Oficina de Coleta e análise de dados na Web

Leonardo Nascimento
Computational Social Scientist
Federal University of Bahia

15% Programador,
20% Químico,
25% Psicólogo,
40% Cientista Social

Clique nas imagens para acessar

Públicos refratados: grupos de extrema-direita brasileiros na plataforma Telegram

Surge em 2018 junto ao Programa de Pós-graduação em Ciências Sociais (PPGCS) e o Instituto de Ciência, Tecnologia e Inovação (ICTI/UFBA)

Reunir pesquisadores interessados em métodos digitais/computacionais e ciências sociais

Pesquisadores

Eric Brasil
Digital Historian
Unilab

Letícia Cesarino
Digital anthropologist
Federal University of Santa Catarina

Rosana Moore
Pós Doc LABHDUFBA

Paulo Fonseca
Digital sociologist
Federal University of Bahia

Pedro Moraes
Data engineer
Ibotirama Sistemas

Tarssio Barreto
Data scientist
BIT::Analytics

Equipe

Anna Carollyne dos Santos Vieira (UFBA) - Bacharelado Interdisciplinar em Humanidades
Beatriz Leal Fraga (UFBA) - Bacharelado em Ciências Sociais
Daniel de Sena Bastos (UFBA) - Bacharelado Interdisciplinar em CTI
Emily Caroline de Vasconcelos (USP) - Bacharelado em Ciências Sociais
Ingrid da Silva Barreiros (UFBA) - Bacharelado Interdisciplinar em CTI
Jéssica Santana de Oliveira - Bacharelado Interdisciplinar em CTI
Juciane Pereira de Jesus(UFBA) - Mestrado em ciências sociais
Maria Luiza Scheren (UFSC) - Bacharel em Antropologia
Louise Lima Karczeski (UFSC) - Mestre em Antropologia Social
Matheus da Costa Martins - Bacharelado em Ciências Sociais
Pedro Mores - Engenharia de Dados
Rosana Moore - Doutora em Sociologia - Bolsista Pós-doc CNPQ
Sofia Schurig - Comunicação Social (FACOM/UFBA)
Tarssio Barreto - Cientista de Dados
Thamirys Albuquerque Cunha (UFBA) - Bacharelado em Ciência, Tecnologia e Inovação

Parcerias

FGV DAPP
DIRETORIA DE ANÁLISE DE POLÍTICAS PÚBLICAS

NOVA
UNIVERSIDADE NOVA DE LISBOA

Hub

INTERNETLAB

DEMOCRACIA DIGITAL

Análise dos ecossistemas de desinformação no Telegram durante o Processo Eleitoral Brasileiro de 2022

L4BHD UFB4

INTERNETLAB

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

PPGAS/UFSC

Objetivo geral do curso

*Entender o que é **mineração de dados nas ciências humanas e sociais; seu significado cultural e epistemológico**; e os limites e possibilidades da sua utilização para os diferentes campos das humanidades.*

Objetivo específicos

1 - Entender as etapas metodológicas de um processo de coleta de dados na web;

2 - Interpretar criticamente as fontes e as possibilidades e limites que elas oferecem;

3 - Aprender, a partir de uma coleta de dados do instagram, como estruturar seu primeiro projeto de coleta de dados na web.

Algumas definições...

“A mineração de dados é o estudo da coleta, limpeza, processamento, análise e obtenção de insights úteis a partir de dados.

(AGGARWAL, Charu C. Data Mining. Springer, 2015, p.1)

Wikipedia

processo de descoberta de padrões em grandes conjuntos de dados envolvendo métodos na interseção de:

- aprendizado de máquina,
- técnicas estatísticas, e
- sistemas de banco de dados.

O que estes objetos têm em comum?

Pedra

Argila

Papiro

Papel

Cilindro de cera

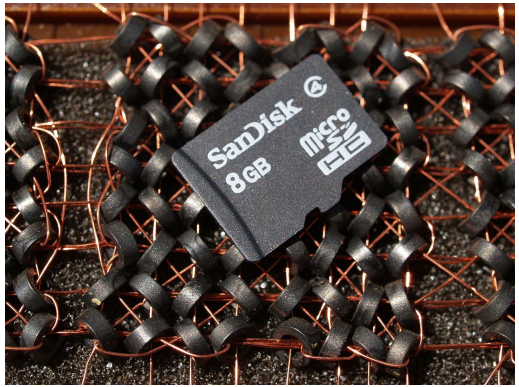
Fita Cassete

Disco de Vinil

https://en.wikipedia.org/wiki/Writing

(Resposta: eles são analógicos)

O que estes objetos têm em comum?



