

apresentado em

Faixa 2

Conhecimentos básicos sobre dados para repórteres de IA

O que abordaremos neste módulo

- Uma breve e incompleta história de como os dados evoluíram
- Dados no âmbito da IA
- Três exemplos:
 - Dados estruturados alimentando algoritmos de tomada de decisão
 - Dados longitudinais que alimentam suas timelines nas redes sociais
 - Dados de texto para grandes modelos de linguagem

Vamos começar com um exercício



Que tipo de dados existem?

Vamos analisar um tweet.

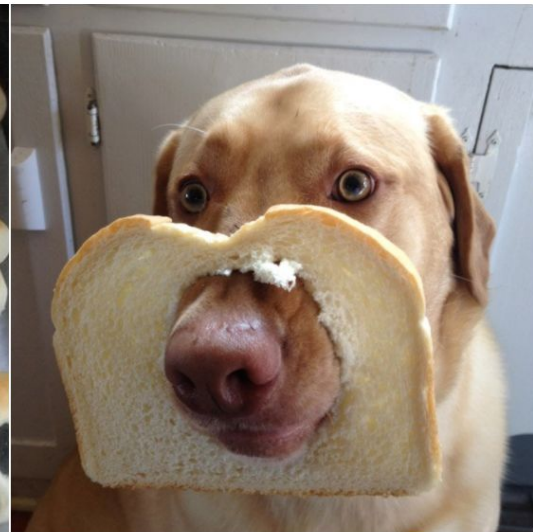
1. Leia isto.
2. Dê uma risada.
3. Depois me diga que tipo de dados esse tweet pode gerar.



Yael
@elle91

Follow

Purebread dogs vs. inbread dogs



RETWEETS

1,446

LIKES

2,758



12:25 PM - 22 Feb 2017

22

1.4K

2.8K

Pessoa que postou:

- Link da imagem
- Nome de exibição
- Nome de usuário no Twitter

Texto

Mídia

Engajamento:

- Retuítes
- Curtidas
- Respostas (mais abaixo)

Data



Yael
@elle91

+ Follow

Purebread dogs vs. inbread dogs



RETWEETS

1,446

LIKES

2,758



12:25 PM - 22 Feb 2017



22



1.4K



2.8K

O mesmo tweet,
formato diferente!
Todo o site X.com é
alimentado por fluxos
de dados como este.

APIs ou Interface
de Programação de
Aplicativos*

*palavra sofisticada para
um fluxo de dados em um
formato que
robôs/computadores
podem entender 🤖

```
1  {
2    "contributors": null,
3    "truncated": false,
4    "text": "Purebred dogs vs. inbred dogs https://t.co/e1nC0d7EkR",
5    "is_quote_status": false,
6    "in_reply_to_status_id": null,
7    "id": 834454023260532737,
8    "favorite_count": 2760,
9    "source": "<a href='\"http://twitter.com/download/android\"' rel='\"nofollow\"'>Twitter for
10   * Android</a>",
11    "retweeted": false,
12    "coordinates": null,
13    "entities": {
14      "symbols": [],
15      "user_mentions": [],
16      "hashtags": [],
17      "urls": [],
18      "media": [ {
19        "expanded_url": "https://twitter.com/elle91/status/834454023260532737/photo/1",
20        "display_url": "pic.twitter.com/e1nC0d7EkR",
21        "url": "https://t.co/e1nC0d7EkR",
22        "media_url_https": "https://pbs.twimg.com/media/C5STfk1WQAEzRet.jpg",
23        "id_str": "834454001643044865",
24        "sizes": {
25          "large": {
26            "h": 413, "resize": "fit", "w": 550
27          },
28          "small": {
29            "h": 413, "resize": "fit", "w": 550
30          },
31          "medium": {
32            "h": 413, "resize": "fit", "w": 550
33          }
34        }
35      } ]
36    }
37  }
```



