Análise, construção e visualização de dados

Prof. Dr. Álvaro Campos Ferreira

alvaro.ferreira@idp.edu.br

Webscrapping

Webscrapping

É legal ou ilegal fazer webscrapping?

- É possível acidentalmente realizar muitos requests e isso é visto como um ataque
- Seu IP pode ser banido para algum site ou serviço

idp

Webscrapping

É perfeitamente legal se feito de forma responsável!

- Robots.txt

  https://twitter.com/robots.txt

- Requests
- Delays

idp

Webscrapping

Existem dados em tabelas que podem ser utilizados diretamente pelo Pandas.

```
import pandas as pd
url = ‘https://pt.wikipedia.org/wiki/COVID-19’
html = pd.read_html(url)
```

Webscrapping

Para selecionar apenas a tabela que queremos, usamos o argumento match na função read_html().

```
import pandas as pd
url = ‘https://pt.wikipedia.org/wiki/COVID-19’
html = pd.read_html(url,match=’Frequência’)
```

Webscrapping

Como obter os dados da página que desejamos estudar?

```
import requests
res = requests.get('https://www.google.com/')
```

Webscrapping

Esses dados não estão estruturados. Para estrutura-los, usamos BeautifulSoup.

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(res.text,"lxml")
```

idp

Webscrapping

Dados obtidos da página estão na linguagem HTML, que divide os dados no que chama de tags.

Para explorar as tags de um site, usa-se a ferramenta de desenvolvimento do navegador, F12 no Firefox.

idp

Webscrapping

Uma vez identificadas as tags de interesse, filtramos o resultado:

```
soup.findAll('div')
for tag in soup.findAll('div'):
    print(tag.text)
```

INSTITUTO BRASILEIRO DE ENSINO,
DESENVOLVIMENTO E PESQUISA