8GB (front) vs 8B (back)

Floppy disks (8”, 5 1/4”, 3 1/2”)

Compact disk

https://en.wikipedia.org/wiki/Writing

(Resposta: eles são digitais)

O que aconteceu?

...nova materialidade digital dos bits

(Brasil & Nascimento, 2020)

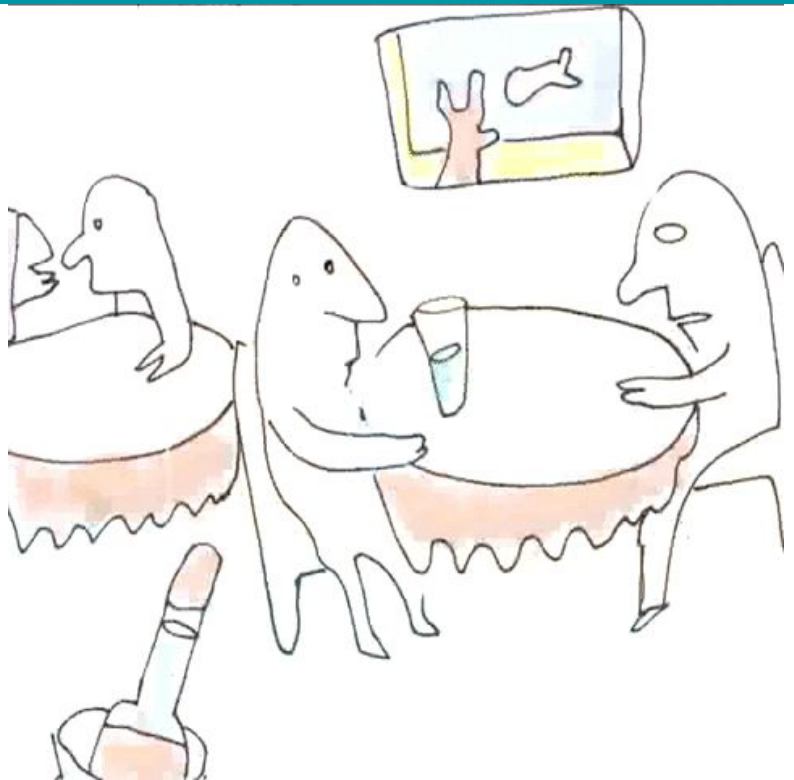
Folha de S. Paulo is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like Aldus PageMaker including versions of Lorem Ipsum

Rematerialização

Se concordarmos que, no “setor da informática, não há nada de virtual” (Vinck, 2016: 37) - pois os arquivos digitais “ocupam espaço” em servidores, cabos, antenas, hard disk drives etc. -, a desmaterialização não é senão uma rematerialização (Vinck, 2016: 36).

"Digitalização do eu na vida cotidiana"

(Nascimento, 2020)



Algoritmização de processos sociais

Traços digitais

(Howison et al. 2011, p. 769)

Dataficação (Cukier & Mayer-Schoenberger, 2013)

Era do “Big Data”

The **co-evolution** of **storage** capacity, **transmission** capacity, and **processing** capacity

Visualcapitalist.com (2021)

Big data são dados muitas vezes...

incompletos, inacessíveis, não representativos, flutuantes, algoritmicamente confusos, sujos e sensíveis - tudo isso vem do fato de que esses dados não foram coletados por pesquisadores para pesquisadores"

(Bit by Bit - Matthew J. Salganik, p.41)

Digital trace data

1. São dados encontrados (mesmo os extraídos), ao invés dados de produzidos para a pesquisa (ad hoc) através de instrumentos de pesquisa;

2. São dados relacionados/baseados em eventos ao invés de dados resumidos ou sintetizados;

3. Como os eventos ocorrem em um período de tempo, são dados dados longitudinais.

Howison et al. 2011, p. 769

Por que Minerar Dados na Web?