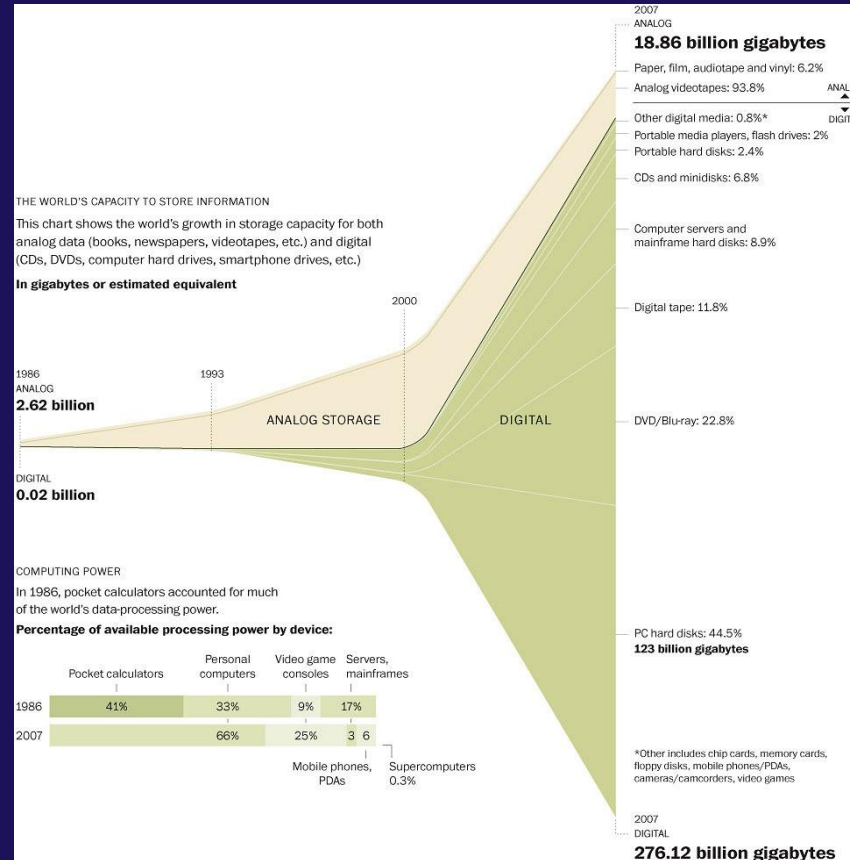
Os dados estão em toda parte!

A internet é alimentada por dados



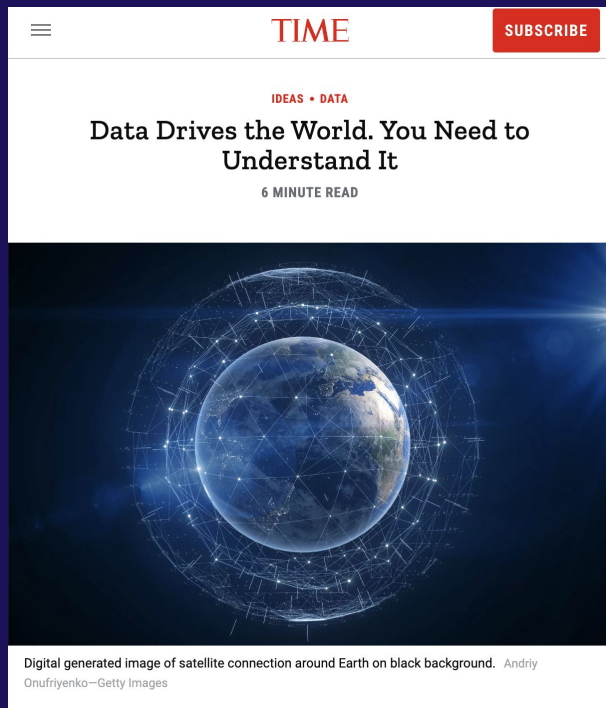
O que aconteceu
com os dados?

