Processamento de linguagem natural com Python

Coordenador
Fernando Ferreira

INSTITUTO INFNET 1994

Institucional

Reitor

Prof. Eduardo Ramos
Eduardo.ramos@Infnet.edu.br

Coordenação do Curso

Prof. Fernando Ferreira
Fernando.gFerreira@infnet.edu.br

Gerente Acadêmica

Ana Curi
ana.curi@Infnet.edu.br

Coordenador/Professor

Prof. Fernando Ferreira

fernando.gferreira@infnet.edu.br

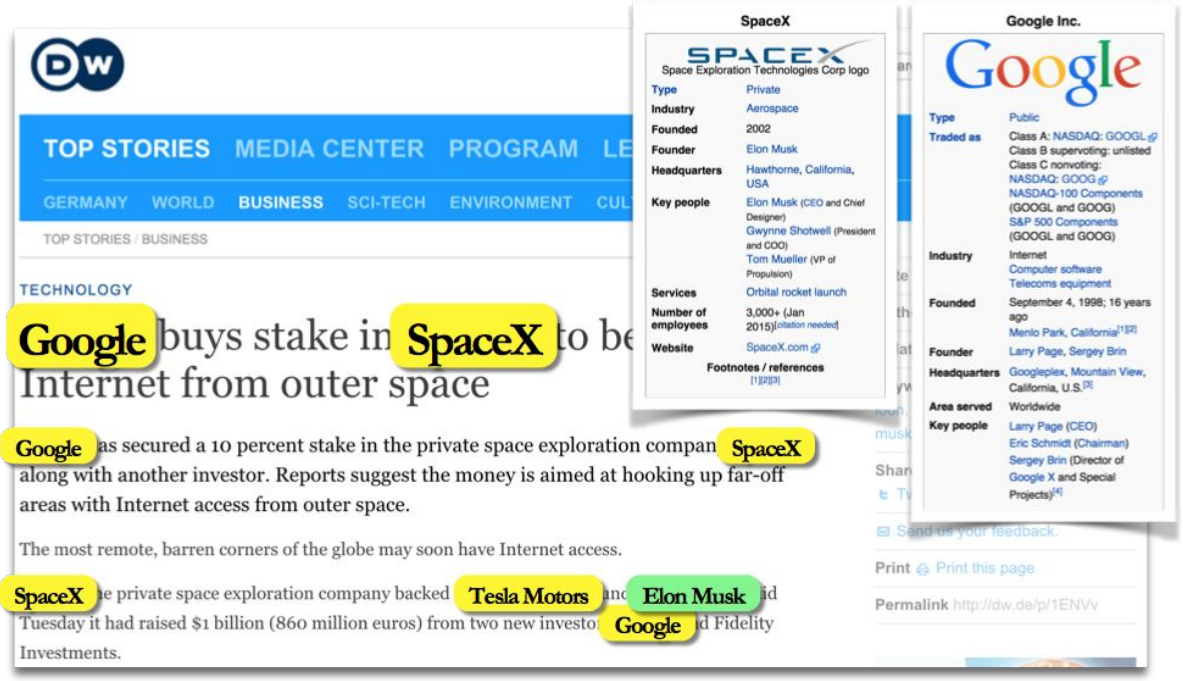
https://www.linkedin.com/in/nandoferreira/