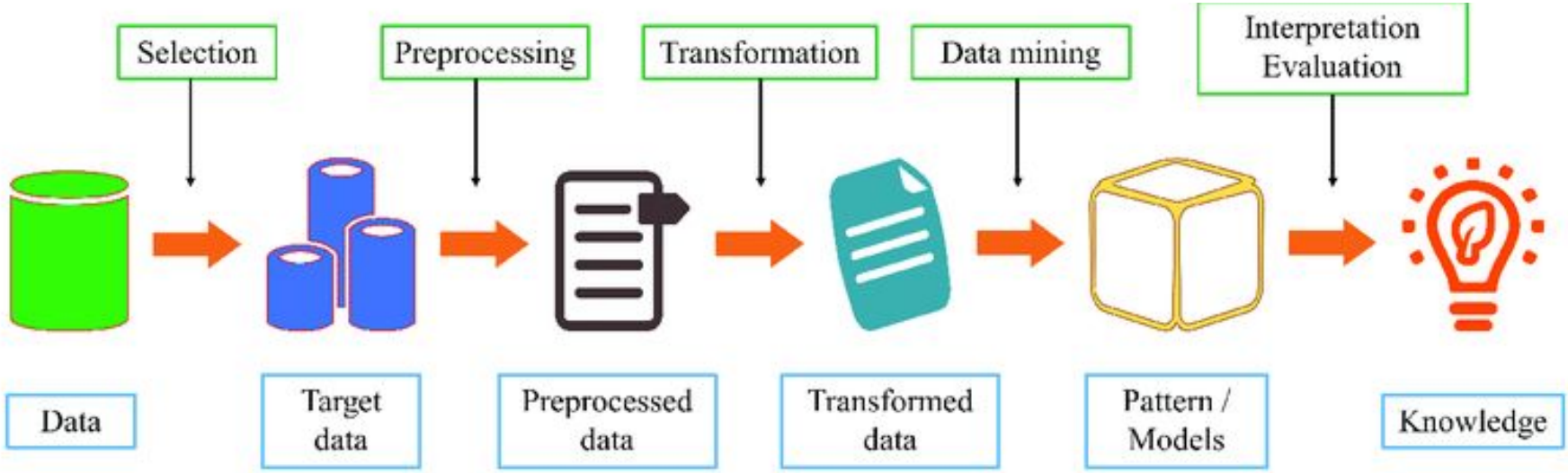
Na era analógica, a recolha de dados sobre comportamento (...) era cara e, portanto, relativamente rara. Agora, na era digital, os comportamentos de milhares de milhões de pessoas são registrados, armazenados e analisáveis.

(Bit by Bit - Matthew J. Salganik, p.13)

“a era digital cria novas oportunidades para pesquisa social”

(Bit by Bit - Matthew J. Salganik, p.2)

Etapas da Mineração de Dados na Web



Complexificando...

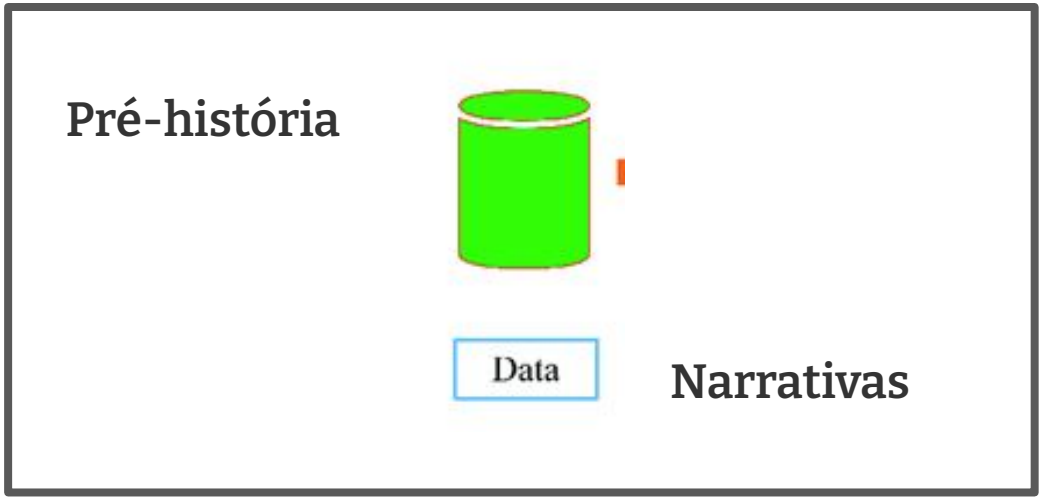
Todo e qualquer dado possui uma “pré-história”...

"Raw data is an oxymoron"

"Raw data is an oxymoron"

O que os dados estão tentando representar?