Há cada vez mais dados



Fonte: Washington Post (2011)

Agora há muito mais do que apenas dados governamentais



Fonte: IBM via Time



Fonte: CNET

Exemplo de pegada digital: WSJ



How Pizza Night Can Cost More in Data Than Dollars

Even a low-key evening at home can mean handing over a trove of personal information to high-tech companies

By Stephanie Stamm, Tripp Mickle and Jessica Kuronen

Published April 10, 2018 at 5:30 a.m. ET

The smartphones, Facebook accounts and other technology products deeply embedded in modern life help people get more things done every day. They also gather more information about us than we often realize.

But Facebook's crisis [over how it handles and protects user data](#) has led some to ask: What data am I giving up?

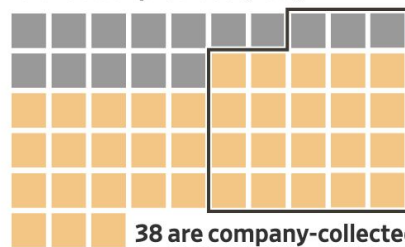
Imagine "Sally" sets up a pizza-and-movie night with her friend "Kristen." The Wall Street Journal reviewed privacy statements to assess just how much data could be unknowingly shared on top of the price of that pepperoni pie.

O custo dos dados

Os repórteres descobriram que um simples pedido de pizza gerou pelo menos 53 tipos de dados diferentes. Os dados detalhados refletem as informações que as empresas podem coletar de acordo com suas políticas de privacidade, termos de serviço e documentos relacionados.

Data points collected in this scenario

15 are user-provided (28%)



23 items
come from
Facebook

38 are company-collected (72%)

Fonte: [WSJ](#)

Exemplo de pegada digital: servidores de metadados



Vista aérea de alguns dos datacenters do Facebook

Big Data

Está fazendo muita gente ganhar muito dinheiro

 Meta Platforms (Facebook) 

#7	\$941.51 B	
Rank	Marketcap	Country
\$366.37	-0.30%	170.66%
Share price	Change (1 day)	Change (1 year)

 Internet

 Tech

 AI

Categories

Market cap

Revenue

Earnings

More ▾

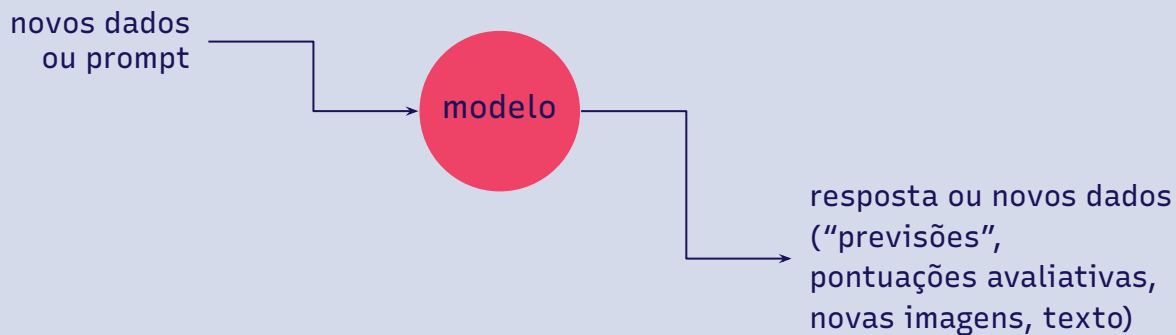
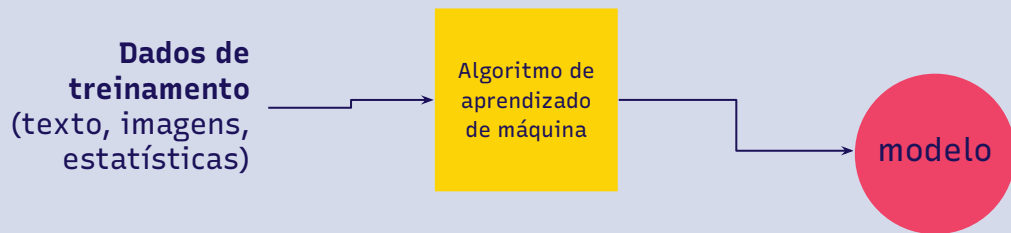
Market capitalization of Meta Platforms (Facebook) (META)