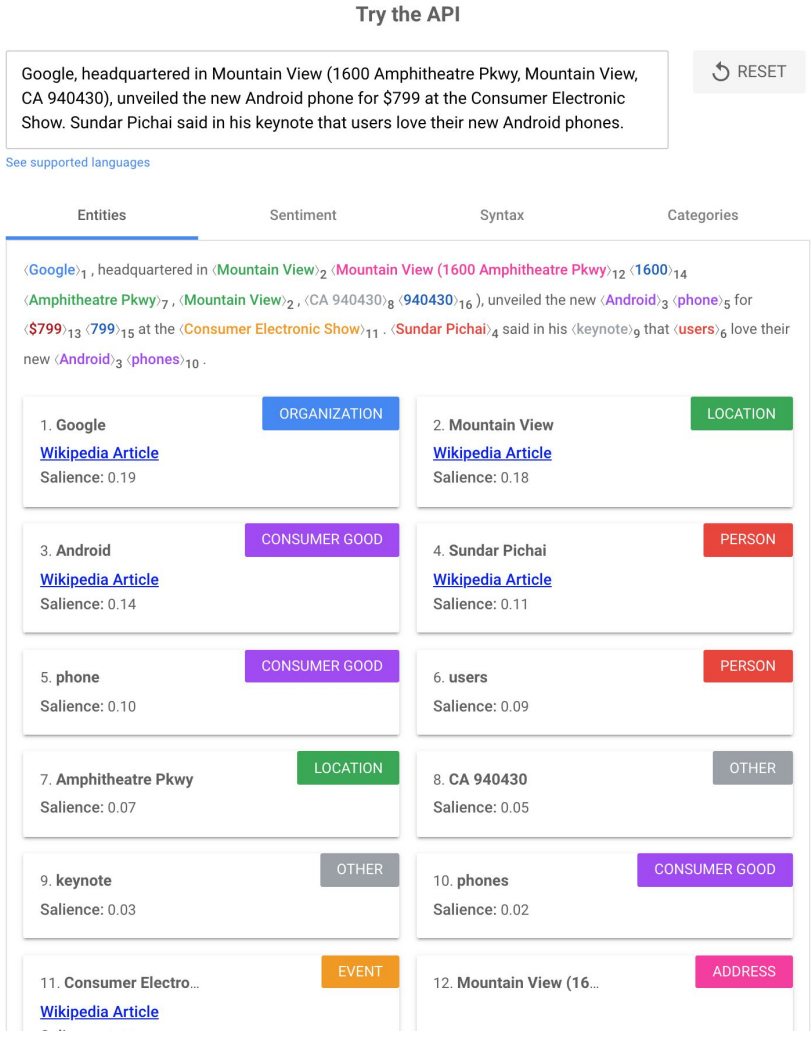
Mini-bio

- 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

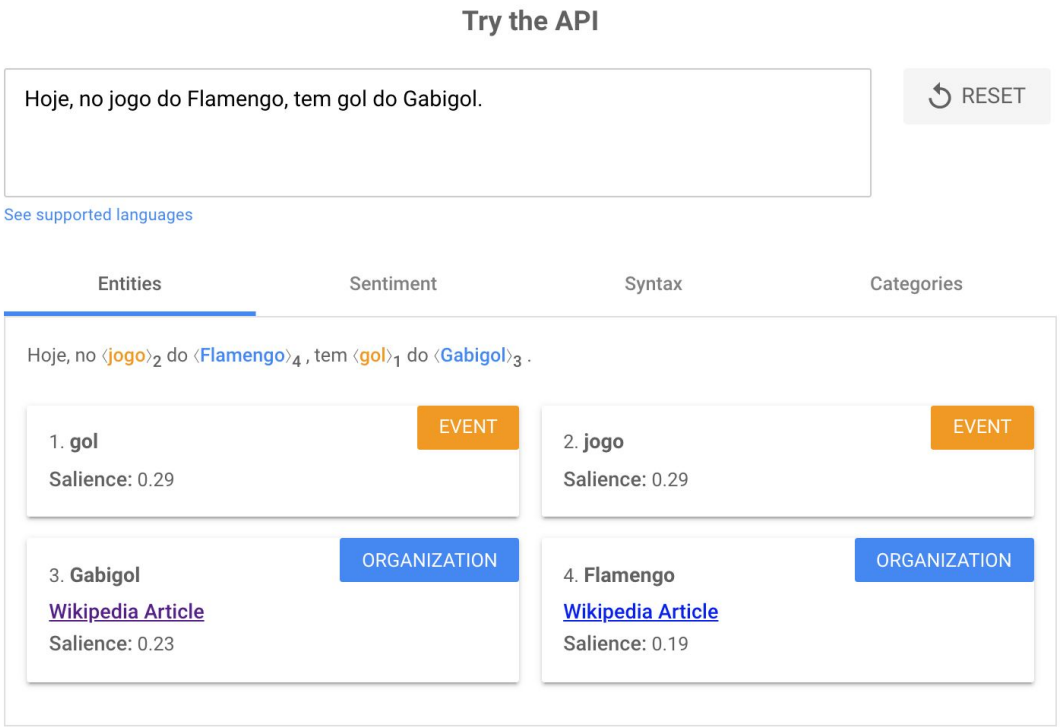
NER: Named Entity Recognition

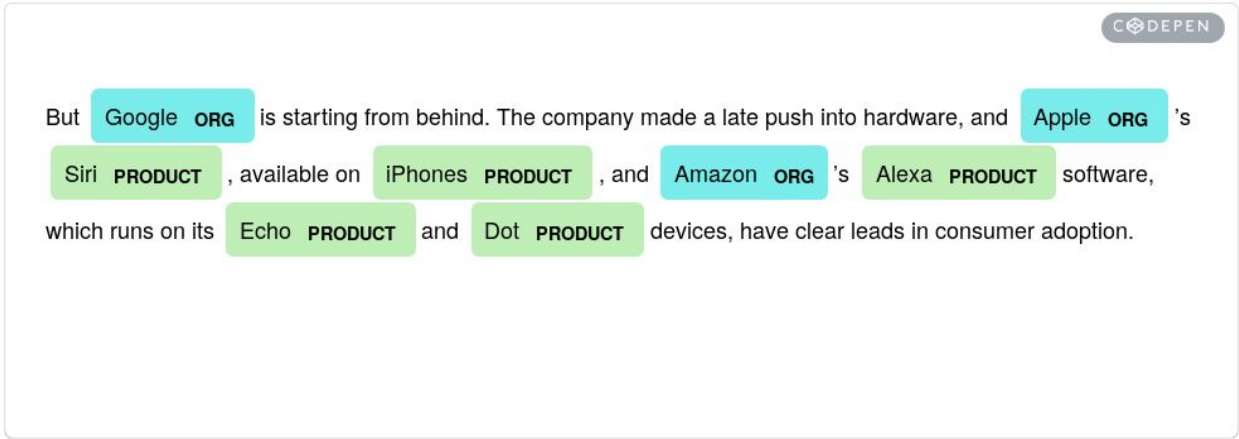
Relembrando





Gerado com Google Cloud Natural Language Demo


Try the API
Hoje, no jogo do Flamengo, tem gol do Gabigol.
RESET
See supported languages
Entities
Sentiment
Syntax
Categories
Hoje, no ⟨jogo⟩2 do ⟨Flamengo⟩4 , tem ⟨gol⟩1 do ⟨Gabigol⟩3 .
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


CODEPEN
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**

Fonte: **Word Embeddings for Named Entity Recognition, Braune, F. 2017**

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

Fonte: CoNLL-2003

supervisionada

Dado o segmento, treinar o classificador para indicar:

