



# IA e Processamento de Linguagem Natural

Thalita de Melo



# Inteligência Artificial e Machine Learning

Temos diversas subáreas dentro da inteligência artificial:

- Deep learning
- Visão computacional
- Processamento de linguagem natural



# PLN - Processamento de Linguagem Natural

O objetivo principal do PNL é permitir que os computadores compreendam, interpretem e gerem texto ou fala da mesma forma que os seres humanos.

Alguns exemplos de uso são:

- Análise Morfológica
- Tradução Automática
- Geração de Linguagem Natural
- Sumarização de Texto
- Extração de Informações
- Reconhecimento de Entidades (NER)



# Reconhecimento de Entidades (NER)

Digamos que você recebeu um dataset onde é preciso encontrar certas entidades como cidade, estado...

- Exemplo 1: FAZENDA NATANAEL, ZONA RURAL, RIO PARDO DE MINAS/MG
- Exemplo 2: FAZENDA ALEGRIA, VOLTA REDONDA-RJ

Uma opção é utilizar o Regex, outra seria utilizar a IA Generativa (chatgpt, bard...) e temos também o PLN, mais especificamente o NER.

Qual escolher?

# Reconhecimento de Entidades (NER)

When Sebastian Thrun **PERSON** started working on self - driving cars at Google **ORG** in 2007 **DATE** , few people outside of the company took him seriously . “ I can tell you very senior CEOs of major American **NORP** car companies would shake my hand and turn away because I was n’t worth talking to , ” said Thrun **PERSON** , in an interview with Recode **ORG** earlier this week **DATED** .



# Usos do NER

- Categorização de Notícias
- Separação de dados não estruturados
- Reconhecimento de entidades em consultas médicas
- Chatbots (identificação de palavras chaves)
- Classificação de e-mails

# Ensinando a Máquina

Em alguns casos os modelos pré-treinados não são o suficiente

Modelo usado: pt\_core\_news\_lg

```
frase = "CERÂMICA J.A, RODOVIA PI, SENTIDO DE BARRAS A CABACEIRAS, KM 7, Nº S/N, ZONA RURAL, BARRAS/PI. "  
doc = nlp(frase)  
  
for entidade in doc.ents:  
    print("palavra:", entidade.text)  
    print("entidade:", entidade.label_)
```

  

```
palavra: RODOVIA PI  
entidade: LOC  
palavra: SENTIDO  
entidade: LOC  
palavra: CABACEIRAS  
entidade: LOC  
palavra: Nº S/N  
entidade: MISC  
palavra: BARRAS  
entidade: LOC  
palavra: PI  
entidade: LOC
```

# Ensinando a Máquina

Em alguns casos os modelos pré-treinados não são o suficiente

Modelo usado: en\_core\_web\_lg

```
import spacy

nlp = spacy.load("en_core_web_lg")

referencia = "Abberger, K. (2006). \"Another Look at the Ifo Business Cycle Clock.\" Journal of \\  
Business Cycle Measurement and Analysis, 2005/3.\\  
Abberger, K. and W. Nierhaus (2008). \"Die ifo Konjunkturuhr: Ein Präzisionswerk \\  
zur Analyse der Wirtschaft.\" ifo Schnelldienst, 61(23), 16-24. \\  
Agresti, A. and B. Mojon (2001). \"Some Stylized Facts on the Euro Area Business \\  
Cycle.\" Working paper series no. 95, ECB."

doc = nlp(referencia)

for entidade in doc.ents:
    print('palavra:', entidade.text)
    print('label: ', entidade.label_)
    print('-----')
```

```
palavra: Abberger
label: PERSON
-----
palavra: K.
label: PERSON
-----
palavra: 2006
label: DATE
-----
palavra: Another Look at the Ifo Business Cycle Clock
label: WORK_OF_ART
-----
palavra: Journal of Business Cycle Measurement and Analysis
label: ORG
```





# Doccano

Doccano é um “marcador” de textos open-source. Com ele podemos fazer anotações (marcações) para classificação, rotulagem, entre outras coisas. É possível criar rotulos para dados de análise de sentimentos, NER, sumarização de texto, etc.

<https://github.com/doccano/doccano>





# spaCy

O spaCy é uma biblioteca open-source desenvolvida para o processamento de linguagem natural. Atualmente suporta mais de 73 idiomas e possui diversos modelos pré-treinados. Com o spaCy podemos fazer análises de sentimento, tokenização de textos, reconhecimento de entidades (NER), classificação de textos, análise morfológica ...

O spaCy é treinado como um detetive linguístico para reconhecer diferentes partes da linguagem. Assim como um detetive aprende a distinguir pistas importantes, o spaCy é treinado em grandes conjuntos de dados de texto para entender padrões e características das palavras.

<https://spacy.io/>

# spaCy

# O Processo

✓

🔍

⇅

📅

💬

✂

🗑

📄

1 of 1410 |< < > >|

CERÂMICA J.A, RODOVIA PI, SENTIDO DE BARRAS A CABACEIRAS, KM 7, Nº

S/N, ZONA RURAL, BARRAS/PI.

•CITY •STATE

Progress

Total 1410

Complete 29

3%

Label Types

CITY STATE

Key Value

No data available



# Dados com label

```
[{'id': 87, 'text': 'CERÂMICA J.A, RODOVIA PI, SENTIDO DE BARRAS A CABACEIRAS, KM 7, Nº  
S/N, ZONA RURAL, BARRAS/PI', 'label': [[84,90,"CITY"],[91,93,"STATE"]], 'Comments': []},]
```

# Resultados

```
[8] # Teste
    texto_de_teste = "FAZENDA DAS PALMEIRAS, NA ZONA RURAL, RIO VERMELHO/MG"

    # Processar texto de teste
    doc = nlp(texto_de_teste)

    # Extrair entidades nomeadas reconhecidas
    for ent in doc.ents:
        print(ent.text, ent.start_char, ent.end_char, ent.label_)
```

```
RIO VERMELHO 38 50 CITY
MG 51 53 STATE
```

# Resultados

```
[ ] # Teste 4
    texto_de_teste = 'Amanpreet Singh and Narina Thakur, "A review of supervised machine learning algorithms", \
        3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016.'

    doc = nlp(texto_de_teste)

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

Amanpreet Singh 0 15 PERSON

Narina Thakur 20 33 PERSON

A review of supervised machine learning algorithms", 36 88 WORK

International Conference on Computing for Sustainable Global Development ( 94 168 ORG

2016 179 183 YEAR



# Obrigada!

Linkedin: <https://www.linkedin.com/in/thalita-de-melo-soares/>

Github: <https://github.com/thalita-de-melo/apresentacao-analytica>

Exemplos usados:

[https://colab.research.google.com/drive/1m\\_zUu5H2QCeZ-gAk9nURp01WtqdME-Vz?usp=sharing](https://colab.research.google.com/drive/1m_zUu5H2QCeZ-gAk9nURp01WtqdME-Vz?usp=sharing)

<https://colab.research.google.com/drive/1SHuTP-luLptDivjkT4dUhd0ffkHDky24?usp=sharing>