Market cap: \$941.51 Billion

As of January 2024 **Meta Platforms (Facebook)** has a market cap of **\$941.51 Billion**. This makes Meta Platforms (Facebook) the world's **7th** most valuable company by market cap according to our data. The market capitalization, commonly called market cap, is the total market value of a publicly traded company's outstanding shares and is commonly used to measure how much a company is worth.

Ciclo de vida da IA

Em que parte do ciclo de vida estão os dados de treinamento?

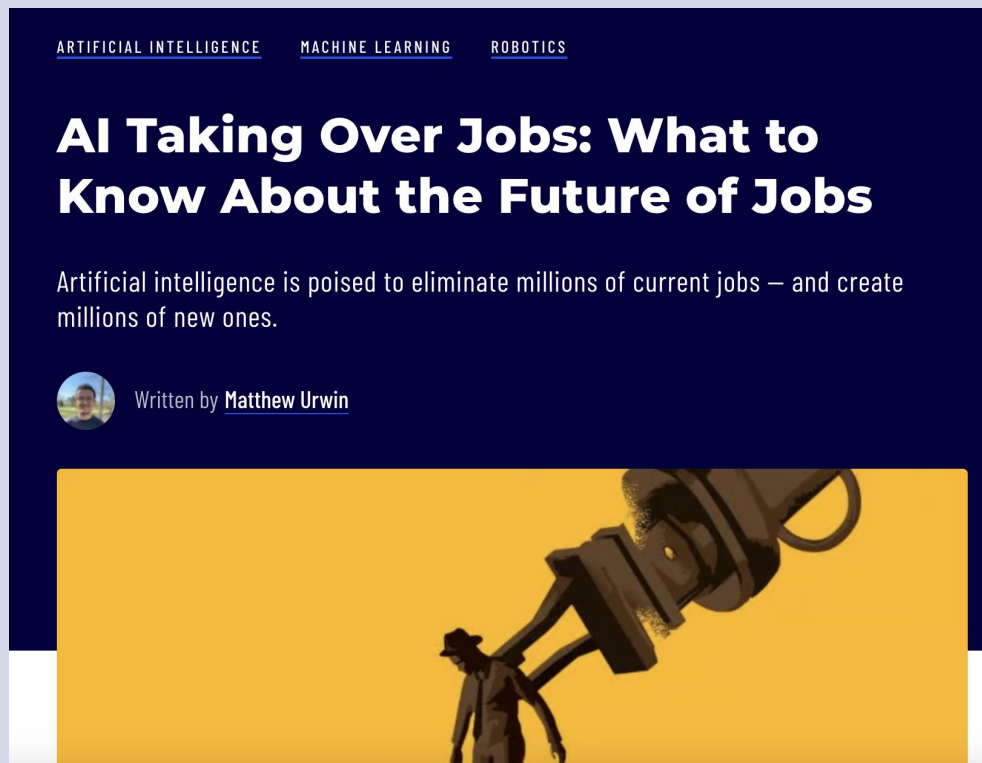


Prevemos com 95%
de certeza que
esta imagem
retrata o papa

ou

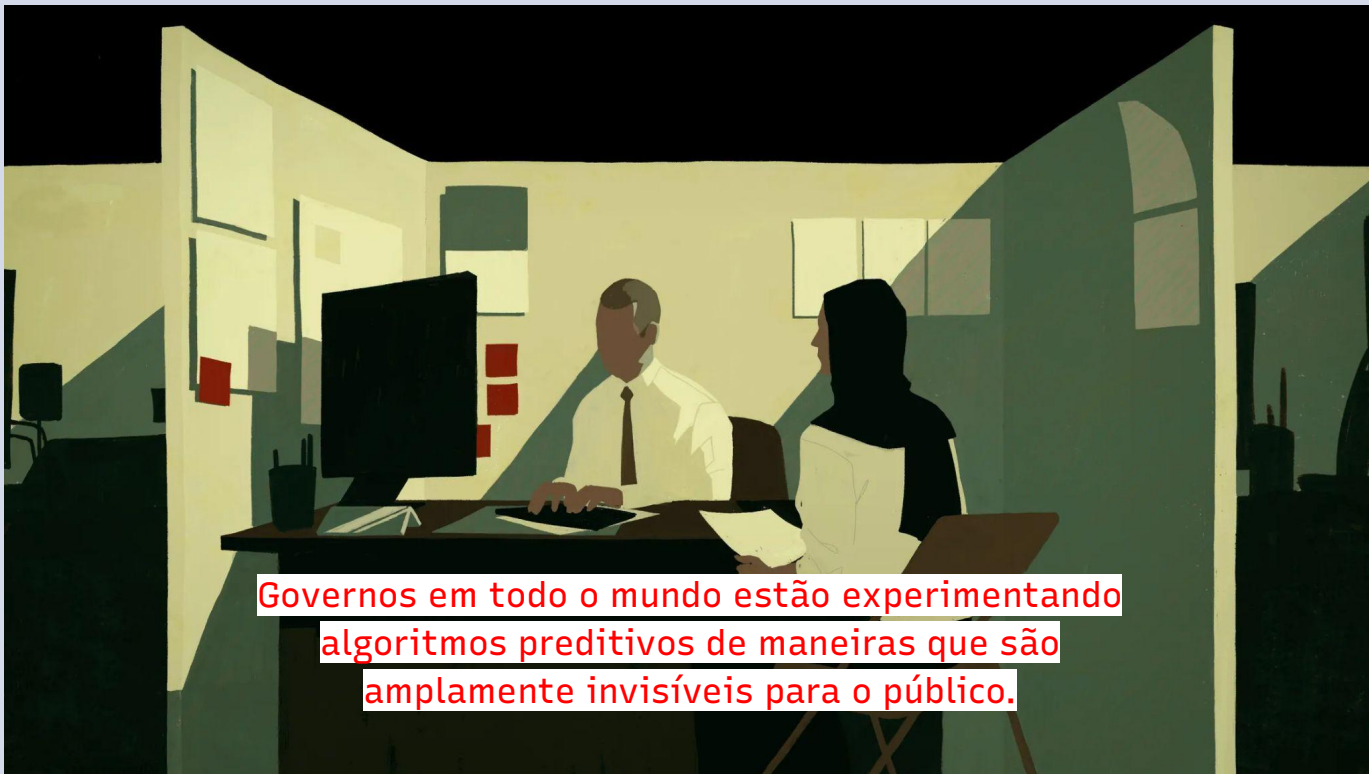


IA: os dados estão mudando nossa sociedade



Source: <https://builtin.com/artificial-intelligence/ai-replacing-jobs-creating-jobs>

