|  |  |
| --- | --- |
| Curso: | Jurimetria |
| Disciplina: | Raspagem (web scraping) de Tribunais de Justiça e Administrativos |

**Ementa**

Esta disciplina ensina como baixar dados processuais por meio da internet. Nela você aprenderá os principais conceitos de transferência de dados pela internet. Igualmente, verá como inspecionar uma página web e realizar requisições. Por fim, aprenderá como automatizar as buscas nas páginas de forma iterativa e como estruturar esses dados em tabelas.

**Conteúdo Programático**

**Unidade 1: Noções básicas de transferência de dados pela internet**

1 - Introdução à internet

2 - Internet vs WWW

3 - Inspeção de uma página web

3.1 inspector

3.2 Network e Headers

3.2.1 URL

3.2.2 Status

3.2.3 Tamanho

3.3.4 Parâmetros de requisição

3.3.5 Parâmetro de resposta

4 - As três etapas de web scraping: requisição, extração, iteração

**Unidade 2: Requisição**

1 - Protocolos de transmissão de dados

1.1 TCP/IP

1.2 DNS

1.2 Emails: smpt/imap/pop3

1.3 Transferência de arquivos: (s)ftp

1.4 Transferência de páginas: http(s)

2 - O protocolo http

2.1 API

2.2 Requisições

4.2.1 GET

4.2.2 POST

2.3 Tipos de dados

2.4 Entendendo cookies

2.5 Headers

Unidade 3: Extração

1. HTML e XML

1.1 Decifrando html

1.2 Manipulação do DOM: Document Object Model

2. Noções de CSS

3. Noções de Javascript

4. Xpath

3.1 Parseamento

3.2 Busca (find e find all)

3.3 Extração: text e atributos

3.4 Regex

5. Estruturação com dplyr e tibble

Unidade 4: Iteração

1. Criando funções de requisição

2. Iterando com purrr

3. Criando funções de parseamento

4. Dando nomes seguros

**Referências Bibliográficas**

DAMIANI, Athos at al. **Ciência de Dados em R**. [s.l.]: curso-r, 2022.

DE JESUS FILHO, Jose. **stj: Baixa e organiza decisões do Superior Tribunal de Justiça**. [s.l.: s.n.], 2022. Disponível em: <<https://github.com/jjesusfilho/stj>>.

JESUS FILHO, José de ; TRECENTI, Julio. **Coleta e organização de dados do Tribunal de Justiça de São Paulo**. [s.l.: s.n.], 2020. Disponível em: <<https://jjesufilho.github.io/tjsp/>>.

Wickham, Hadley, 2019. *Stringr: Simple, Consistent Wrappers for Common String Operations*. <https://CRAN.R-project.org/package=stringr>.

———. 2020. *Httr: Tools for Working with URLs and HTTP*. <https://CRAN.R-project.org/package=httr>.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. <https://CRAN.R-project.org/package=dplyr>.

Wickham, Hadley, Jim Hester, and Jeroen Ooms. 2020. *Xml2: Parse XML*. <https://CRAN.R-project.org/package=xml2>.

Wickham, Hadley, Jeroen Ooms, and Kirill Müller. 2021. *RPostgres: ’Rcpp’ Interface to ’PostgreSQL’*. <https://CRAN.R-project.org/package=RPostgres>