



# Processamento de linguagem natural com Python

Coordenador  
Fernando Ferreira





# Institucional

---

## Reitor

**Prof. Eduardo Ramos**

[Eduardo.ramos@lnfnet.edu.br](mailto:Eduardo.ramos@lnfnet.edu.br)

## Coordenação do Curso

**Prof. Fernando Ferreira**

[Fernando.gFerreira@lnfnet.edu.br](mailto:Fernando.gFerreira@lnfnet.edu.br)

## Gerente Acadêmica

**Ana Curi**

[ana.cur@lnfnet.edu.br](mailto:ana.cur@lnfnet.edu.br)

# Coordenador/Professor



Prof. Fernando Ferreira

[fernando.gferreira@infnet.edu.br](mailto:fernando.gferreira@infnet.edu.br)

<https://www.linkedin.com/in/nandoferreira/>

## Mini-bio

- Doutor pelo Programa de Engenharia Elétrica da Universidade Federal do Rio de Janeiro (Coppe-UFRJ) - ênfase em Inteligência Computacional.
- Sócio-fundador da empresa Twist (<https://twist.systems/>) - residente do Parque Tecnológico da UFRJ.
- Coordenador do MBA de Data Science e do MIT de Inteligência Artificial do Instituto Infnet.
- Professor de disciplinas de desenvolvimento ágil de software na graduação da Infnet. Professor do MBA de Business Analytics e Big Data da Fundação Getúlio Vargas.
- Membro da colaboração internacional CERN/UFRJ.
- [Twist] Empresa membro-nato do CDIA.Rio - Rede de Ciência de Dados & Inteligência Artificial do Rio de Janeiro

# NER: Named Entity Recognition

---



# Relembrando



**DW**

**TOP STORIES** **MEDIA CENTER** **PROGRAM** **LE**

GERMANY WORLD BUSINESS SCI-TECH ENVIRONMENT CUL

TOP STORIES / BUSINESS

TECHNOLOGY

## Google buys stake in SpaceX to be Internet from outer space

Google has secured a 10 percent stake in the private space exploration company SpaceX along with another investor. Reports suggest the money is aimed at hooking up far-off areas with Internet access from outer space.

The most remote, barren corners of the globe may soon have Internet access.

SpaceX, the private space exploration company backed by Tesla Motors Inc. and Fidelity Investments, Tuesday it had raised \$1 billion (860 million euros) from two new investors, Elon Musk and Google.

**SpaceX**

Space Exploration Technologies Corp logo

Type Private

Industry Aerospace

Founded 2002

Founder Elon Musk

Headquarters Hawthorne, California, USA

Key people Elon Musk (CEO and Chief Designer)  
Gwynne Shotwell (President and COO)  
Tom Mueller (VP of Propulsion)

Services Orbital rocket launch

Number of employees 3,000+ (Jan 2015)[citation needed]

Website SpaceX.com

Footnotes / references [1][2][3]

**Google Inc.**

Google

Type Public

Traded as Class A: NASDAQ: GOOGL  
Class B supervoting: unlisted  
Class C nonvoting:  
NASDAQ: GOOG  
NASDAQ-100 Components (GOOGL and GOOG)  
S&P 500 Components (GOOGL and GOOG)

Industry Internet  
Computer software  
Telecoms equipment

Founded September 4, 1995; 16 years ago

Founder Larry Page, Sergey Brin

Headquarters Googleplex, Mountain View, California, U.S.<sup>[4]</sup>

Area served Worldwide

Key people Larry Page (CEO)  
Eric Schmidt (Chairman)  
Sergey Brin (Director of Google X and Special Projects)<sup>[4]</sup>

Shareholder Musk Shareholder Musk

Send us your feedback.

Print Print this page

Permalink <http://dw.de/p/1ENVv>

## Try the API

Google, headquartered in Mountain View (1600 Amphitheatre Pkwy, Mountain View, CA 940430), unveiled the new Android phone for \$799 at the Consumer Electronic Show. Sundar Pichai said in his keynote that users love their new Android phones.