IA: os dados estão mudando nossa sociedade



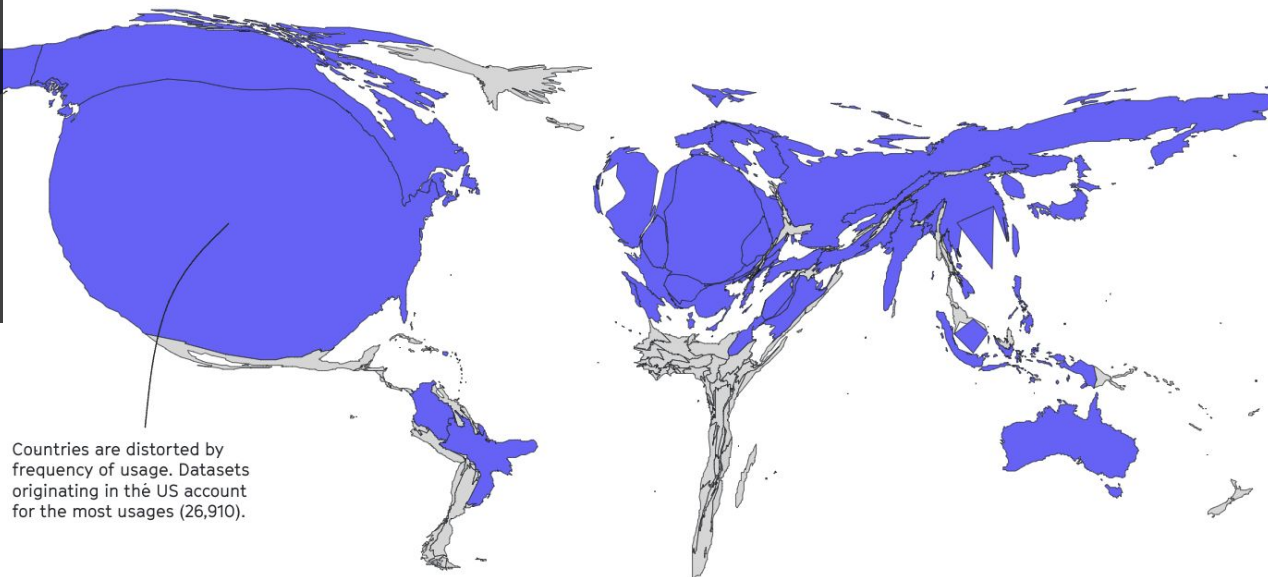
Governos em todo o mundo estão experimentando algoritmos preditivos de maneiras que são amplamente invisíveis para o público.

Muitos dos dados estão reproduzindo preconceitos

Em milhares de artigos, são os mesmos conjuntos de dados de apenas alguns países que são usados com mais frequência para avaliar o desempenho de modelos de aprendizado de máquina em todos os lugares.

Frequency of dataset usage by country

● Usage of datasets from here ● No usage of datasets from here



Dados estruturados alimentando algoritmos de tomada de decisão



De acordo com o programa Old Age Samman Allowance de Haryana, na Índia, pessoas com 60 anos ou mais, cuja renda somada à do cônjuge não exceda 300.000 rúpias (US\$ 3.600) por ano, têm direito a uma pensão mensal de 2.750 rúpias (US\$ 33).

Em junho de 2020, o governo local começou a usar um sistema algorítmico recém-criado – o Repositório de Dados de Identidade Familiar ou o banco de dados Parivar Pehchan Patra (PPP) – para determinar a elegibilidade das pessoas que pediram esse benefício de assistência social.



Dhuli Chand led a wedding procession to prove to government officials he was alive and should be paid his pension. Image courtesy of The Reporters' Collective.

O PPP é um ID exclusivo de oito dígitos fornecido a cada família no estado e contém detalhes sobre nascimento e morte, casamento, emprego, propriedade e imposto de renda, entre outros dados, dos membros dessa família. Ele mapeia as informações demográficas e socioeconômicas de cada família, vinculando vários bancos de dados do governo para verificar se eles estão aptos ou não a programas de assistência social.

O estado afirmou que o PPP criou “dados autênticos, verificados e confiáveis de todas as famílias” e tornou obrigatório para os cidadãos acessarem todos os programas de assistência social.

Mas, na prática, o PPP marcou erroneamente Chand como “falecido”, negando sua pensão por vários meses. Pior ainda, as autoridades não alteraram seu status de “falecido” mesmo depois de ele ter se reunido com autoridades pessoalmente várias vezes.



Dhuli Chand led a wedding procession to prove to government officials he was alive and should be paid his pension. Image courtesy of The Reporters' Collective.

