

Raspagem de dados | Twitter



Usando Python, mas não assusta.

Gustavo Sarti

sociólogo | educador

Redes sociais
@gmsarti
gmsarti@gmail.com

Coletivo
Máquina Tudo
maquinatudo.myportfolio.com



Por que o Twitter?

Relevância para o debate

+

Mais fácil de raspar dados

Raspagem de dados da internet (ou *web scraping*) é o processo de coletar dados públicos da internet.

O processo é diferente em cada site ou plataforma e exige um pouco de conhecimento específico sobre o site que será usado.

E esse Python?

Linguagem de programação um pouco mais fácil, mas muito poderosa e ótima para trabalhar com dados

Programar é uma boa habilidade de se ter. No mínimo ela aumenta sua autonomia enquanto pesquisador.

Raspar dados da internet nos possibilita acessar mais dados do que conseguiríamos processar em ferramentas de planilha.

Jupyter notebook

O artigo acadêmico está obsoleto e o futuro está aqui

Jupyter Notebook

O Jupyter Notebook é um app de código aberto que roda no seu navegador no computador. Nele é possível criar e compartilhar documentos com código (em algumas linguagens de programação), equações, visualizações e texto narrativo.

Os principais usos são: limpeza e transformação de dados, simulação numérica, modelagem estatística, visualização de dados, aprendizado de máquina etc.

Se um dia você quiser instalá-lo em seu computador eu recomendo que faça isso usando a [Anaconda](#). Mas vamos usar esses recursos todos sem precisar instalar nada e usando o poder computacional dos outros.

colab

Google Colab

<https://colab.research.google.com/notebooks/intro.ipynb>

kaggle

Kaggle

<https://www.kaggle.com/>

Começando os trabalhos

Nossa ideia é fazer um sistema que colete informações no Twitter e nos entregue alguma informação útil.

Para isso existe um ciclo típico de trabalho com dados, geralmente iniciando de uma *Questão de pesquisa*.

Reduzir tempo despendido em cada etapa mais mecânica libera mais tempo para análise e reflexão.



Máquina de salsicha

Um sistema que recolhe dados não estruturados e nos entrega alguma informação

Link para o notebook ->

Raspar_e_Analisar.ipynb

Arquivo Editar Ver Inserir Ambiente de execução Ferramentas Ajuda Todas as alterações foram salvas

+ Código + Texto

Twint

É a biblioteca que efetivamente faz a raspagem dos dados

Se você quiser se aprofundar sobre como usar e configurar pode ler na [documentação nesta página aqui.](#)

```
1 !twint -s "7 de setembro" --since 2021-08-25 -o /content/drive/MyDrive/Datasets/sete_de_setembro.csv --c
```