RESET

See supported languages

Entities

Sentiment

Syntax

Categories

⟨Google⟩<sub>1</sub>, headquartered in ⟨Mountain View⟩<sub>2</sub> ⟨Mountain View (1600 Amphitheatre Pkwy)⟩<sub>12</sub> ⟨1600⟩<sub>14</sub> ⟨Amphitheatre Pkwy⟩<sub>7</sub>, ⟨Mountain View⟩<sub>2</sub>, ⟨CA 940430⟩<sub>8</sub> ⟨940430⟩<sub>16</sub>, unveiled the new ⟨Android⟩<sub>3</sub> ⟨phone⟩<sub>5</sub> for ⟨\$799⟩<sub>13</sub> ⟨799⟩<sub>15</sub> at the ⟨Consumer Electronic Show⟩<sub>11</sub>. ⟨Sundar Pichai⟩<sub>4</sub> said in his ⟨keynote⟩<sub>9</sub> that ⟨users⟩<sub>6</sub> love their new ⟨Android⟩<sub>3</sub> ⟨phones⟩<sub>10</sub>.

1. Google	ORGANIZATION	2. Mountain View	LOCATION
<a href="#">Wikipedia Article</a> Salience: 0.19		<a href="#">Wikipedia Article</a> Salience: 0.18	
3. Android	CONSUMER GOOD	4. Sundar Pichai	PERSON
<a href="#">Wikipedia Article</a> Salience: 0.14		<a href="#">Wikipedia Article</a> Salience: 0.11	
5. phone	CONSUMER GOOD	6. users	PERSON
Salience: 0.10		Salience: 0.09	
7. Amphitheatre Pkwy	LOCATION	8. CA 940430	OTHER
Salience: 0.07		Salience: 0.05	
9. keynote	OTHER	10. phones	CONSUMER GOOD
Salience: 0.03		Salience: 0.02	
11. Consumer Electro...	EVENT	12. Mountain View (16...	ADDRESS
<a href="#">Wikipedia Article</a> - ..			

## Try the API

Hoje, no jogo do Flamengo, tem gol do Gabigol.

RESET

[See supported languages](#)

Entities

Sentiment

Syntax

Categories

Hoje, no **jogo**<sub>2</sub> do **Flamengo**<sub>4</sub>, tem **gol**<sub>1</sub> do **Gabigol**<sub>3</sub>.

1. **gol**

EVENT

Salience: 0.29

2. **jogo**

EVENT

Salience: 0.29

3. **Gabigol**

ORGANIZATION

[Wikipedia Article](#)

Salience: 0.23

4. **Flamengo**

ORGANIZATION

[Wikipedia Article](#)

Salience: 0.19

But Google ORG is starting from behind. The company made a late push into hardware, and Apple ORG's Siri PRODUCT, available on iPhones PRODUCT, and Amazon ORG's Alexa PRODUCT software, which runs on its Echo PRODUCT and Dot PRODUCT devices, have clear leads in consumer adoption.

Achar menções a entidades em um texto e rotular com uma categoria.

Exemplo:

**Trump** attacks **BMW** and **Mercedes**  
**U.N.** official **Ekeus** heads for **Baghdad**



# Rótulos

---

- Pessoas PER
- Locais LOC
- Organizações ORG
- Nomes NAM
- Outros MISC

# Dados Rotulados

**Trump** <PER> attacks **BMW** <ORG> and **Mercedes** <ORG>  
**U.N.** <ORG> official **Ekeus** <PER> heads for **Baghdad** <LOC>

Surface	POS	Sh-synt	Tag
U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

# supervisionada



Dado o segmento, treinar o classificador para indicar:

