5.Crawling

October 27, 2019

1 Crawling

Este notebook apresenta os seguintes tópicos:

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.

[1]:
```
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
```

[1]:
```
200
```

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

[2]:
```
response.text[:100]
```

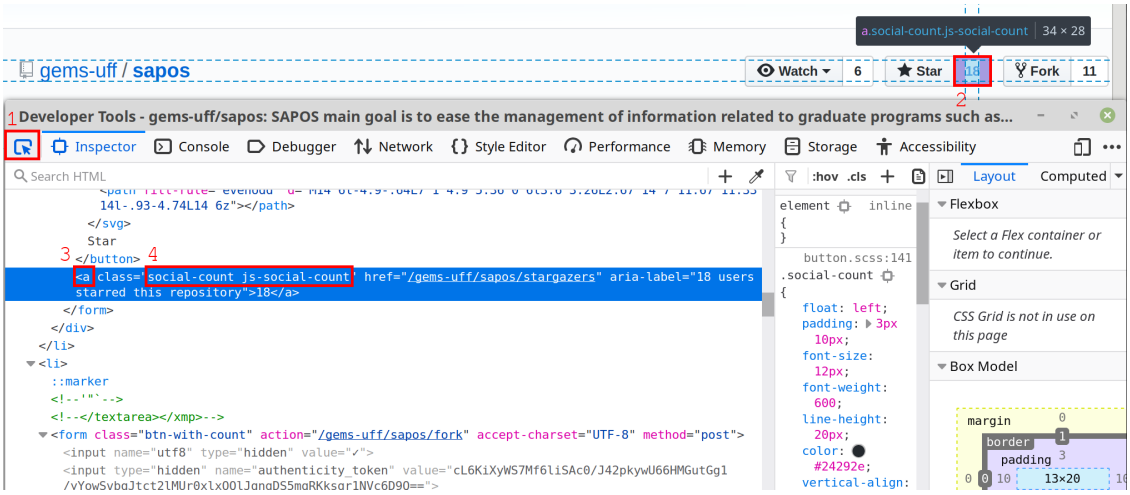
[2]:
```
'\n\n\n\n\n\n<!DOCTYPE html>\n<html lang="en">\n  <head>\n    <meta
charset="utf-8">\n  <link rel="dns-prefetch'
```

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

[3]:
```
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.

[4]:
```
soup.select("a.social-count")
```

[4]:
```
[<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:

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

[5]:
```
[<a aria-label="18 users starred this repository" class="social-count js-social-
count" href="/gems-uff/sapos/stargazers">
       18
     </a>]
```

[6]:
```
_[0].text.strip() + " estrelas"
```

[6]:
```
'18 estrelas'
```

Pelo href terminado em `stargazers`:

[7]:
```
soup.select('a.social-count[href$="stargazers"]')
```

[7]:
```
[<a aria-label="18 users starred this repository" class="social-count js-social-
count" href="/gems-uff/sapos/stargazers">
        18
      </a>]
```

[8]:
```
_[0].text.strip() + " estrelas"
```

[8]:
```
'18 estrelas'
```

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

[9]:
```
soup.select('a.social-count[href$="watchers"]')[0].text.strip() + " watchers"
```

[9]:
```
'6 watchers'
```

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

[10]:
```
'11 forks'
```

1.1 Exercicio 5

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

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

Continua: 6.API.v3.pdf