Web Scraping

curso-r

Sobre a Curso-R

A empresa

curso-r

R6 consultoria e treinamentos

www.curso-r.com

Ministrantes

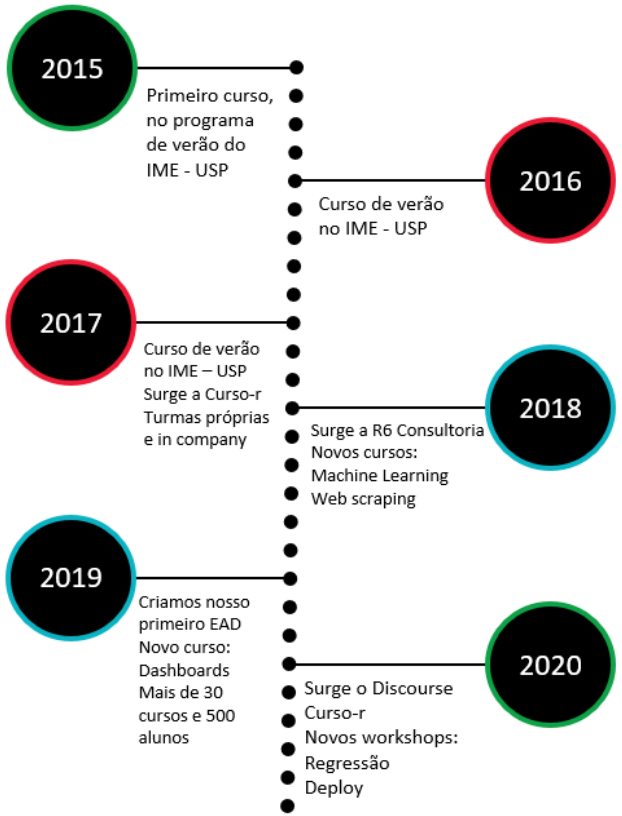
Julio Trecenti

Doutorando em Estatística pelo IME-USP. Secretário-geral da Associação Brasileira de Jurimetria (ABJ). Conselheiro do CONFE. Sócio da Terranova Consultoria. Trabalha com web scraping, arrumação de dados, construção de modelos preditivos, APIs, pacotes em R e dashboards em Shiny.

Caio Lente

Mestrando em Ciência da Computação no IME-USP e cientista de dados na Terranova Consultoria. Começou a se apaixonar pelo R em 2016 e agora não fala em outra coisa. Metido a designer, maníaco da organização e metade texano.

Linha do tempo



curso-r

Nossos cursos

Programação em R

R para Ciência de dados I

R para Ciência de dados II

Introdução ao R com C++

Extração de dados

Web scraping

Comunicação e automação

Dashboards com R

Deploy

Modelagem

Regressão Linear

Machine Learning

XGBoost

Deep Learning

curso-r

Sobre o curso

Web scraping

Introdução

- O que é e quando fazer web scraping
- O ciclo do web scraping
- Noções de protocolo HTTP
- Acessando dados de APIs

Web scraping

Web scraping de fato

- Como fazer requisições HTTP e baixar páginas web a partir do R
- Como estruturar dados de arquivos .xml, .html e .json
- Como iterar algoritmos no R