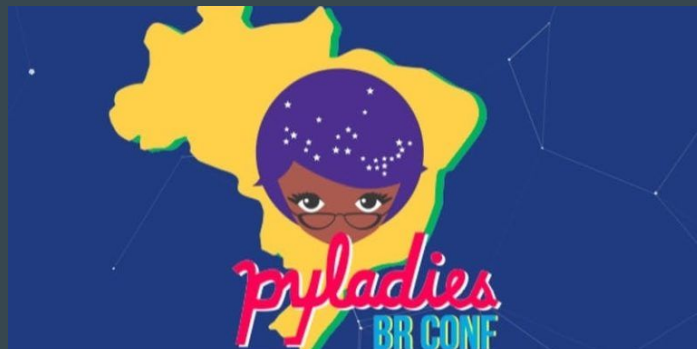
ID	Data	Time	Conteúdo
1430440075821830144	2021-08-25	08:01:05	+0000 <@reluday> @OmarAzizSenador Vocês são farinha do mesmo saco. T
1430439628537012230	2021-08-25	07:59:18	+0000 <@eliezerxp> RT @Isentoosi: Após João Doria afastar o coronel
1430439623839387653	2021-08-25	07:59:17	+0000 <@eliezerxp> RT @BrazillFight: SILAS MALAFAIA "Pela primeira v
1430439584375132163	2021-08-25	07:59:07	+0000 <@eliezerxp> RT @BolsonaroSP: O povo de Deus é contra o aborto
1430439056140390404	2021-08-25	07:57:01	+0000 <@LeoRezendeNog> OMBRO AMIGO DA CPI A MORAES/ PM NO 7 DE SETEM
1430438709946687493	2021-08-25	07:55:39	+0000 <@ciba_ze> @Victorinogustav #Dia7VaiserGiganchegada #7deSetembr
1430438573476618243	2021-08-25	07:55:06	+0000 <@gualbarino> Isso foi ontem. Precisamos fazer MUITO maior em
1430438482938322946	2021-08-25	07:54:45	+0000 <@BlogdoNoblat> Empresários do 7 de Setembro golpista e ricos
1430435656652795905	2021-08-25	07:43:31	+0000 <@mariolima1994> @ForçaAcip @PMarcos07 Podemos fazer sim. O po
1430435610204987396	2021-08-25	07:43:20	+0000 <@rogtavares> Então é isso o filho bananinha do pr fake esti
1430434357949382656	2021-08-25	07:38:21	+0000 <@Andrae16573935> @rubensvalente Estou apreensiva com eles dev
1430433833267208193	2021-08-25	07:36:16	+0000 <@Jorgebernucchi> @carlosaviana Os homens de Deus sabem o que v
1430433795346518017	2021-08-25	07:36:07	+0000 <@JakeMorceguinho> @UOLNoticias Bom que ele já dá um golpe agr
1430433105446416385	2021-08-25	07:33:23	+0000 <@allandsantos> Dia 7 de Setembro é o dia da INDEPENDÊNCIA BR
1430432942225084419	2021-08-25	07:32:44	+0000 <@Marcio_R_Cruz> @valter_morais 🇺🇵. Temos que encantar 7 de se
1430431710274064386	2021-08-25	07:27:50	+0000 <@dconline> Governo de SC não fará desfile de 7 de Setembro em
1430431433630359554	2021-08-25	07:26:44	+0000 <@MarcoTe38757191> Cadeia, confisco de bens e multas astroômic
1430431407541719047	2021-08-25	07:26:38	+0000 <@H_T_Junior> @ricograziano @aldorebello 7 de setembro será ind
1430431189408571395	2021-08-25	07:25:46	+0000 <@JP31239464> @nikolas_dm Nikolas você vai pro 7 de setembro o
1430429507886927879	2021-08-25	07:19:05	+0000 <@Jovane03667013> @jairbolsonaro bom dia Sr. Presidente, convi
1430429161169068038	2021-08-25	07:17:42	+0000 <@blogdoeri> EDITORIAL: O objetivo das manifestações de 7 de s
1430428251839766529	2021-08-25	07:14:05	+0000 <@RenatoKleberPr1> Sabem qual é a minha maior expectativa nest
1430427399347482624	2021-08-25	07:10:42	+0000 <@ROSADOS083383166> O PESSOAL DA ESQUERDA NÃO DEVE COMPARECER A
1430426286342684675	2021-08-25	07:06:17	+0000 <@EnsopaErli> MAIS UM: Outro coronel convoca para ato em 7 de
1430425137371881474	2021-08-25	07:01:43	+0000 <@Merson18706415> @aureliofinardi @lordivan22 7 de setembro na
1430423952065802244	2021-08-25	06:57:00	+0000 <@blogdoeri> Esta porra loquice de 7 de setembro será o velório
1430423839457091586	2021-08-25	06:56:33	+0000 <@AlicePetrucci4> @AguilarEdil Bom dia! Todos os dias o 7 de
1430423586234306561	2021-08-25	06:55:33	+0000 <@olivarefabi> Bolsonaro confirma presença em Brasília e na Pa
1430423441090433025	2021-08-25	06:54:58	+0000 <@luisduran> @Conservadora191 Ele quer mostrar que é bonzinho
1430422558394945540	2021-08-25	06:51:28	+0000 <@philos_em_gotas> @rosepaixao_on Espero que vc tenha razão e
1430422517525659649	2021-08-25	06:51:18	+0000 <@olivarefabi> Movimento Fora Bolsonaro contesta decisão de Do
1430421915303297024	2021-08-25	06:48:55	+0000 <@Isabela52864568> Opinião - Vinícius Torres Freire: Empresári
1430421581700927489	2021-08-25	06:47:35	+0000 <@olivarefabi> Vídeo: Tenente-coronel mostra que está longe de
1430421575719964675	2021-08-25	06:47:34	+0000 <@AguilarEdil> Bom dia Hoje é 7 de Setembro?
1430421275374149632	2021-08-25	06:46:22	+0000 <@rosepaixao_on> Peguem suas pipocas e assistam o desfile no s
1430420841108549633	2021-08-25	06:44:39	+0000 <@eugeniomsangior> Pelo 7 de Setembro, mais um coronel da PM e
1430419901567709184	2021-08-25	06:40:55	+0000 <@Patriot78989419> https://t.co/CQCZF5mPA0 Vai lá sabido 🇺🇵
1430419887520980992	2021-08-25	06:40:51	+0000 <@alvaro2018> @jairbolsonaro @jairbolsonaro só aceite conversa
143041955767620136	2021-08-25	06:39:33	+0000 <@CleversonJohn> @FairlyNight @STF_oficial uma vergonha, @TSEJu
1430419275597160448	2021-08-25	06:38:25	+0000 <@Eraldo30655272> @ALLuSapelli 7 de setembro chegando e zema t

52s

Aprofundando

Existe um risco em seguir mantendo afastadas as redes sociais de ferramentas digitais de pesquisa e análise. Programar em Python ou R nos dá muitas opções fantásticas, mas não são as únicas. Vale a pena conhecer ferramentas como Tableau, Power BI ou mesmo os bons e velhos editores de planilhas.

Para lembrar como os notebooks de hj funcionam ficam alguns vídeos.



Power BI

Contatos

[Twitter](#) | [LinkedIn](#)
[Github](#) | [Youtube](#)

Raspagem de dados do Twitter usando Python © 2021 by Gustavo Sarti is licensed under CC BY 4.0.

To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

boa sorte ;)