Dados longitudinais alimentando
suas timelines nas redes sociais



Dados usados pelo Facebook

Pastas incluídas em um arquivo de dados do Facebook que pode ser baixado

O algoritmo do Facebook é um sistema de classificação que usa aprendizado de máquina para organizar o conteúdo nos feeds dos usuários. Dados usados:

- Dados que você insere no Facebook
- Dados que você produz ao usar o Facebook
- Dados que são rastreados sobre você por páginas com botões do Facebook

Para ter uma ideia da profundidade desses dados, os usuários podem baixar seus dados seguindo as instruções do Facebook:

https://www.facebook.com/help/1701730696756992?helpref=hc_global_nav

Mais informações sobre o assunto aqui:

Como o Facebook te observa online ([BuzzFeed News](#))

3 maneiras simples pelas quais fornecemos uma grande quantidade de informações pessoais ao Facebook e a aplicativos aleatórios ([BuzzFeed News](#))

- about_you
- ads
- apps_and_websites
- calls_and_messages
- comments
- events
- following_and_followers
- friends
- groups
- https/
- likes_and_reactions
- location_history
- marketplace
- messages
- other_activity
- pages
- payment_history
- photos_and_videos
- posts
- profile_information
- saved_items
- search_history
- security_and_information
- your_places



Post title

Exemplo: algoritmo do Facebook

The top post on a Facebook user's news feed, shown as the biggest box, is a prized position based on thousands of data points related to the user and post itself, such as the poster, reactions and comments.



Fonte: [The Washington Post](#)

Show your age

Exemplo: algoritmo do Facebook

Trip photos

Since 2018, the algorithm has elevated posts that **encourage interaction**, such as ones popular with friends. This broadly prioritizes posts by **friends and family** and **viral memes**, but also **divisive content**.

Zoe A. and 60.8K others

10.7K comments



Like



Comment



Kai J. and 45.2K others



Like



Comment



Engaged!

Rap name

Fonte: [The Washington Post](#)

Player takes political stand

Spicy opinion piece

Exemplo: algoritmo do Facebook

Each user's feed reflects their expressed interests. For a subset of **extremely partisan** users, today's algorithm can turn their feeds into echo chambers of **divisive content** and **news**, of varying reputability, that support their outlook.

Fonte: [The Washington Post](#)

Dados de texto para processamento de linguagem natural



Exemplo: Google

A previsão de texto na pesquisa do Google é baseada em inúmeras pesquisas.

Quantas pessoas de coração partido pesquisaram o nome de Ryan Gosling no Google para descobrir se ele era casado?! (Além disso, muitas pessoas provavelmente também pesquisaram o nome de um famoso + esposa)



Alguns conceitos

- 1. Corpus**
- 2. Tokenização**
- 3. Radicalização
ou lematização**

Alguns conceitos

- 1. Corpus**
- 2. Tokenização**
- 3. Radicalização
ou lematização**



Alguns conceitos

1. Corpus
2. Tokenização
3. Radicalização ou lematização



Alguns conceitos

1. **Corpus**
2. **Tokenização**
3. **Radicalização
ou lematização**

O objetivo da radicalização ou lematização, é reduzir as formas flexionais e, às vezes, as formas derivadas de uma palavra a uma forma básica comum.

Por exemplo:

sou, são, é → ser

carro, carros, carro's, carros' → carro

O resultado desse mapeamento do texto será algo como:

os carros do menino são de cores diferentes
→ o carro do menino ser cor diferente

Alguns conceitos

1. **Corpus**
2. **Tokenização**
3. **Radicalização
ou lematização**

Radicalização versus lematização

Exemplo: "vi"

(Como em "Eu vi um pássaro, ou seria um avião? Acho que pode ter sido a Supergirl.")

Radical:

"vi" → v

Lema:

vi → ver*

* se o token for um verbo!

Analizando o texto

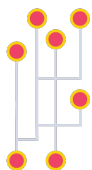
Depois de preparar seu corpus em tokens (como radicais ou lemas), você pode analisar o texto. Isso pode incluir:

- Criar listas das palavras mais utilizadas
- Prever quais palavras são mais prováveis de se seguirem (assim como em uma boa pesquisa!)
- Detectar o sentimento em um texto





AI Spotlight Series



**The AI
Accountability
Network**



Pulitzer Center