- Este segmento é uma **Entidade Nomeada?**
- Dar o **rótulo** correto

## Classificação

**Tarefa:** **Trump** attacks BMW and

Mercedes

**P:** **Trump** é uma entidade nomeada?

**R: SIM,** Trump é uma **PESSOA**



MIT Technology Review

Artificial intelligence / Machine learning



# OpenAI's new language generator GPT-3 is shockingly good—and completely mindless

≡ WIRED

SUBSCRIBE

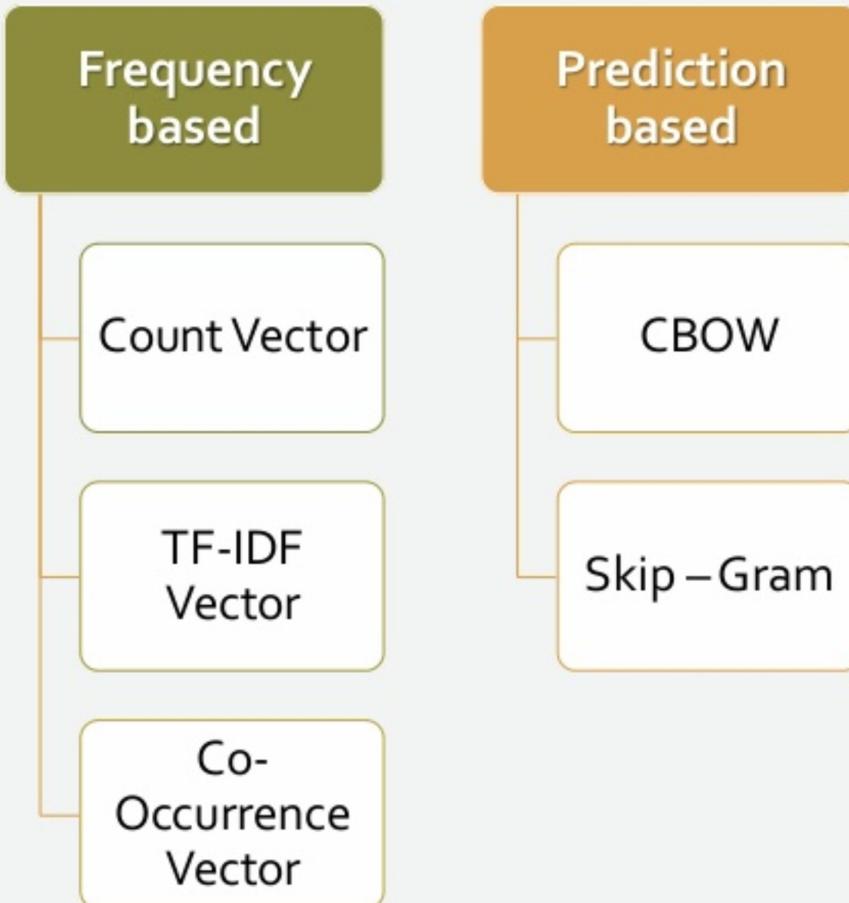
## Did a Person Write This Headline, or a Machine?

GPT-3, a new text-generating program from OpenAI, shows how far the field has come—and how far it has to go.