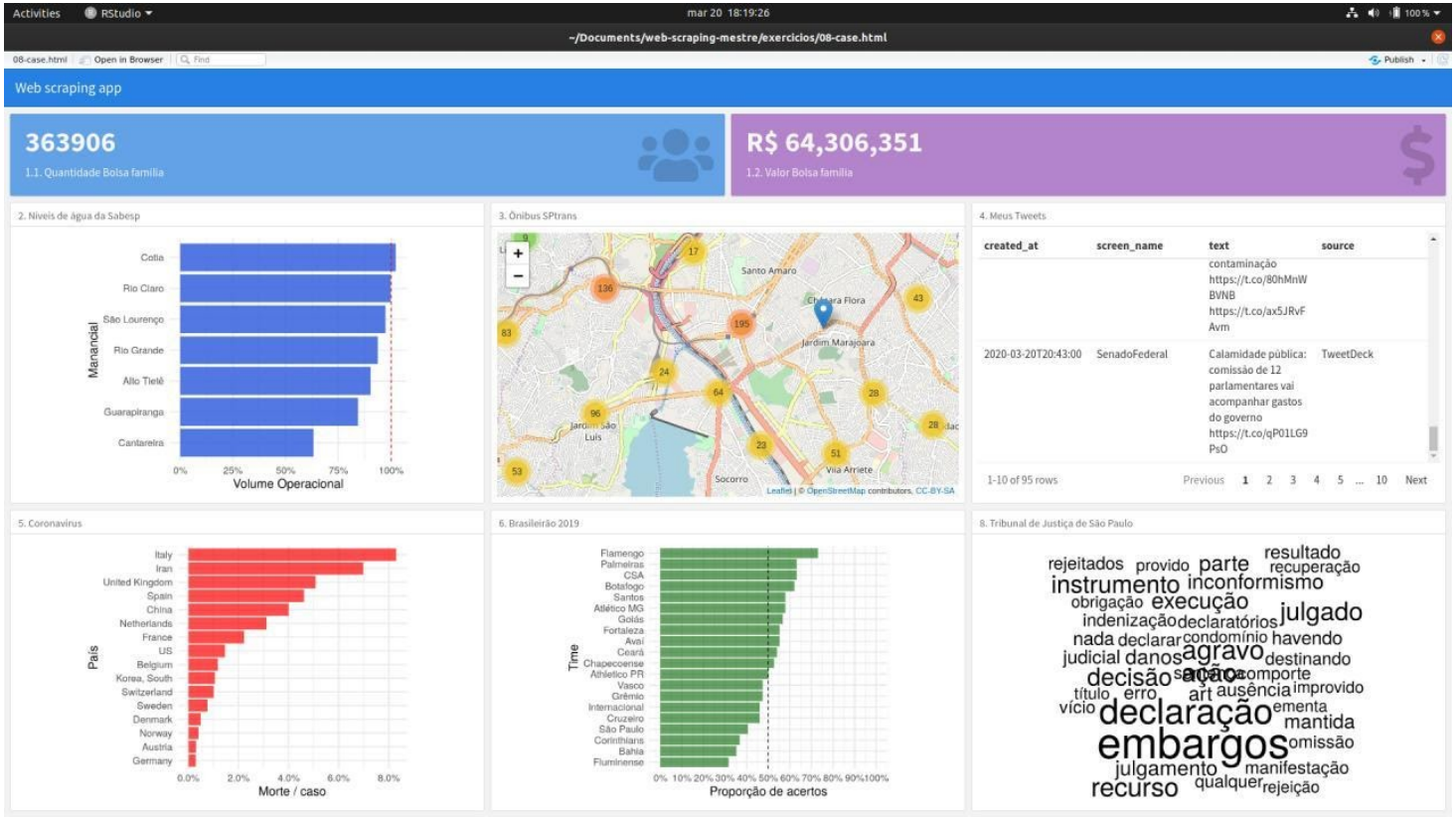
Web scraping

Tópicos

- Tratamento de erros
- Paralelização
- Páginas dinâmicas com Selenium

Resultados

No final, você terá dados suficientes para montar um dashboard como esse!



Dinâmica

- Pelo menos um exemplo prático por aula
    - Foco: APIs públicas, sites públicos
- Exercícios para casa, com entrega facultativa
- Trabalho final, com entrega obrigatória
    - As pessoas que fizerem os melhores trabalhos receberão uma bolsa para fazer qualquer curso da Curso-R
    - Mais detalhes sobre o trabalho final nas próximas aulas

Tirando dúvidas

- **Não existe dúvida idiota**.
- Nem sempre é trivial fazer a pergunta certa para que outra pessoa esclareça a sua dúvida. Neste curso, **vamos mostrar melhores práticas na hora de fazer perguntas sobre programação**.
- Fora do horário de aula ou monitoria:
    - perguntas gerais sobre o curso deverão ser feitas no Classroom.
    - perguntas sobre R, principalmente as que envolverem código, deverão ser enviadas no nosso discourse. Se envolver web scraping, é importante especificar a página que está querendo acessar e como você faria para encontrá-la manualmente.
- Veja aqui dicas de como fazer uma boa pergunta.

Por que usar o discourse?

- Muito melhor para escrever textos que possuem códigos. Com ele, podemos usar o pacote `{reprex}`!
- Saber pesquisar sobre erros e fazer a pergunta certa é essencial para aprender e resolver problemas de programação.
- No discourse, teremos mais pessoas acompanhando e respondendo as dúvidas.
- Em um ambiente aberto, as suas dúvidas vão contribuir com a comunidade.

https://discourse.curso-r.com/

Material

Para baixar o material,

- Baixe a pasta do curso atual (pelo arquivo zip ou clonando)
- Se estiver numa pasta compactada, descompacte o arquivo numa pasta
- Abra o projeto (arquivo `.Rproj`)
- Rode

```
if (!require(CursoR))
  remotes::install_github("curso-r/CursoR")

CursoR::atualizar_material()
```

Mais informações

https://curso-r.github.com/main-web-scraping