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

Classificação

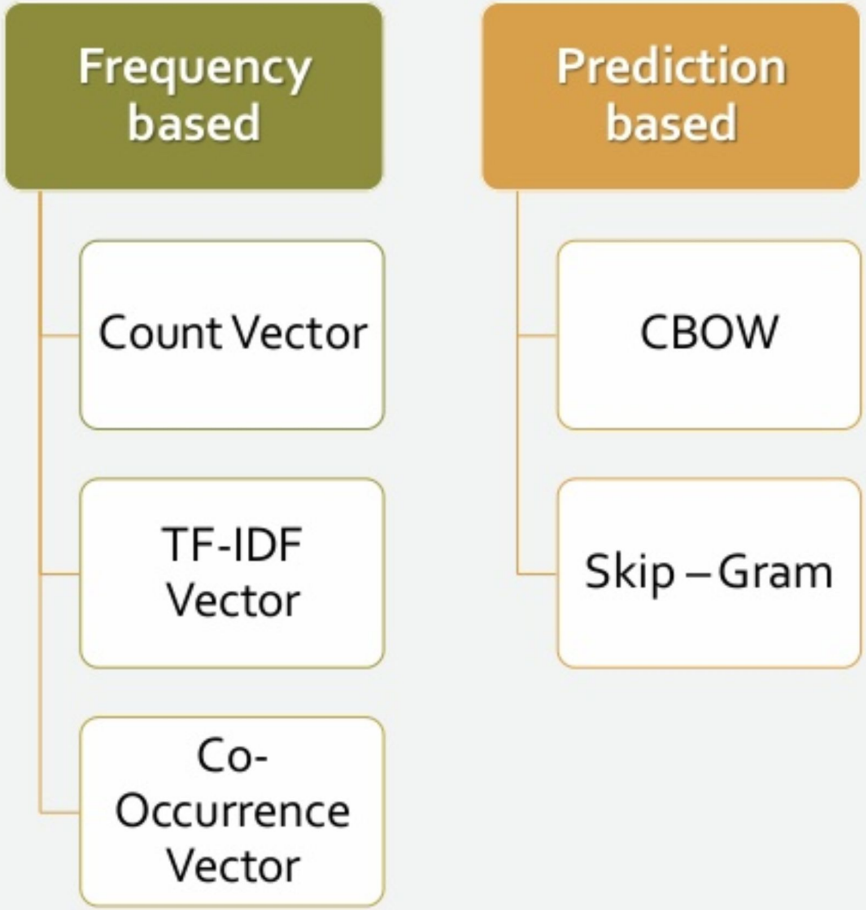
Tarefa: **Trump** attacks BMW and Mercedes

P: Trump é uma entidade nomeada?

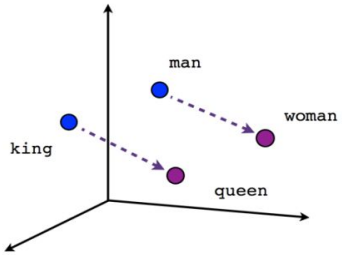
R: SIM, Trump é uma **PESSOA**


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.

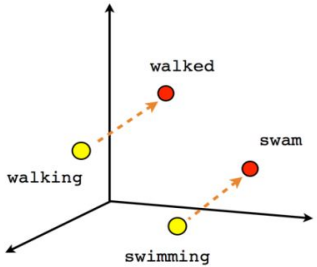
Predicted Word Embedding


Frequency based
Count Vector
TF-IDF Vector
Co-Occurrence Vector
Prediction based
CBOW
Skip – Gram