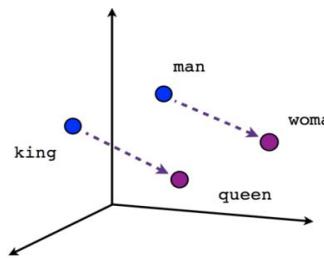
# Predicted Word Embedding

---

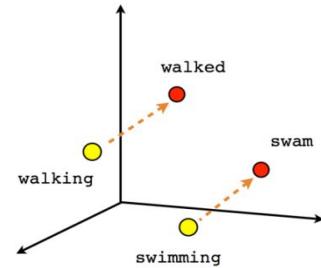




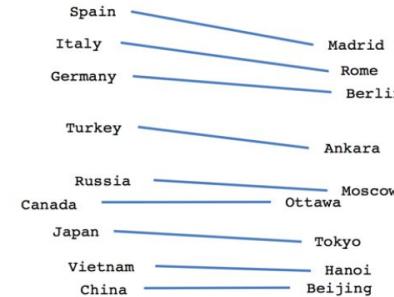
# Word2Vec



Male-Female

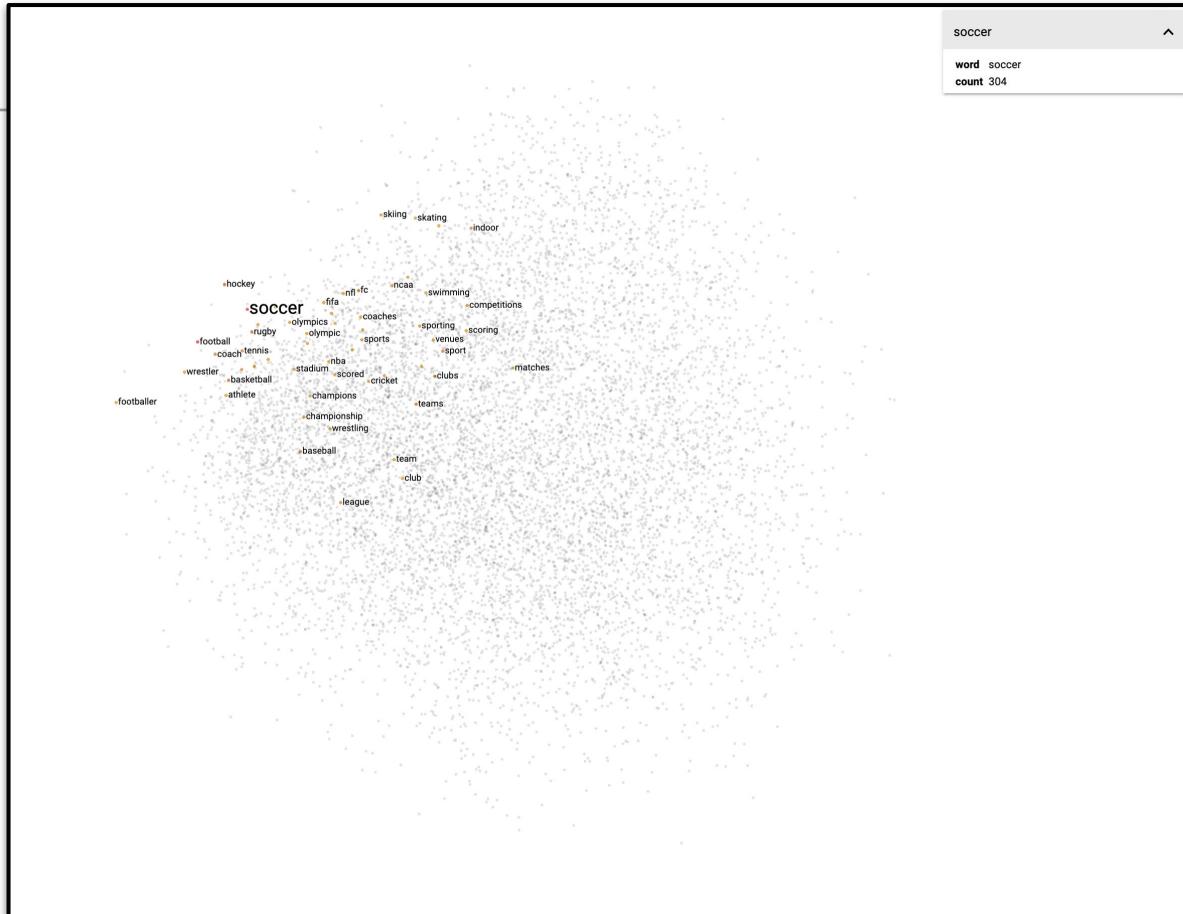


Verb tense

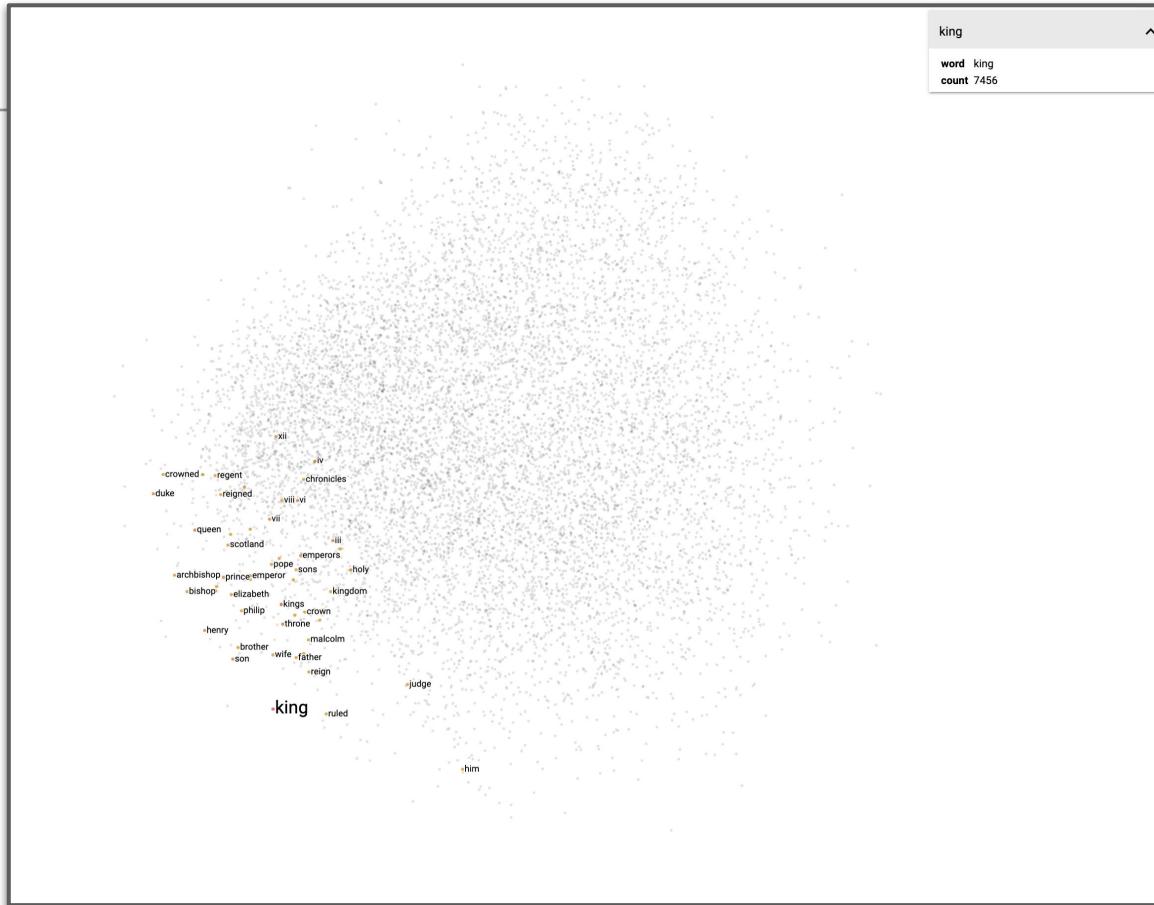


Country-Capital

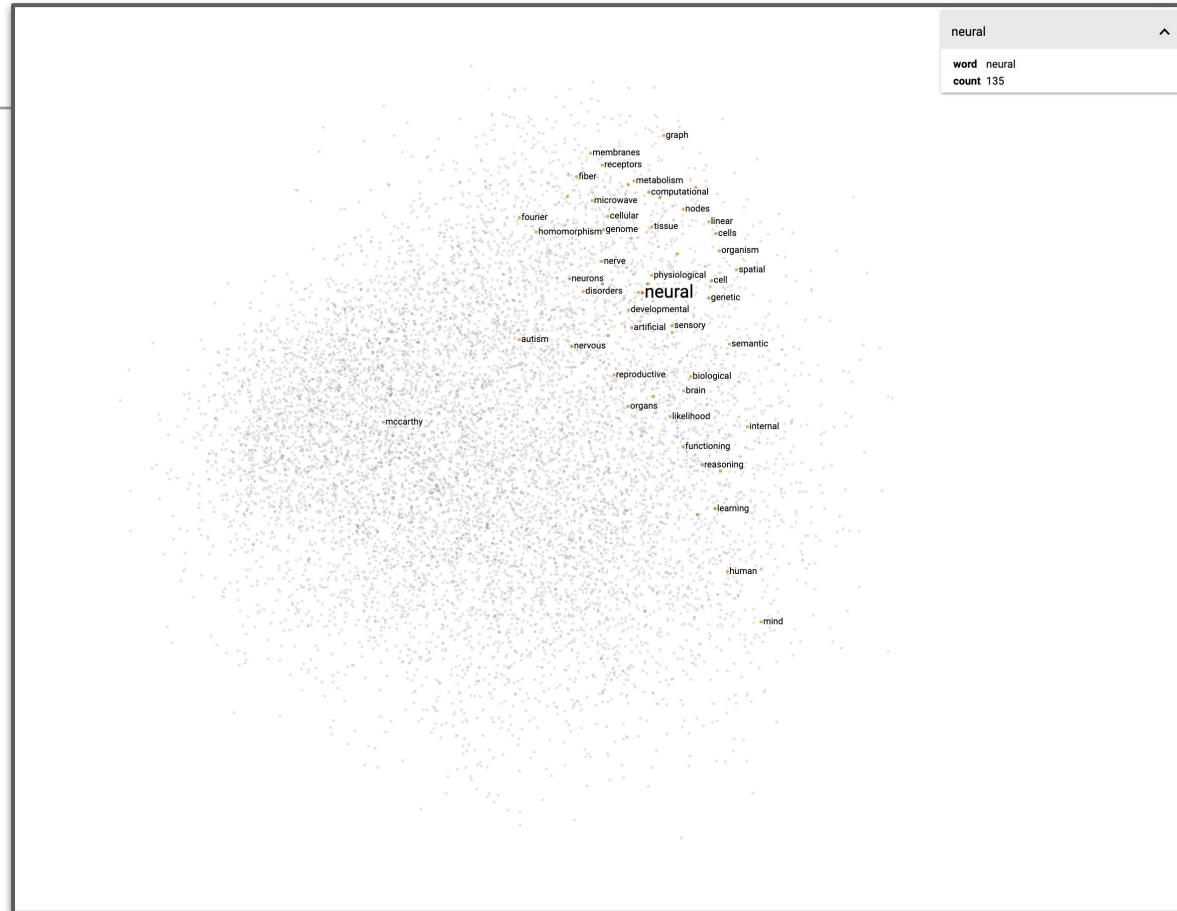
Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De



<https://projector.tensorflow.org/>



<https://projector.tensorflow.org/>



## Doc2vec

- documents to vector space

## tweet2vec

- There are a lot of noisy text and informal language structure.

## item2vec

- dealing with item and user similarity is at heart of lot of recommendation algorithms

## Lda2vec

- this embedding technique tries to marry best of both worlds, word2vec and LDA

# spacy

- 
- Tokenization
  - Part-of-speech (POS) tagging
  - Dependency Parsing
  - Sentence Boundary Detection (SBD)
  - Named Entity Recognition (NER)
  - Similarity
  - Rule-based matching
  - Training



# spaCy

```
$ pip install spacy  
$ python -m spacy download en
```

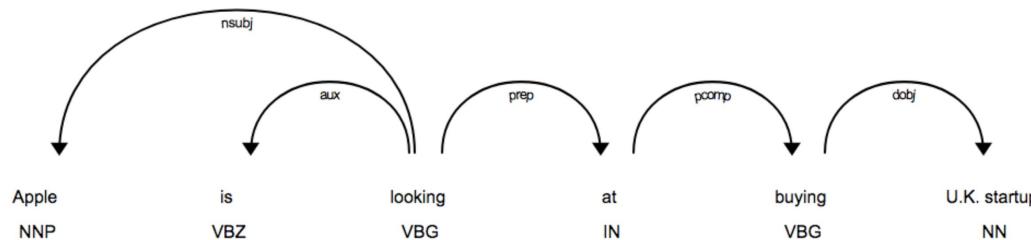
```
import spacy  
  
nlp = spacy.load('en')  
doc = nlp(u'Apple is looking at buying U.K. startup')
```



# spaCy

```
doc = nlp(u"Apple is looking at buying U.K. startup")

for token in doc:
    print(token.text, token.pos_, token.tag_)
```



# spaCy



```
doc = nlp(u"Apple is looking at buying U.K. startup")

for ent in doc.ents:
    print(ent.text, ent.start_char, ent.end_char, ent.label)
```

Apple 0 5 ORG

U.K. 27 31 GPE

Apple **ORG** is looking at buying **U.K. GPE** startup



# spaCy

```
dog, cat, banana = nlp(u"dog cat banana")
```

```
dog.similarity(cat) → 0.80
```

```
cat.similarity(dog) → 0.80
```

```
dog.similarity(banana) → 0.24
```