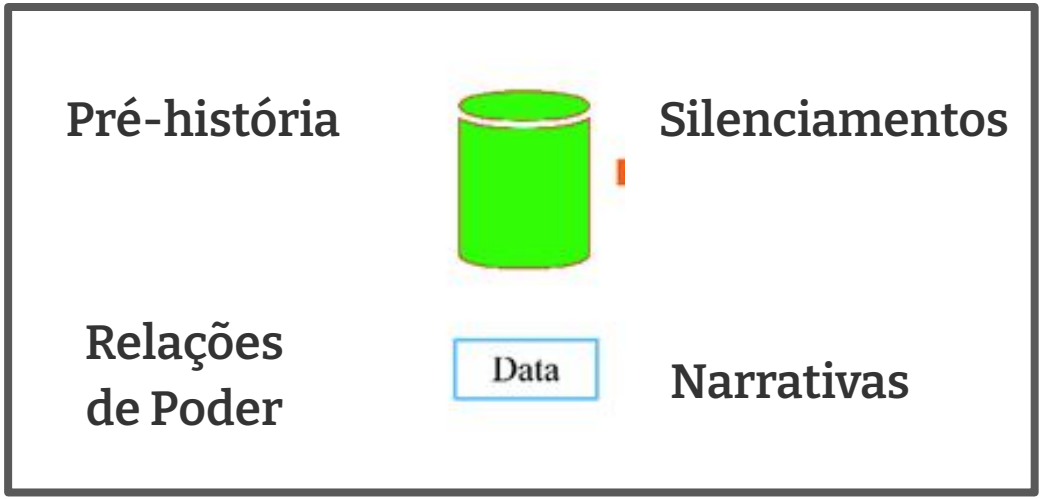
“Tente compreender as condições sob as quais o conjunto de dados surgiu. Sempre há uma história por trás dos dados. Sempre há algum tipo de narrativa - Smari McCarthy, Making Data Speak ”



Link: http://94.130.88.78/films/sm%C3%A1ri-mccarthy-making-data-speak)

O que os dados estão tentando representar?

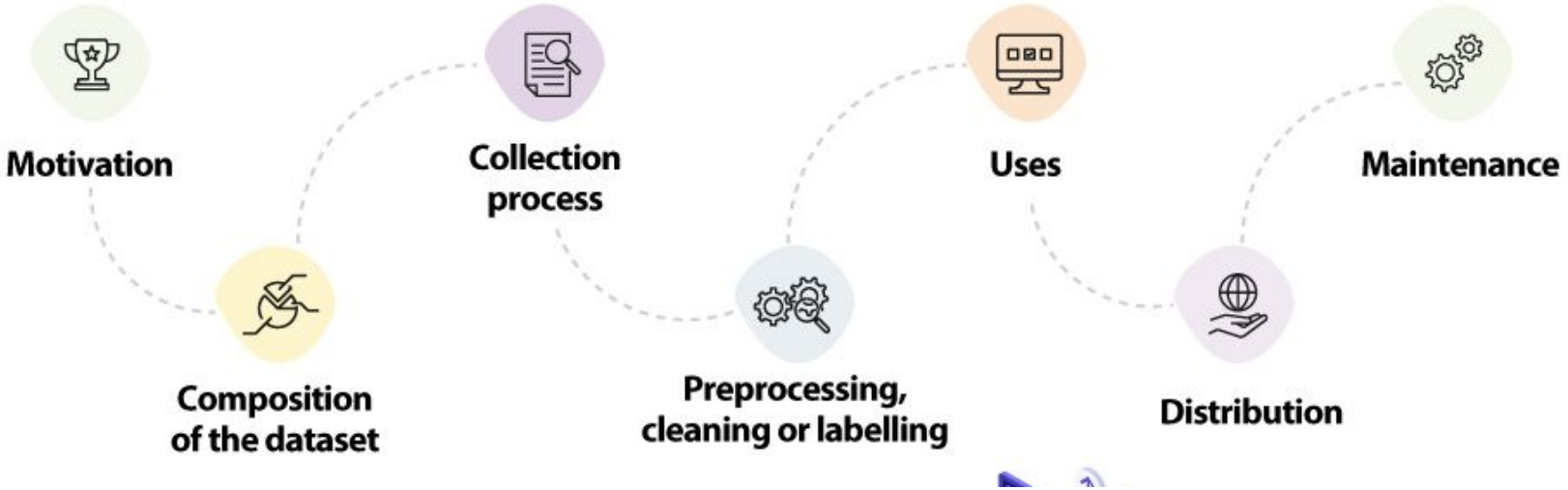
Tente obter o contexto, entender quem está gerando os dados, por que, quais podem ser os erros, que tipo de falhas podem haver, o que está sendo omitido e o que não está, o que está sendo explicitamente mantido preciso e assim por diante... - Smari McCarthy, Making Data Speak



Link: http://94.130.88.78/films/sm%C3%A1ri-mccarthy-making-data-speak)

DATASHEETS FOR DATASETS

Datasheets for datasets are sheets that assist in the documentation of data.
Each datasheet includes a series of questions related to the life cycle of the data:



https://arxiv.org/abs/1803.09010

https://vimeo.com/63958844

Datasheets for Datasets: motivação

Qual a motivação para o surgimento de um conjunto de dados?

