5.Crawling

October 25, 2019

Para entrar no modo apresentação, execute a seguinte célula e pressione -

[1]:
```
%reload_ext slide
```

```
<IPython.core.display.Javascript object>
```

1 Crawling

Este notebook apresenta os seguintes tópicos:

- Section 1 - Crawling
- Section 1.1 - Exercício 5

Nesta seção, faremos requisições da página inicial de um repositório no GitHub e tentaremos extrair informações dela.

Lembrar de iniciar o servidor de proxy:

`python proxy.py`

A página usada é a mesma que usamos para mostrar requisição com requests. Ou seja, podemos usar o mesmo código para fazer a requisição.

[2]:
```
import requests
SITE = "http://localhost:5000/" # Se não usar o proxy, alterar para https://
 ↪github.com/

response = requests.get(SITE + "gems-uff/sapos")
response.status_code
```

[2]:
```
200
```

O conteúdo do HTML pode ser obtido pelo atributo `response.text`.

[3]:
```
response.text[:100]
```

[3]:
```
'\n\n\n\n\n\n<!DOCTYPE html>\n<html lang="en">\n  <head>\n    <meta
charset="utf-8">\n  <link rel="dns-prefetch'
```

Para extrair informações do HTML, podemos usar a biblioteca `BeautifulSoup`.

[4]:
```
from bs4 import BeautifulSoup
soup = BeautifulSoup(response.text, 'html.parser')
```

Essa célula parseou HTML para o objeto `soup`, que nos permite invocar métodos para buscar elementos do DOM

Para descobrirmos o que buscar, podemos usar a função de "Inspecionar elemento" do navegador e observar `id`, `class` e elementos que queremos.

Usando o elemento e a classe, podemos usar um seletor do BeautifulSoup para obter o número de estrelas.

[5]:
```
soup.select("a.social-count")
```

[5]:
```
[<a aria-label="6 users are watching this repository" class="social-count"
href="/gems-uff/sapos/watchers">
       6
     </a>,
 <a aria-label="18 users starred this repository" class="social-count js-social-
count" href="/gems-uff/sapos/stargazers">
       18
     </a>,
 <a aria-label="11 users forked this repository" class="social-count"
href="/gems-uff/sapos/network/members">
       11
     </a>]
```

O seletor usado trouxe mais elementos do que gostaríamos. Precisamos filtrar ainda mais. Nesse caso, podemos filtrar pela classe `.js-social-count` ou pelo `href`.

Pela classe:

[6]:
```
soup.select("a.social-count.js-social-count")
```

[6]:
```
[<a aria-label="18 users starred this repository" class="social-count js-social-
count" href="/gems-uff/sapos/stargazers">
       18
     </a>]
```

[7]:
```
_[0].text.strip() + " estrelas"
```

[7]:
```
'18 estrelas'
```

Pelo href terminado em `stargazers`:

[8]:
```
soup.select('a.social-count[href$="stargazers"]')
```

[8]:
```
[<a aria-label="18 users starred this repository" class="social-count js-social-
count" href="/gems-uff/sapos/stargazers">
       18
     </a>]
```

[9]:
```
_[0].text.strip() + " estrelas"
```

[9]:
```
'18 estrelas'
```

Usando **href**, também podemos obter watchers e forks:

[10]:
```
soup.select('a.social-count[href$="watchers"]')[0].text.strip() + " watchers"
```

[10]:
```
'6 watchers'
```

[11]:
```
soup.select('a.social-count[href$="members"]')[0].text.strip() + " forks"
```

[11]:
```
'11 forks'
```

1.1 Exercicio 5

Obtenha a lista de arquivos e diretórios da raiz do repositório com seus respectivos commits.

[12]:
```
...
```

[12]:
```
{'app': '83bb2570fabf704062c34e091a66ea73aba7755c',
 'bin': '099f05a532a543805a5bd430777ff5eb95a8d0de',
 'config': 'e7a53305df24d15bf3443c129d3b3ebb497c216f',
 'db': 'ac1ead04e291927e6d348d2c719d54e045446139',
 'doc': '293dfc5fce25e1f9074cd21ea14008e0c223cea8',
 'lib': 'b0cffa9fe39ab27c546bfd5e5fdfcfd78d1c909e',
 'public': '099f05a532a543805a5bd430777ff5eb95a8d0de',
 'script': '7aed384a68f2e9c67dc00df6a0ab3d97670afad6',
 'spec': 'e7a53305df24d15bf3443c129d3b3ebb497c216f',
 'test': '9254d7e524e829170805f0325ae0459dd6d1f979',
 'wiki': '455816cf36fe468f7fb54142d4e2e870d91df3b1',
 '.gitignore': 'b7fc89c74e5bb5fa977c9b364a1cc33da162cfb6',
 '.mailmap': '5fcd366a8289fdf937d839f292d2294374983c8a',
 '.rspec': '8143c1aba4d3a06f83f09c0c1b9ec778e9ee9fd2',
 '.travis.yml': '51561548b3f192a93e0456ba41a3f778e0e5def0',
 'Gemfile': 'd3e4307771d6e7722cb1bec5c967519e252c00d3',
 'Gemfile.lock': 'd3e4307771d6e7722cb1bec5c967519e252c00d3',
 'LICENSE': '57b9e09574da97ee43a87a447babcc2a99bdd750',
 'README.md': '997544ed41ee5c06d722760ebbb61c997343a583',
 'Rakefile': '099f05a532a543805a5bd430777ff5eb95a8d0de',
 'config.ru': '099f05a532a543805a5bd430777ff5eb95a8d0de'}
```

Continua: 6.API.v3.pdf