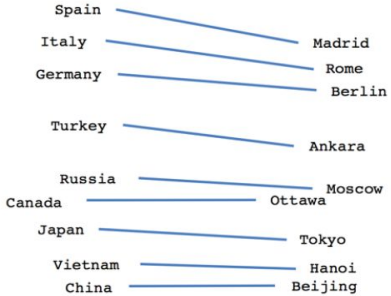
Word2Vec



Male-Female



Verb tense

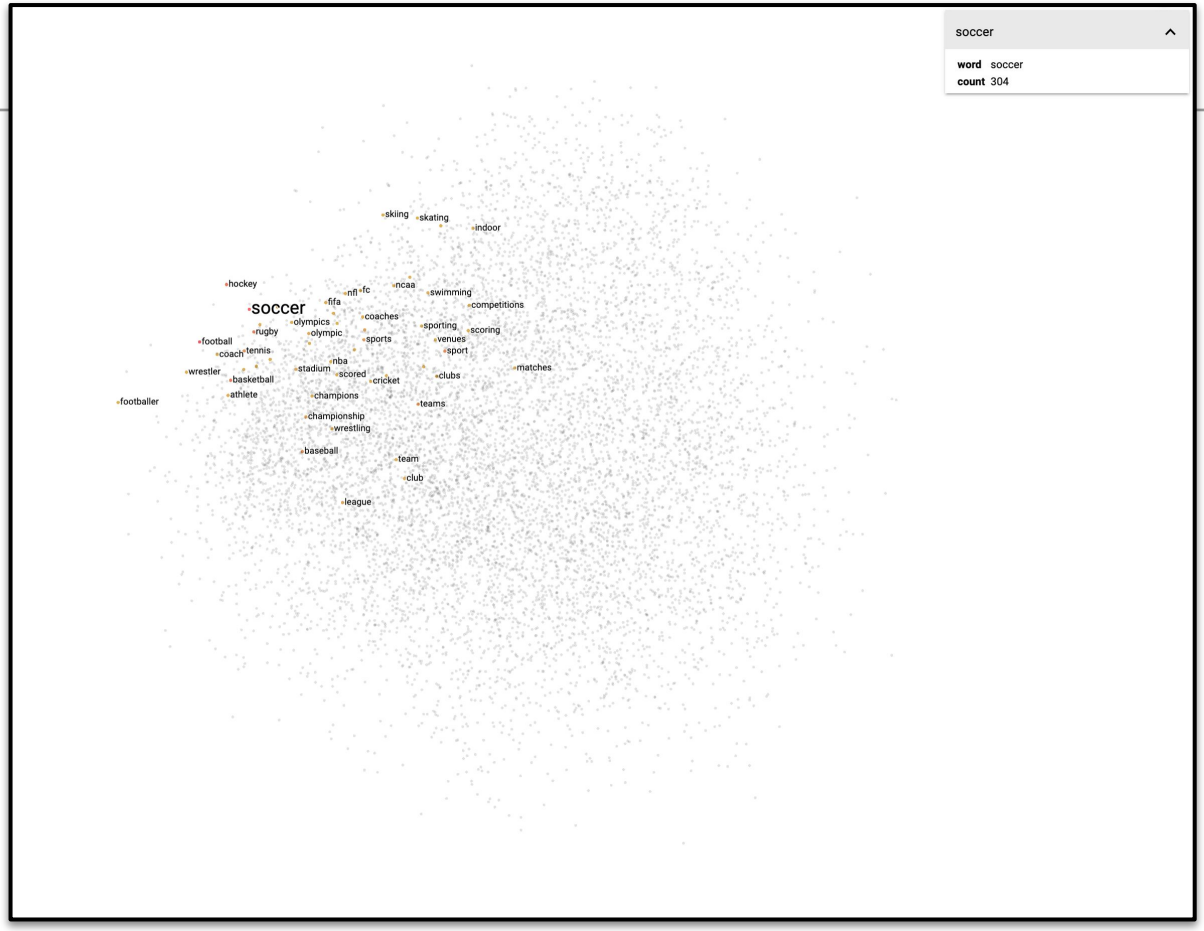


Country-Capital

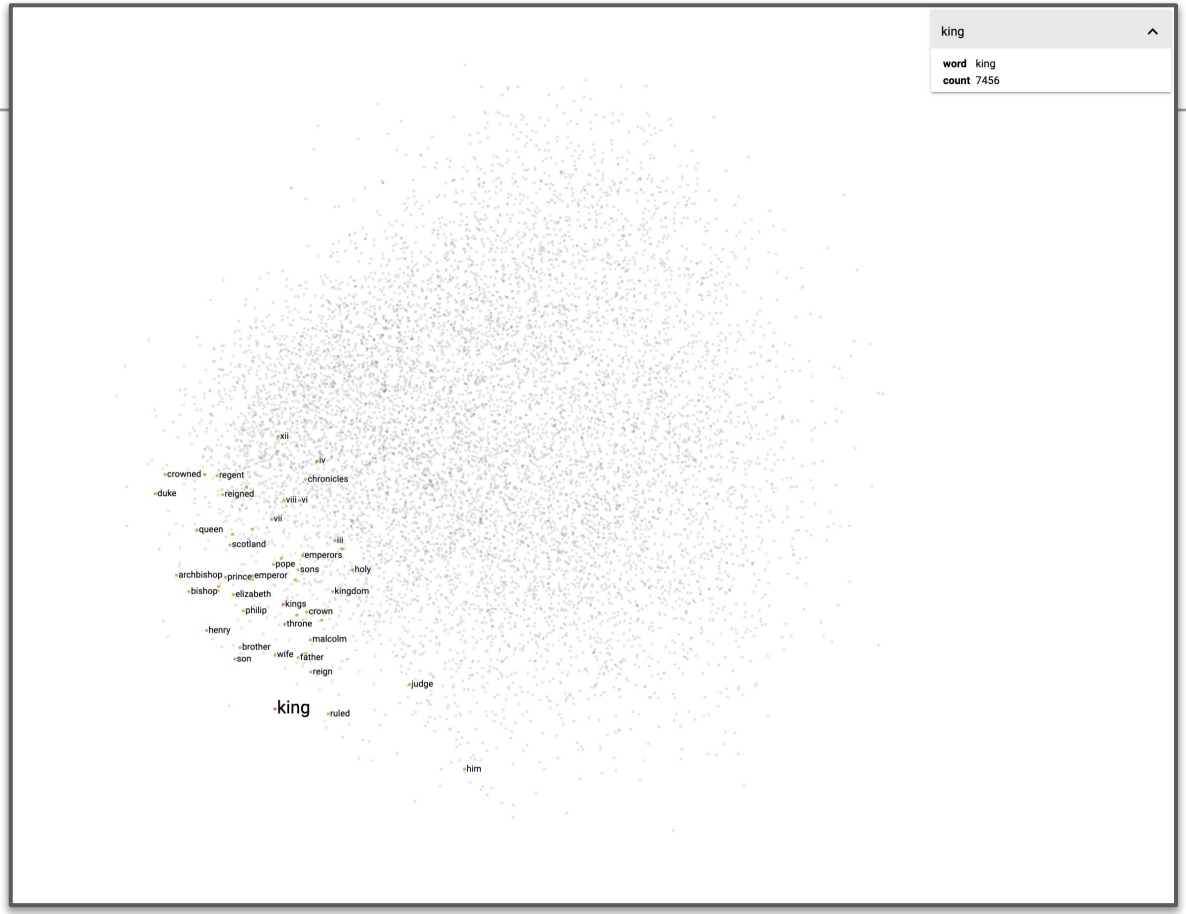
Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese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

MIKOLOV et al. **Distributed Representations of Words and Phrases and their Compositionality**.