Para que propósito o dataset foi criado?

Havia uma tarefa específica em mente?

Havia uma lacuna específica que precisava ser preenchida?

Quem criou ?

Quem financiou?

Datasheets for Datasets: composição

O que representam as instâncias que compõem o dataset (documentos, fotos, pessoas, texto, vídeos)?

Existem erros, fontes de ruído, ou redundâncias nos dados?

O dataset, se visto diretamente, pode ser ofensivo, insultante, ameaçador ou podem causar ansiedade?

Datasheets for Datasets: coleta

Como foram adquiridos os dados?

Que mecanismos ou procedimentos foram usados?

Quem coletou?

Quando coletou?

As pessoas cujos dados estão ali foram informadas da coleta?

Heurística Digital

(Brasil & Nascimento, 2020)

Análise crítica do conteúdo da fonte é fundamental no fazer historiográfico (Gil e Bresciano, 2015: 37)

É preciso investigar os metadados; investigar possíveis fraudes nos documentos; verificá-los por outras fontes; confirmar os conteúdos; avaliar a congruência; detectar erros, etc. (Chaudhuri, 2007 apud Gil e Bresciano, 2015: 38)

Relação online/offline...

“Digital Data Traces das interações com serviços on-line não permitem inferências sobre os reais interesses dos usuários, mas apenas sobre a fatia de seus interesses que querem que seja visto publicamente”

(JUNGHERR, 2015, p.44)

Ferramentas e Linguagens

Python (BeautifulSoup, Scrapy).

R/RStudio

Ferramentas específicas

SQL

API

Apps de Análise Qualitativa de Dados

Desafios e Considerações Éticas

Privacidade/consentimento

Limitações e bloqueios de sites

Respeitar (ou não) os termos de uso e robots.txt.

Desafios e Considerações Éticas

Respeito pelas Pessoas → tratar as pessoas como autônomas e respeitar seus desejos. (Bit by Bit - Matthew J. Salganik, 295)

Beneficência → compreender e melhorar o risco/benefício perfil do seu estudo e, em seguida, decidir se ele atinge o equilíbrio certo. (p. 296)

Justiça → garantir que os riscos e benefícios da investigação sejam distribuído de forma justa. (p. 298)

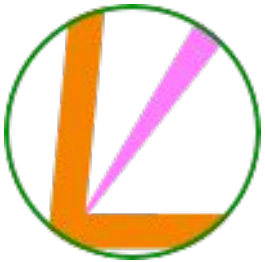
Respeito à Lei e ao Interesse Público → estende o princípio da Beneficência para além dos participantes específicos da investigação, para incluir todas as partes interessadas relevantes. (p. 299)

Minerando dados no Instagram

(clique aqui para o repositório do github)

A importância do Instagram

Coletando dados do Instagram



https://instaloader.github.io/

Versão parcialmente traduzida:

https://github.com/leofn/instaloader

Coletando dados do Instagram

1 - Python 3 instalado na máquina e no PATH (executável via prompt);

2 - Instalar o Instaloader;

3 - Fazer a coleta de um perfil escolhido;

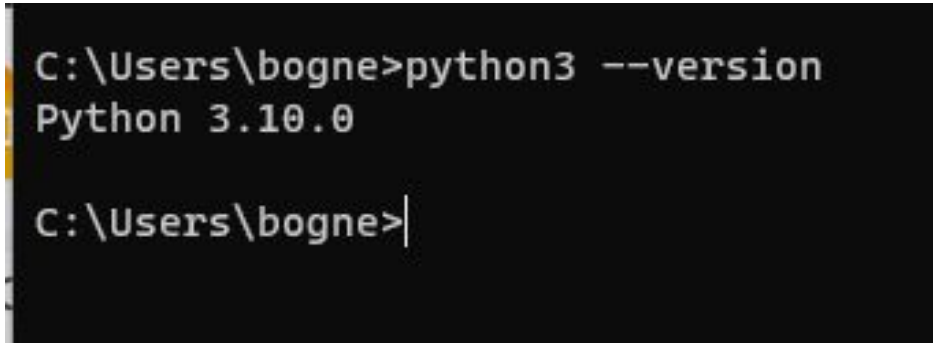
4 - Executar o ETL (Extract, Transform and Load);

5 - Gerar um dataset “tidy”

6 - Realizar análises descritivas simples e modelagem de tópicos

Coletando dados do Instagram

1 - Python 3 instalado na máquina e no PATH (executável via prompt);



