Untitled

July 8, 2020

[102]:
```
!pip install nb_black -q
!pip install pdfminer.six -q
!pip install urllib3 -q
!pip install bs4 -q
```

[103]:
```
%load_ext lab_black
```


0.0.1 Aluno Marco Antonio Moreira Carujo

1 Atividade 1: Coleta de Documentos em PDF

Atividade: Baixar arquivos PDFs da Web e armazená-los como arquivo de texto.

Site: 2019 - Anais do IV Congresso sobre Tecnologias na Educação (https://sol.sbc.org.br/index.php/ctrle)

[118]:
```
from urllib.request import urlopen
from bs4 import BeautifulSoup
from joblib import Parallel, delayed
```

Dentro do site temos dois links um comprido onde mostra apenas os ultimos artigos e um completo onde temos todos os artigos (75 artigos) do ano de 2019.

Podemos baixar ele e ir na tag **<a>** buscar o href de cada um. Dentro do href apenas mudando a parametro/rota view para download podemos baixar cada um dos PDF's.

A tag a possuí uma classe chama **obj_galley_link pdf** o que facilita o encontro da mesma.

[110]:
```
response = urlopen("https://sol.sbc.org.br/index.php/ctrle")  # Comprimido
response = urlopen("https://sol.sbc.org.br/index.php/ctrle/issue/view/486")  #␣
↪Completo
html = response.read()

soup = BeautifulSoup(html, "html.parser")
elements = soup.find_all("a", class_="obj_galley_link pdf")
print("Quantos artigos temos no site?", len(elements), "artigos.")
```

```
Quantos artigos temos no site? 75 artigos.
```

Apenas manipulações de objetos e tranformações de Strings para gerar as seguintes listas: - lista de links de view - lista de links de download - lista de links de nomes de arquivos para ser tanto .pdf quanto .txt

[111]:
```
links_to_view = [el["href"] for el in elements]
links_to_view[:5]
```

[111]:
```
['https://sol.sbc.org.br/index.php/ctrle/article/view/8870/8771',
 'https://sol.sbc.org.br/index.php/ctrle/article/view/8871/8772',
 'https://sol.sbc.org.br/index.php/ctrle/article/view/8872/8773',
 'https://sol.sbc.org.br/index.php/ctrle/article/view/8873/8774',
 'https://sol.sbc.org.br/index.php/ctrle/article/view/8874/8775']
```

[112]:
```
links_to_download = [el.replace("view", "download") for el in links_to_view]
links_to_download[:5]
```

[112]:
```
['https://sol.sbc.org.br/index.php/ctrle/article/download/8870/8771',
 'https://sol.sbc.org.br/index.php/ctrle/article/download/8871/8772',
 'https://sol.sbc.org.br/index.php/ctrle/article/download/8872/8773',
 'https://sol.sbc.org.br/index.php/ctrle/article/download/8873/8774',
 'https://sol.sbc.org.br/index.php/ctrle/article/download/8874/8775']
```

[113]:
```
names_to_download = [el[-9:].replace("/", "-") for el in links_to_download]
names_to_download[:5]
```

[113]:
```
['8870-8771', '8871-8772', '8872-8773', '8873-8774', '8874-8775']
```

As função **download** e **transforma em texto** foram retiradas/inspiradas no material da aula.

A função worker tem como objetivo encapsular a tarefa de baixar e transformar para ser paraleliada.

[117]:
```
import requests
from pdfminer.high_level import extract_text
import os


def download(url, nome):
    resposta = requests.get(url)
    if resposta.status_code == 200:
        with open(os.path.join("pdf", nome + ".pdf"), "wb") as f:
            f.write(resposta.content)


def transforma_em_texto(nome):
    pdf_text = extract_text(os.path.join("pdf", nome + ".pdf"), codec="utf-8")
    with open(os.path.join("text", nome + ".txt"), "w", encoding="utf-8") as f:
        f.write(pdf_text)
```

```
def worker(i):
    link = links_to_download[i]
    name = names_to_download[i]
    download(link, name)
    transforma_em_texto(name)
    return f'Arquivo {name} pdf e txt gerado e salvo!'
```

Execução da celula seguir irá baixar paralelamente 75 artigos e transformar em texto. Os PDFs serão salvos no diretorio *pdf* e os arquivos txt no diretorio *txt*.

[121]:
```
%%time
os.makedirs("pdf", exist_ok=True)
os.makedirs("text", exist_ok=True)

Parallel(verbose=1, n_jobs=-1)(delayed(worker)(i) for i in␣
 ↪range(len(names_to_download)))
```

```
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 12 concurrent workers.
[Parallel(n_jobs=-1)]: Done  26 tasks      | elapsed:    8.2s

Wall time: 19.2 s

[Parallel(n_jobs=-1)]: Done  75 out of  75 | elapsed:   19.1s finished
```

[121]:
```
['Arquivo 8870-8771 pdf e txt gerado e salvo!',
 'Arquivo 8871-8772 pdf e txt gerado e salvo!',
 'Arquivo 8872-8773 pdf e txt gerado e salvo!',
 'Arquivo 8873-8774 pdf e txt gerado e salvo!',
 'Arquivo 8874-8775 pdf e txt gerado e salvo!',
 'Arquivo 8875-8776 pdf e txt gerado e salvo!',
 'Arquivo 8876-8777 pdf e txt gerado e salvo!',
 'Arquivo 8877-8778 pdf e txt gerado e salvo!',
 'Arquivo 8878-8779 pdf e txt gerado e salvo!',
 'Arquivo 8879-8780 pdf e txt gerado e salvo!',
 'