Introdução ao processamento de linguagem natural com

spaCy

Live de Python #226

Roteiro

1. Processamento de linguagem

Afinal, que raios é isso e pra que serve?

2. SpaCy

Instalação e conceitos básicos da biblioteca

3. Ferramentas de linguagem

Tags, morfologia e lematização

4. Um tiquinho sobre linguística

O básico para sobrevivência

picpay.me/dunossauro

apoia.se/livedepython

pix.dunossauro@gmail.com

Ajude o projeto <3

A2n, Ademar Peixoto, Adilson Herculano, Adriana Cavalcanti, Adriano Ferraz, Alexandre Harano, Alexandre Lima, Alexandre Souza, Alexandre Takahashi, Alexandre Villares, Alex Lima, Allan Almeida, Alynne Ferreira, Alysson Oliveira, Ana Carneiro, Andre Azevedo, André Rafael, Aquiles Coutinho, Arnaldo Turque, Aurelio Costa, Bruno Batista, Bruno Divino, Bruno Freitas, Bruno Guizi, Bruno Lopes, Bruno Ramos, Caio Felix, Caio Nascimento, Carina Pereira, Christiano Morais, Clara Battesini, Dandara Sousa, Daniel Freitas, Daniel Haas, Daniel Santos, Danilo Segura, David Couto, David Kwast, Delton Porfiro, Diego Farias, Diego Guimarães, Dilenon Delfino, Dino Aguilar, Diogo Paschoal, Douglas Bastos, Douglas Zickuhr, Eduardo Tolmasquim, Elton Silva, Emanuel Betcel, Emerson Rafael, Eneas Teles, Erick Ritir, Érico Andrei, Eugenio Mazzini, Euripedes Borges, Everton Silva, Fabiano Tomita, Fabio Barros, Fábio Barros, Fabio Castro, Fábio Thomaz, Fabricio, Fabricio Araujo, Felipe Rodrigues, Fernanda Prado, Fernando Florëncio, Firehouse, Flávio Meira, Flavkaze, Gabriel Barbosa, Gabriel Mizuno, Gabriel Nascimento, Gabriel Simonetto, Geandreson Costa, Guilherme Cabrera, Guilherme Felitti, Guilherme Gall, Guilherme Ostrock, Guilherme Piccioni, Guilherme Silva, Gustavo Suto, Harold Gautschi, Heitor Fernandes, Henrique Junqueira, Hugo Cosme, Igor Taconi, Ismael Ventura, Italo Silva, Jairo Jesus, Jairo Lenfers, Janael Pinheiro, João Paulo, Joelson Sartori, Johnny Tardin, Jônatas Silva, José Barbosa, José Gomes, Joseíto Júnior, Jose Mazolini, José Pedro, Juan Gutierrez, Juliana Machado, Julio Franco, Júlio Gazeta, Júlio Pereira, Julio Silva, Kaio Peixoto, Kaneson Alves, Leandro Miranda, Leonardo Mello, Leonardo Nazareth, Leon Solon, L. Perciliano, Luancomputacao Roger, Luã Vacaro, Lucas Adorno, Lucas Carderelli, Lucas Mello, Lucas Mendes, Lucas Nascimento, Lucas Schneider, Lucas Simon, Lucas Teixeira, Lucas Valino, Luciano Ratamero, Luciano Silva, Luciano Teixeira, Luiz Junior, Luiz Lima, Luiz Paula, Maicon Pantoja, Maiquel Leonel, Marcelino Pinheiro, Márcio Martignoni, Marcio Moises, Marco Mello, Marcos Gomes, Marco Yamada, Maria Clara, Maria Gabriela, Marina Passos, Matheus Cortezi, Matheus Oliveira, Matheus Silva, Matheus Vian, Mauricio Fagundes, Mauricio Nunes, Mírian Batista, Mlevi Lsantos, Murilo Andrade, Murilocunha, Murilo Viana, Nando Sangenetto, Natan Cervinski, Nicolas Teodosio, Osvaldo Neto, Otávio Carneiro, Patricia Minamizawa, Patrick Felipe, Paulo D., Paulo Tadei, Pedro Henrique, Pedro Pereira, Pedro Silva, Peterson Santos, P Muniz, Priscila Santos, Rafael Lopes, Rafael Romão, Ramayana Menezes, Regis Santos, Regis Tomkiel, Rene Bastos, Ricardo Silva, Ricarte Jr, Riverfount, Robson, Robson Maciel, Rodrigo Alves, Rodrigo Cardoso, Rodrigo Freire, Rodrigo Messias, Rodrigo Quiles, Rodrigo Ribeiro, Rodrigo Vaccari, Rodrigo Vieira, Rogério Lima, Rogério Nogueira, Rogério Sousa, Ronaldo Silva, Ronaldo Silveira, Rui Jr, Samanta Cicilia, Sebastião Tolentino, Talita Rossari, Tay Turnner, Thaynara Pinto, Thi, Thiago Araujo, Thiago Borges, Thiago Curvelo, Thiago Moraes, Thiago Salgado, Thiago Souza, Tiago Minuzzi, Tiago Souza, Tony Dias, Tony Santos, Tyrone Damasceno, Valcilon Silva, Valdir Tegon, Vcwild, Vinícius Bastos, Vinicius Stein, Vitor Luz, Vladimir Lemos, Walter Reis, Wesley Mendes, Willian Lopes, Wilson Duarte, Wilson Neto, Wilson Rocha, Xico Silvério, Yury Barros

Obrigado você

Processamento de linguagem natural

PLN

Máquinas vs Linguagem

Máquinas não sabem ler, máquinas não entendem contextos linguísticos.

A área de estudo do processamento de linguagem natural é estudar como fazer com que o computador seja capaz de "entender" as línguas naturais.

Por exemplo: Conseguir entender um documento e extrair informações do mesmo.

Para que seja possível catalogar, categorizar e organizar os próprios documentos.

https://en.wikipedia.org/wiki/Natural_language_processing

A divisão das áreas do PLN

Embora na live de hoje falaremos somente sobre processamento de texto escrito, o PLN pode ser aplicado em diversas outras fonte de dados, como:

- Sons:
  - Reconhecimento de fala [**Vosk**]
  - Conversão de texto em fala [**Coqui**]
- Imagens
  - Reconhecimento óptico (texto em imagens) [**Tesseract**]

exemplo de som: twitter.com/dunossauro/status/1340780846521970690

Processamento de texto

O processamento de texto pode ser usado para diversos seguimentos. Por exemplo:

- Classificar e extrair texto
- Processar e analisar texto
- Chatbots
- Softwares de tradução
- Corretores ortográficos
- etc...

Trabalhando com texto

spaCy

NLTK

OPEN NLP™

Stanza

Instalação e uso básico

spaCy

spaCy

Spacy é uma biblioteca de código aberto para processamento de linguagem natural.

- Licença MIT
- Primeira versão em 2016 (1.0)
- Atualmente na versão 3.4.3
- Suporta 73 idiomas
- Mantido pela explosion.ia e pela comunidade
- Pronto para usar em produção *

https://spacy.io/usage/spacy-101#whats-spacy

pip install spacy

Instalação

O básico necessário

```python
from spacy import blank

nlp = blank('pt')

doc = nlp('Eduardo foi a feira.')

token = doc[0]  # Eduardo

span = doc[:2]  # Eduardo foi
```

O básico necessário

```
1   from spacy import blank
2
3   nlp = blank('pt')
4
5   doc = nlp('Eduardo foi a feira.')
6
7   token = doc[0]  # Eduardo
8
9   span = doc[:2]  # Eduardo foi
```

Importando um modelo em branco e dizendo para usar as regras de português.

O básico necessário

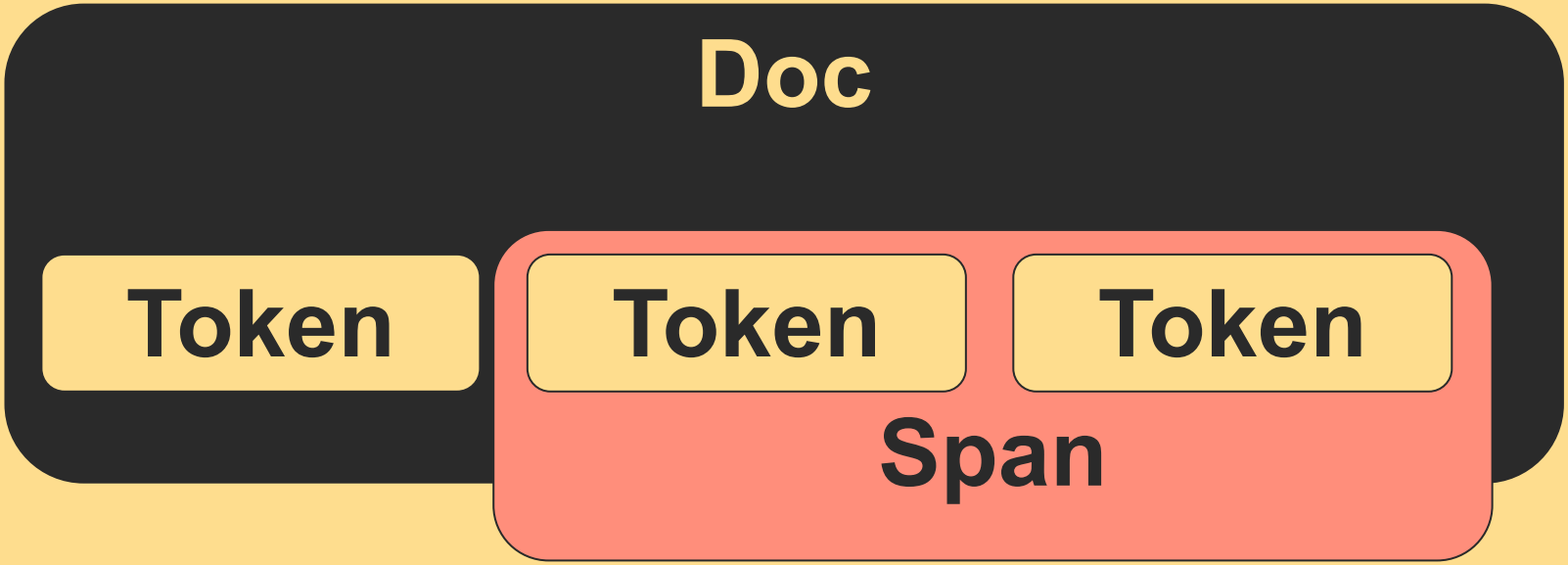
```
1   from spacy import blank
2
3   nlp = blank('pt')
4
5   doc = nlp('Eduardo foi a feira.')
6
7   token = doc[0]  # Eduardo
8
9   span = doc[:2]  # Eduardo foi
```

Doc: É um documento de texto completo

Token: Uma unidade da língua

Span: Uma fatia de um Doc

Containers básicos



https://course.spacy.io/pt/chapter1

Carregando um modelo pré-treinado

```
1   from spacy import load
2
3   texto = 'Eduardo, fazerdor de vídeos no Youtube.'
4
5   nlp = load('pt_core_news_lg')
6   doc = nlp(texto)
```

Carregando um modelo pré-treinado

```
1   from spacy import load
2
3   texto = 'Eduardo, fazerdor                    e.'
4
5   nlp = load('pt_core_news_lg')
6   doc = nlp(texto)
```

O modelo precisa ser baixado ANTES

python -m spacy download <modelo>

https://spacy.io/usage/models

Baixar um modelo para a língua

Diferentes modelos

Existem três variações do mesmo modelo para português:

- pt_core_news_**sm**: Small (pequeno)
- pt_core_news_**md**: Medium (médio)
- pt_core_news_**lg**: Large (Grande)

Quanto menor o modelo, mais rápido ele será. Quanto maior, maior o nível de acurácia do modelo.

python -m spacy download pt_core_news_lg

https://spacy.io/usage

Baixar um modelo para a língua

Experimentando o modelo

```python
from spacy import load

nlp = load('pt_core_news_lg')

doc = nlp(
  'Eduardo foi a feira. Comprou dois pasteis.'
)

doc.sents
# [Eduardo foi a feira., Comprou dois pasteis.]
```

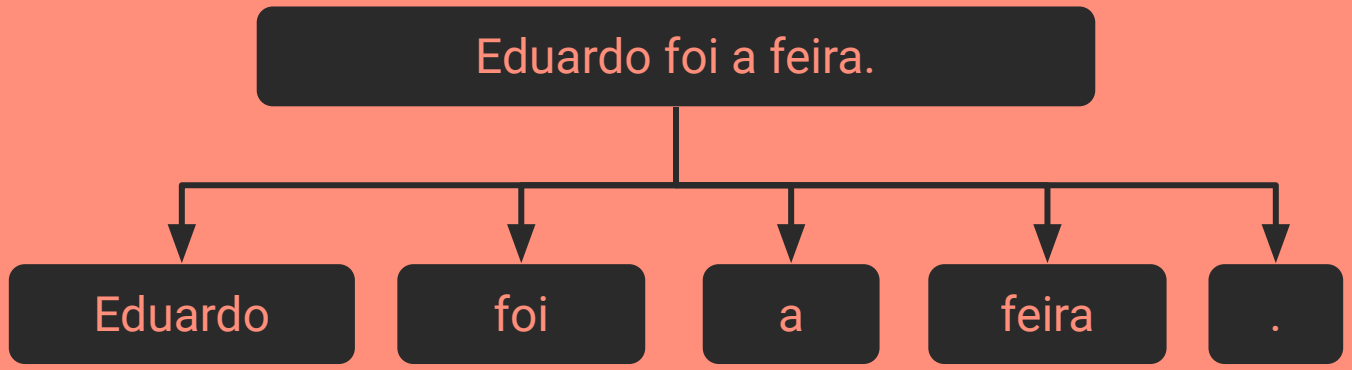
Trabalhando com línguas

Ferramentas

Tokenização doc = 'Eduardo foi a feira.'

A base de trabalhar com documentos de texto é a tokenização. Tokenizar é a arte de separar unidades de linguagem.

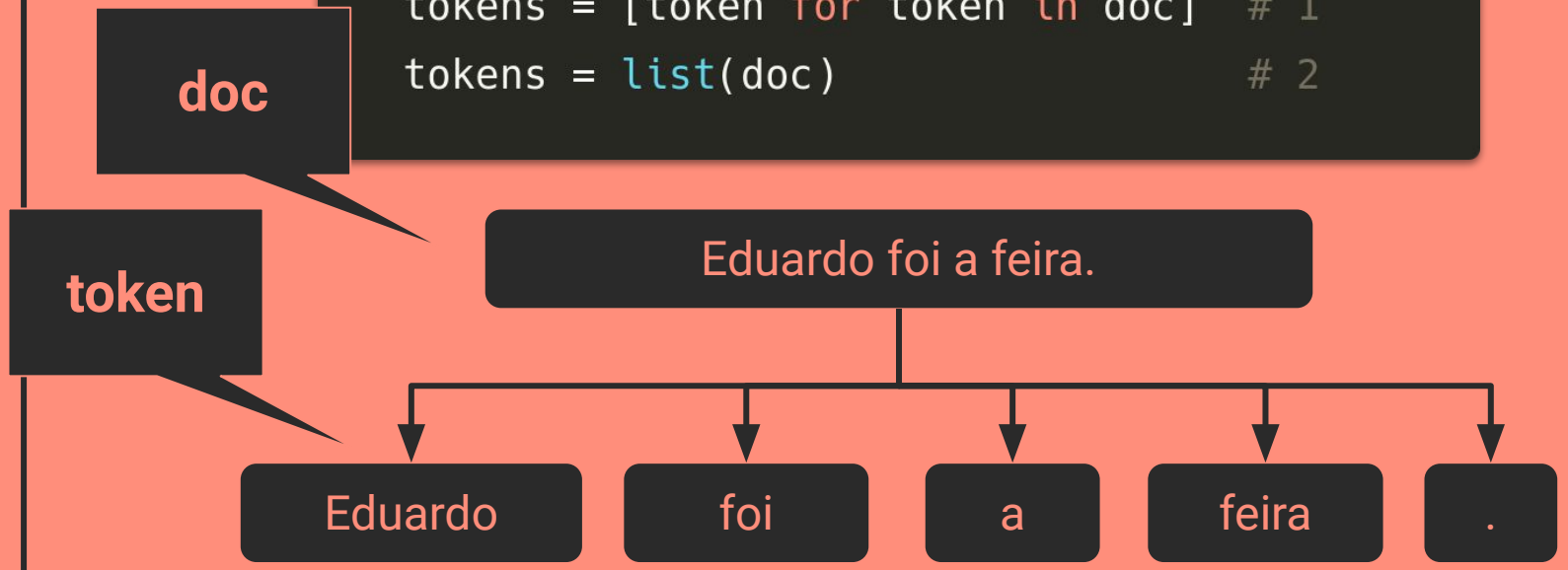
```
tokens = [token for token in doc]  # 1
tokens = list(doc)                 # 2
```



Tokenização

A base de trabalhar com documentos de texto é a tokenização. Tokenizar é a arte de separar palavras.

```
tokens = [token for token in doc]  # 1
tokens = list(doc)                 # 2
```

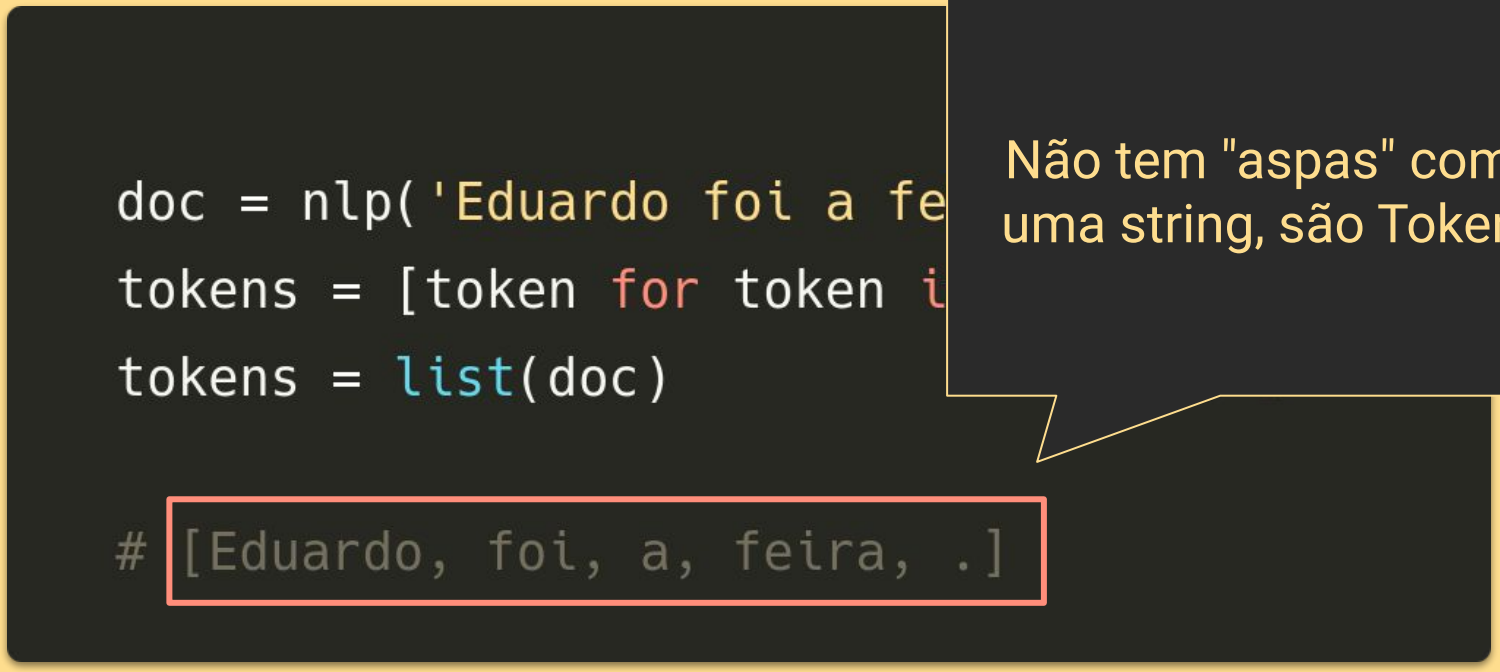


Tokens

```
doc = nlp('Eduardo foi a feira.')
tokens = [token for token in doc]  # 1
tokens = list(doc)                 # 2

# [Eduardo, foi, a, feira, .]
```

Tokens


doc = nlp('Eduardo foi a fe
tokens = [token for token i
tokens = list(doc)
# [Eduardo, foi, a, feira, .]
Não tem "aspas" como uma string, são Tokens

Atributos dos Tokens {exemplo_01.py}

Os tokens são estruturas de dados extremamente poderosas. Os modelos podem deixar diversos atributos importantes neles.

```
1  from spacy import load
2  nlp = load('pt_core_news_lg')
3  doc = nlp('Eduardo faz vídeos.')
4
5  for token in doc:
6    print(token.text, token.shape_, token.is_alpha)
7    # Resultados
8    'Eduardo', 'Xxxxxxx', True
9    'faz', 'xxx', True
10   'vídeos', 'xxxxxx', True
11   '.', '.', False
```

Tipos de análise

Existem dois tipos de análise de textos:

- Morfológica: Estudo das palavras e suas classes gramaticais
- Sintática: Função em que as palavras desempenham em orações

Sintática

Análisando contextos

Atributos léxicos

Com Spacy podemos obter diversas características linguísticas dos tokens. Como:

- Tagueamento POS (Part Of Speech)
- Lematização
- Dependências
- StopWods

Atributos léxicos

Com Spacy podemos obter diversas características linguísticas dos tokens. Como:

- **Tagueamento POS (Part Of Speech)**
  - Classe gramatical
    - Eduardo: Substantivo
    - Foi: Verbo
- **Lematização**
  - Desflexionar palavras
    - Gatos: Gato
    - Gatas: Gato
    - Gata: Gato

Atributos léxicos

Com Spacy podemos obter diversas características linguísticas dos tokens. Como:

- **Dependências**
  - Relações entre palavras
  - "Maria tem"
    - Maria <- Sujeito nominal <- tem
    - Alguém têm, no caso Maria
- **StopWods**
  - Palavras irrelevantes para o contexto
    - "As pessoas foram"
      - "As" é irrelevante para o contexto.

Isso no código {exemplo_02.py}

```
1   doc = nlp('Eduardo faz vídeos.')
2   for token in doc:
3     print(token.text, token.pos_, token.lemma_, token.dep_, token.is_stop)
4
5     # resultado
6     Token    | POS      | Lemma    | Dep      | Stop
7     -----    | -----    | -----    | -----    | -----
8     Eduardo  | PROPN    | Eduardo  | nsubj    | False
9     faz      | VERB     | fazer    | ROOT     | True
10    vídeos   | NOUN     | vídeo    | obj      | False
11    .        | PUNCT    | .        | punct    | False
```

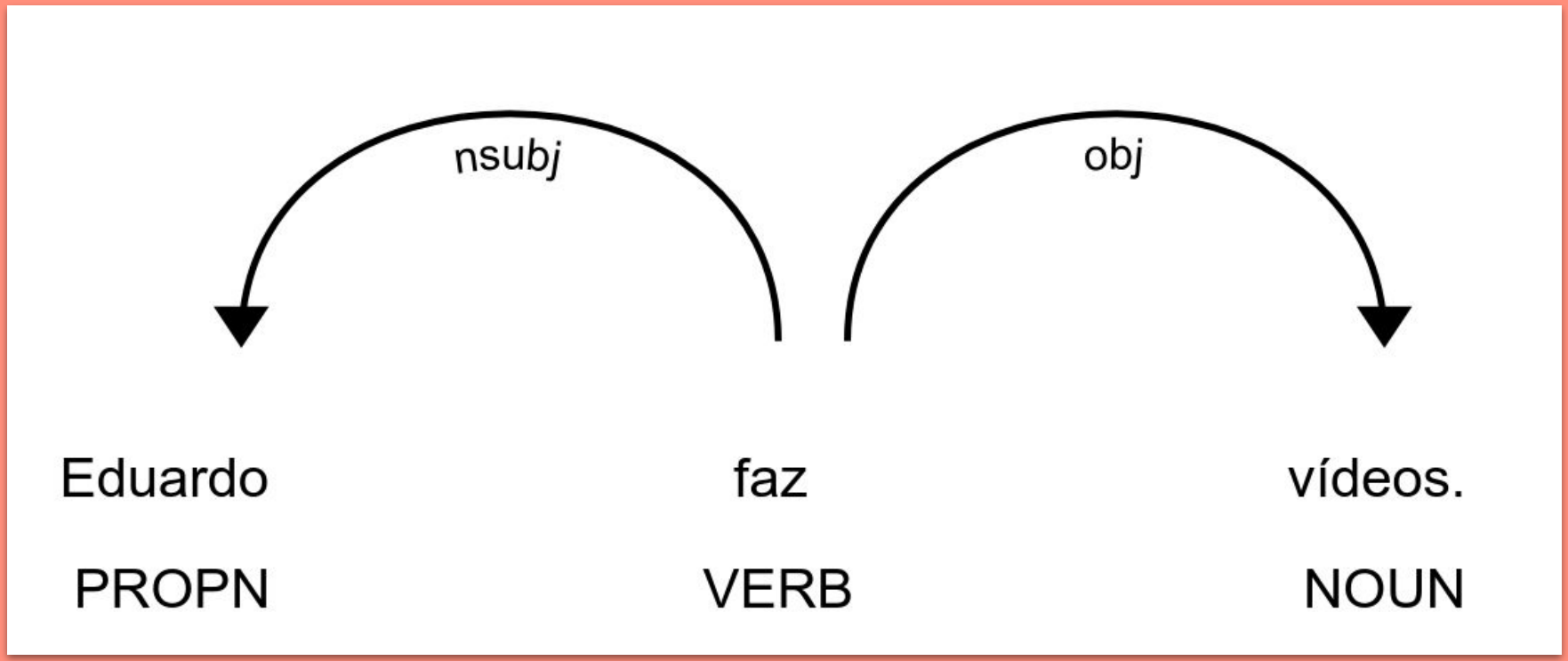
Isso no código {exemplo_02.py}

Token	POS	Lemma	Dep	Stop
Eduardo	PROPN	Eduardo	nsubj	False
faz	VERB	fazer	ROOT	True
vídeos	NOUN	vídeo	obj	False
.	PUNCT	.	punct	False

Visualizar é MUITO mais simples

O spaCy conta com um método de visualização da sintaxe

```
1   from spacy import load, displacy
2
3   nlp = load('pt_core_news_lg')
4   doc = nlp('Eduardo faz vídeos.')
5
6   displacy.serve(doc)
```



http://localhost:5000/

Mas que raios essas coisas querem dizer?

```python
from spacy import explain

# Tags
explain('PROPN')  # 'proper noun'
explain('VERB')   # 'verb'
explain('NOUN')   # 'noun'

# Deps
explain('nsubj')  # 'nominal subject'
explain('obj')    # 'object'
```

POS Tagging (marcação de parte da fala)

Tagging é o nome dado para análise e marcação de determinados tokens.

A marcação é importante para saber

```
for token in nlp('Ele bota a calça.'):
    print(f'{token.text :10} | {token.tag_}')

'''
Ele        | PRON
bota       | VERB
a          | DET
calça      | NOUN
.          | PUNCT
'''
```

As Tags do Universal POS tags

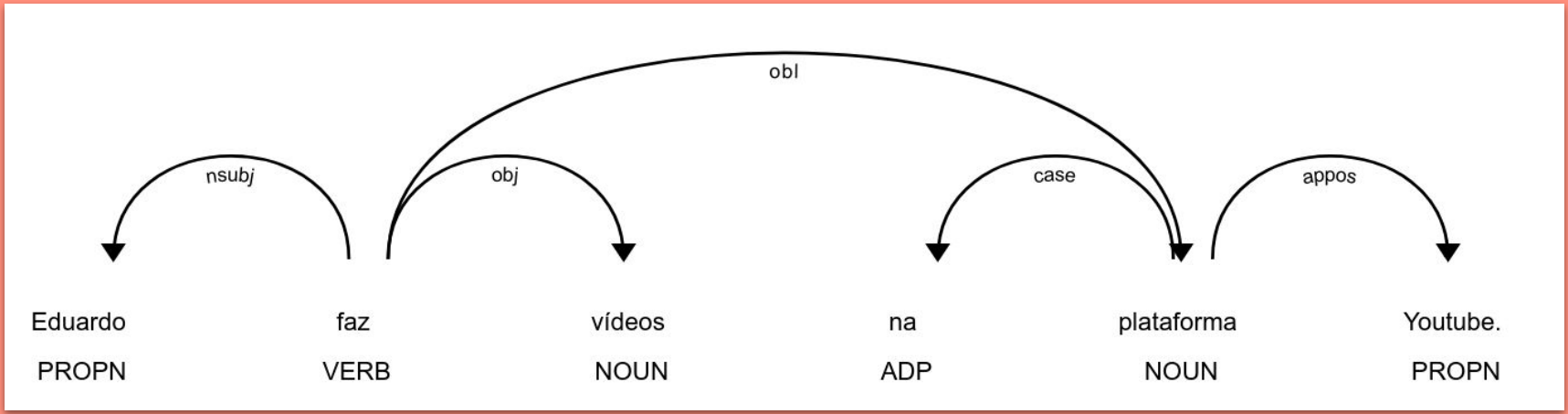
As tags da frase passada

TAG	SIGNIFICADO	EXEMPLOS
PROPN	Pronome	eu, tu, ela, ele, vós, você, …
VERB	Verbo	sair, entrar, ficar, …
DET	Determinante	a, meu, o, muitas, poucas, …
NOUN	Substantivo	zebra, amor, chuva, garfo, …
PUNCT	Pontuação	, . ? ! ;

lista de todas as tags: https://universaldependencies.org/u/pos/

As dependências

As dependências são as relações entre as palavras. Para que uma determinada palavra faça sentido, ela depende de outra.



Lista de todas as dependências: https://universaldependencies.org/u/dep/

As Tags da UDR [Universal Dependecy Relations]

As dependências são as relações entre as palavras. Para que uma determinada palavra faça sentido, ela depende de outra.

TAG	SIGNIFICADO	EXPLICAÇÃO
nsubj	sujeito nominal	Protagonista de algo
obj	objeto / substantivo	Entidade a qual o verbo é destinado
obl	nominal oblíqua	Dependente nominal do verbo
case	marcação de caso	Anexam ou introduzem um objeto
appos	modificador aposicional	atribui um "nome" ao substantivo

lista de todas as tags: https://universaldependencies.org/u/dep/

Morfologia

Mais coisas escondidas em tokens

As tags de morfologia

```
from spacy import load
nlp = load('pt_core_news_lg')
doc = nlp('Eduardo faz vídeos na plataforma Youtube.')

for t in doc:
    print(f'{t.text :10} {t.pos_ :10}', t.morph.to_dict())

"""
Eduardo  PROPN   {'Gender': 'Masc', 'Number': 'Sing'}
faz      VERB    {'Mood': 'Ind', 'Number': 'Sing', 'Person': '3', 'Tense': 'Pres', 'VerbForm': 'Fin'}
vídeos   NOUN    {'Gender': 'Masc', 'Number': 'Plur'}
na       ADP     {'Definite': 'Def', 'Gender': 'Fem', 'Number': 'Sing', 'PronType': 'Art'}
plataforma NOUN    {'Gender': 'Fem', 'Number': 'Sing'}
Youtube  PROPN   {'Gender': 'Fem', 'Number': 'Sing'}
.        PUNCT   {}
"""
```

lista de todas as tags: https://universaldependencies.org/u/feat/

Entidades

Como encontrar elas?

Entidades

Uma das funcionalidades mais legais do spaCy, na minha opinião é a vinculação de entidades.

```
1  from spacy import load
2  nlp = load('pt_core_news_lg')
3  doc = nlp('Eduardo trabalha no Google do Rio de Janeiro.')
4
5  for ent in doc.ents:
6      print(f'{ent.text:20} | {ent.label_}')
7
8  '''
9  Eduardo              | PER
10 Google               | ORG
11 Rio de Janeiro       | LOC
12 '''
```

Forma visual

João PER amava Teresa PER que amava Raimundo PER que amava Maria PER que amava Joaquim PER que amava Lili que não amava ninguém.

João PER foi pra os Estados Unidos LOC , Teresa PER para o convento, Raimundo PER morreu de desastre, Maria PER ficou para tia, Joaquim PER suicidou-se e Lili PER casou com J. Pinto Fernandes PER que não tinha entrado na história.

picpay.me/dunossauro

apoia.se/livedepython

pix.dunossauro@gmail.com

Ajude o projeto <3

Links importantes

- Documentação do spaCy: https://spacy.io
- Lista de modelos: https://spacy.io/usage/models
- Containers: https://spacy.io/api/#architecture-containers
- Para as marcações:
  - Dependências: https://universaldependencies.org/u/dep/
  - POS: https://universaldependencies.org/u/pos/
  - Features: https://universaldependencies.org/u/feat/
  - Conjugação verbal: https://conjuga-me.net/

Outros links:

- Curso da Ines Montani (Fundadora da Explosion): https://course.spacy.io/pt