Coletando dados do Instagram

2 - Instalar o Instaloader;

```
C:\Users\bogne>pip3 install instaloader
Requirement already satisfied: instaloader in c:\users\bogne\appdata\local\programs\python\python310\lib\site-packages (
4.9.5)
Requirement already satisfied: requests>=2.4 in c:\users\bogne\appdata\local\programs\python\python310\lib\site-packages
 (from instaloader) (2.28.2)
Requirement already satisfied: idna<4,>=2.5 in c:\users\bogne\appdata\local\programs\python\python310\lib\site-packages
(from requests>=2.4->instaloader) (3.3)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\bogne\appdata\local\programs\python\python310\lib\site-
packages (from requests>=2.4->instaloader) (1.26.7)
Requirement already satisfied: charset-normalizer<4,>=2 in c:\users\bogne\appdata\local\programs\python\python310\lib\si
te-packages (from requests>=2.4->instaloader) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\bogne\appdata\local\programs\python\python310\lib\site-pac
kages (from requests>=2.4->instaloader) (2021.10.8)
WARNING: You are using pip version 20.3.4; however, version 23.2.1 is available.
You should consider upgrading via the 'C:\Users\bogne\AppData\Local\Programs\Python\Python310\python.exe -m pip install
--upgrade pip' command.

C:\Users\bogne>
```

Coletando dados do Instagram

3 - Fazer a coleta de um perfil escolhido;

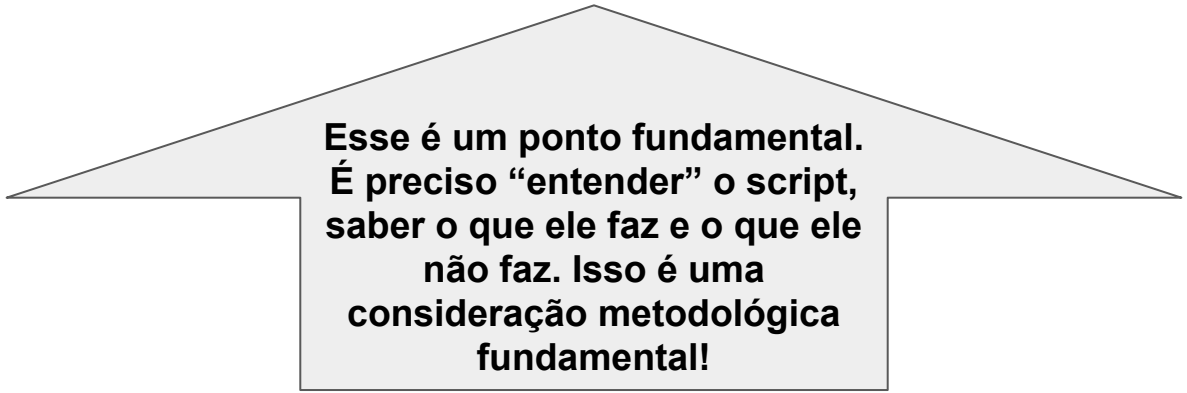
```
D:\instagram>instaloader cienciassociais_pucrio
```

Coletando dados do Instagram

3.1 - É possível usar diversos parâmetros adicionais

Documentação:

https://instaloader.github.io/



Coletando dados do Instagram

3.1 - Pegando todos os posts deste do começo do ano para cá

```
D:\instagram>instaloader --post-filter="date_utc >= datetime(202
3, 1, 1)" perfi1 perfil2 perfil 3
```

(instaloader --post-filter="date_utc >= datetime(2023, 1, 1)" perfi1 perfil2 perfil 3)

ETL

Extrair: recuperar dados brutos de um pool de dados não estruturados e os migra para um repositório de dados temporário;

Limpar: sanitizar os dados extraídos garantindo a qualidade dos dados antes da transformação;

Transformar: estrutura e converte os dados para formatos adequados para determinada finalidade;

Carregar: colocar os dados estruturados em um data warehouse para que possam ser analisados e usados adequadamente;

Analisar: processar os dados para obter insights científicos, comerciais, etc.

Coletando dados do Instagram

4 - Executar o ETL (Extract, Transform and Load) + 5 - Gerar um dataset "tidy"

banco	username	data	likes	comentarios	texto	taggeadoUsername	taggeadoNome	local
2020-08-26_18-59-41_UTC.json.xz	NA	2020-08-26 18:59:41	9	0	Nova página do instagram do Departamento de Ciências So...	NA	NA	NA
2020-08-26_19-04-09_UTC.json.xz	NA	2020-08-26 19:04:09	6	0	Amanhã, quinta-feira (dia 27), vai rolar o evento 'Racismo, ...	NA	NA	NA
2020-08-30_13-27-07_UTC.json.xz	NA	2020-08-30 13:27:07	11	0	Os diálogos interdisciplinares sobre cidades, organizados p...	NA	NA	NA
2020-09-01_18-29-31_UTC.json.xz	NA	2020-09-01 18:29:31	4	0	Estão abertas as inscrições para o PUC HACK, o Hackathon ...	NA	NA	NA
2020-09-03_15-22-52_UTC.json.xz	NA	2020-09-03 15:22:52	15	0	Como chegamos até aqui? Dia 16 de setembro (quarta-feira...	NA	NA	NA
2020-09-04_22-05-16_UTC.json.xz	NA	2020-09-04 22:05:16	8	0	O Programa de Pós-Graduação em Ciências Sociais (PPGCIS,...	NA	NA	NA
2020-09-20_20-25-55_UTC.json.xz	NA	2020-09-20 20:25:55	22	0	Mais um diálogo interdisciplinar sobre cidades, organizados...	NA	NA	NA
2020-09-23_21-06-55_UTC.json.xz	NA	2020-09-23 21:06:55	12	0	O cenário de eleições municipais, combinado com os graves...	NA	NA	NA
2020-09-29_23-32-16_UTC.json.xz	NA	2020-09-29 23:32:16	17	0	CICLO DE DEBATES – ESCOLA MUNICIPAL, ELEIÇÕES E PAN...	NA	NA	NA
2020-10-02_17-23-33_UTC.json.xz	NA	2020-10-02 17:23:33	12	0	O Programa de Pós-Graduação em Ciências Sociais da PUC-...	NA	NA	NA
2020-10-05_18-08-21_UTC.json.xz	NA	2020-10-05 18:08:21	22	0	Hoje começa o III Seminário 'Cidades, Territórios e Direitos'. ...	NA	NA	NA
2020-10-06_23-33-51_UTC.json.xz	NA	2020-10-06 23:33:51	12	0	CICLO DE DEBATES – ESCOLA MUNICIPAL, ELEIÇÕES E PAN...	NA	NA	NA
2020-10-12_21-53-24_UTC.json.xz	NA	2020-10-12 21:53:24	18	1	Como resistir em tempos de fragilização social? O corpo dis...	NA	NA	NA
2020-10-17_13-58-46_UTC.json.xz	NA	2020-10-17 13:58:46	18	0	O Departamento de Ciências Sociais e o Programa de Pós-g...	NA	NA	NA
2020-10-23_00-52-31_UTC.json.xz	NA	2020-10-23 00:52:31	15	0	O cenário de eleições municipais, combinado com os graves...	NA	NA	NA
2020-10-24_20-56-12_UTC.json.xz	NA	2020-10-24 20:56:12	12	0	A Rede Fluminense de Pesquisas sobre Violência, Segurança...	NA	NA	NA
2020-10-26_00-01-37_UTC.json.xz	NA	2020-10-26 00:01:37	29	0	O CENTRAL (Núcleo de Estudos e Projetos de Cidades) e o ...	NA	NA	NA
2020-10-28_23-46-38_UTC.json.xz	NA	2020-10-28 23:46:38	10	0	Nessa sexta-feira, dia 30, tem mais um evento do Seminário...	NA	NA	NA
2020-11-09_22-30-58_UTC.json.xz	NA	2020-11-09 22:30:58	15	1	PESQUISA PERFIL DO ESTUDANTE DE CIÊNCIAS SOCIAIS DA...	NA	NA	NA
2020-11-14_01-40-21_UTC.json.xz	NA	2020-11-14 01:40:21	56	6	O Departamento de Ciências Sociais apresenta o I Seminári...	NA	NA	NA
2020-11-17_16-16-54_UTC.json.xz	NA	2020-11-17 16:16:54	15	0	O Programa de Pós-Graduação em Ciências Sociais (PPGCIS,...	NA	NA	NA
2020-11-21_18-14-37_UTC.json.xz	NA	2020-11-21 18:14:37	20	0	O Departamento de Ciências Sociais da PUC -Rio vem a púb...	NA	NA	NA
2020-11-23_13-05-39_UTC.json.xz	NA	2020-11-23 13:05:39	12	0	O Programa de Pós-Graduação em Ciências Sociais (PPGCIS,...	NA	NA	NA
2020-12-06_22-53-19_UTC.json.xz	NA	2020-12-06 22:53:19	23	0	O Programa de Pós-Graduação em Ciências Sociais (PPGCIS,...	NA	NA	NA
2020-12-09_22-28-35_UTC.json.xz	NA	2020-12-09 22:28:35	19	0	O diálogo inter-religioso, as elaborações e produções criativ...	NA	NA	NA

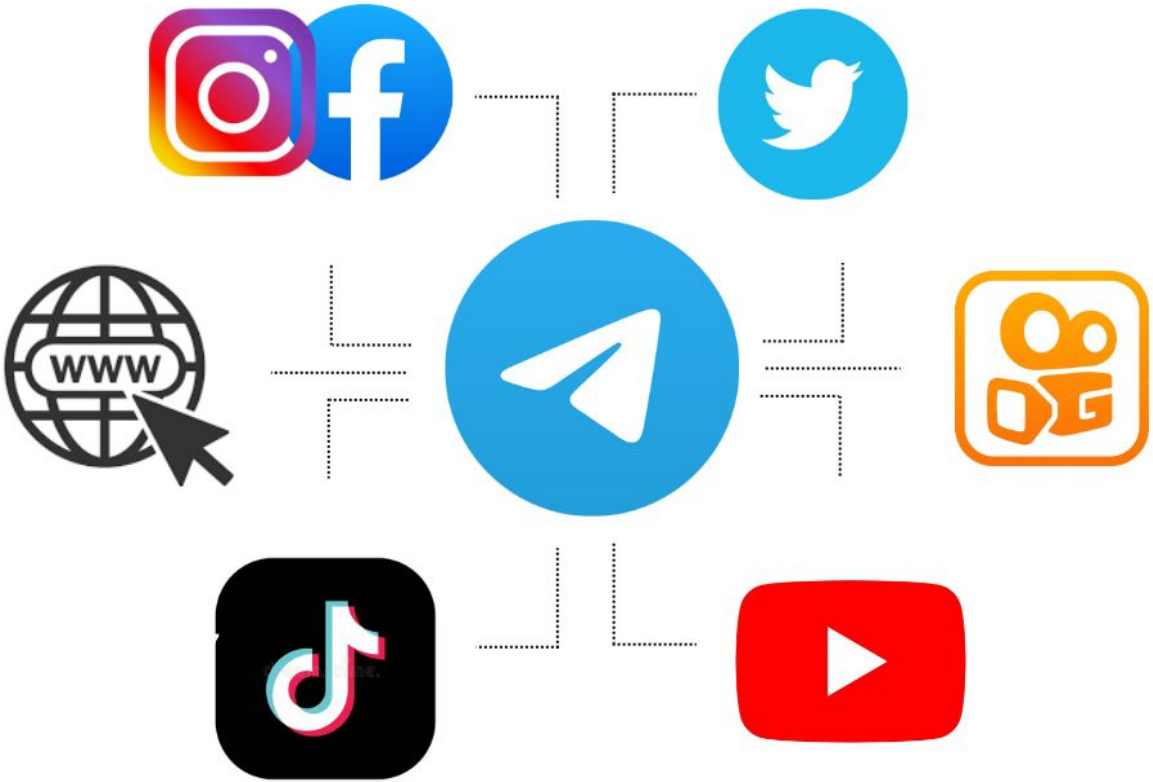
ETL enquanto metodologia

What things are worth counting?

Que coisas valem a pena contar?

Gramática dos dados ←→ Gramática das plataformas

Ecossistema multiplataforma



Perguntas?

https://github.com/LABHDUFBA

https://www.youtube.com/c/LABHDUFBA

@labhdufba

@labhdufb

Referências Bibliográficas

ALVES, Paulo César Borges; NASCIMENTO, Leonardo Fernandes. **Novas fronteiras metodológicas nas ciências sociais**. Salvador: EDUFBA - Editora da Universidade Federal da Bahia, 2018.

HOWISON, James; WIGGINS, Andrea; CROWSTON, Kevin. Validity Issues in the Use of Social Network Analysis with Digital Trace Data. **Journal of the Association for Information Systems**, v. 12, n. 12, 29 dez. 2011. Disponível em: <http://aisel.aisnet.org/jais/vol12/iss12/2>.

JUNGHERR, Andreas. **Analyzing Political Communication with Digital Trace Data**. Springer, 2015.

NASCIMENTO, Leonardo F. **Sociologia digital.** Salvador: EDUFBA - Editora da Universidade Federal da Bahia, 2020. Disponível em: <http://repositorio.ufba.br/ri/handle/ri/32746>. Acesso em: 19 out. 2021.

SALGANIK, Matthew J. **Bit by bit.** Princeton: Princeton University Press, 2018.