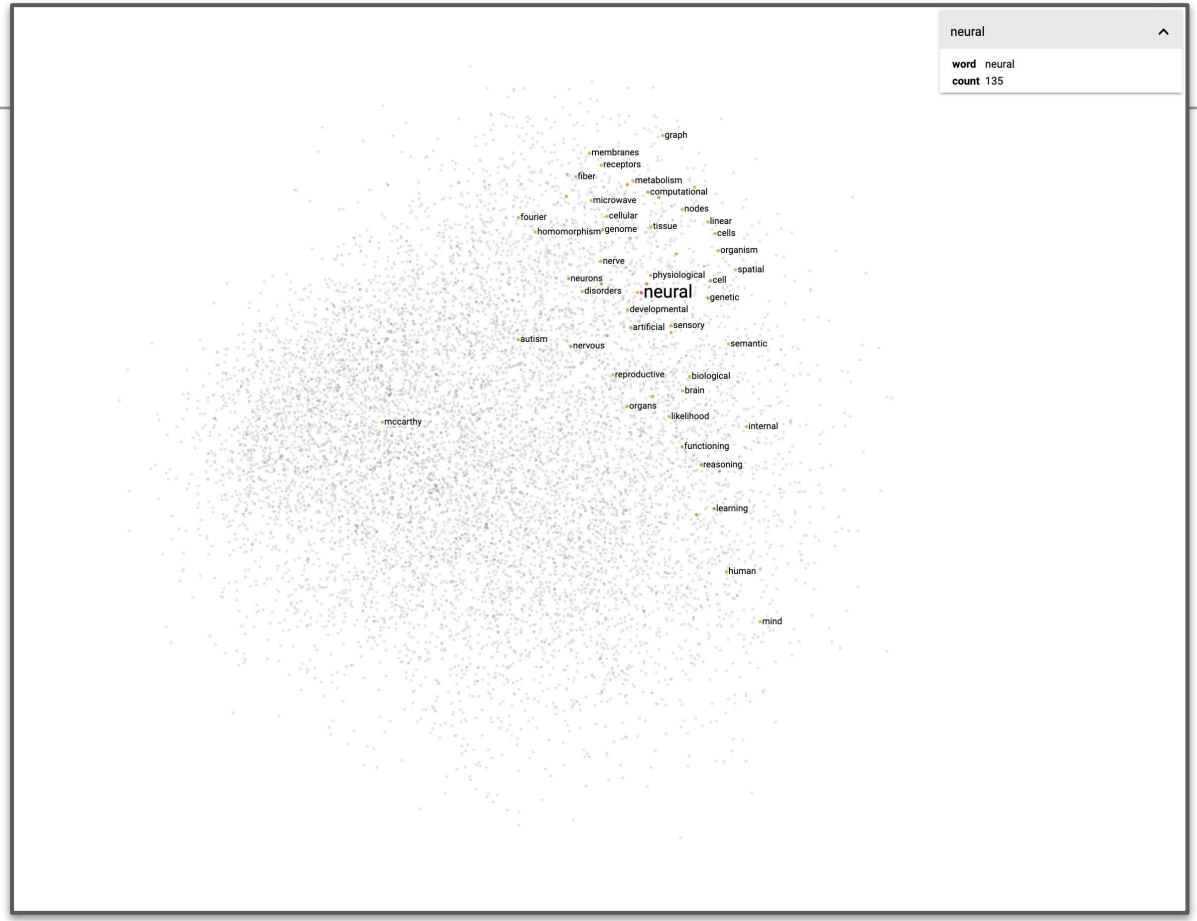
https://projector.tensorflow.org/



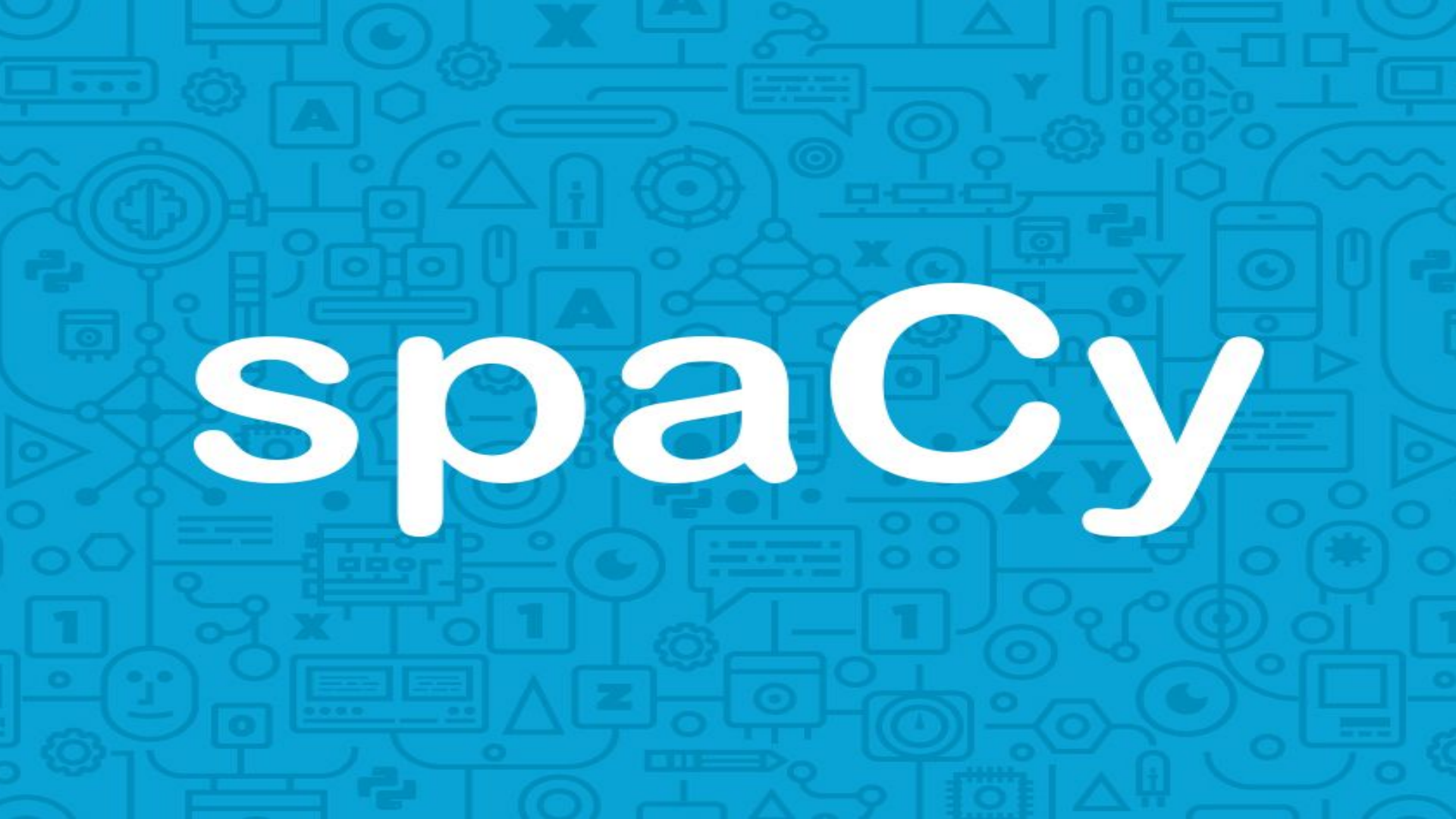
https://projector.tensorflow.org/



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

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
```

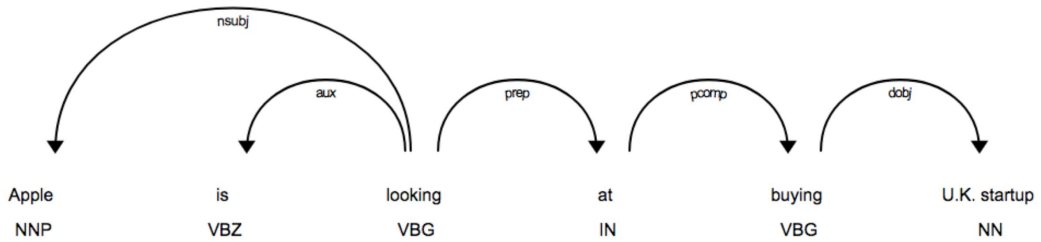
```
import spacy

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_)
```

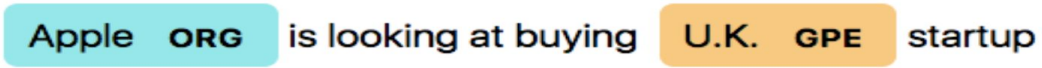


spaCy

```
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
```



spaCy

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

dog.similarity(cat)  0.80
cat.similarity(dog)  0.80
dog.similarity(banana)  0.24
```