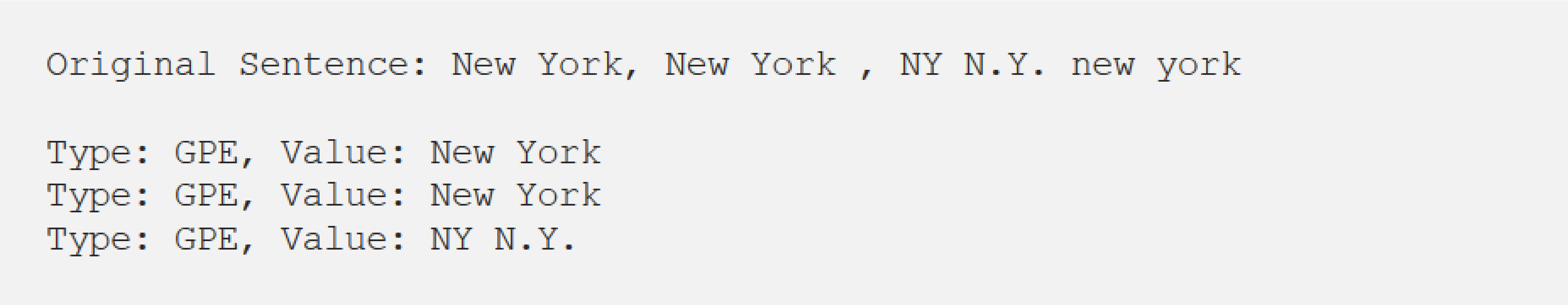
Implementação



```
Original Sentence: The university was founded in 1885 by Leland and
Jane Stanford in memory of their only child, Leland Stanford Jr., who
had died of typhoid fever at age 15 the previous year. Stanford was a
former Governor of California and U.S. Senator; he made his fortune
as a railroad tycoon. The school admitted its first students on
October 1, 1891,[2][3] as a coeducational and non-denominational
institution.


Type: DATE, Value: 1885
Type: GPE, Value: Leland
Type: PERSON, Value: Jane Stanford
Type: PERSON, Value: Leland Stanford Jr.
Type: DATE, Value: age 15 the previous year
Type: ORG, Value: Stanford
Type: GPE, Value: California
Type: GPE, Value: U.S.
Type: ORDINAL, Value: first
Type: DATE, Value: October 1, 1891,[2][3
```

Implementação



```
Original Sentence: New York, New York , NY N.Y. new york

Type: GPE, Value: New York
Type: GPE, Value: New York
Type: GPE, Value: NY N.Y.
```

Exercício

- Faça uma extração de entidades nomeadas a partir de obras literárias
  - Utilize as instruções para carga do conjunto de dados em https://www.nltk.org/book/ch02.html
- Faça visualizações por WordCloud
  - Por Entidade
  - Por Palavras que são entidades
- Gere estatísticas
  - Total de pessoas distintas numa obra
  - Total de lugares numa obra
- Compare duas obras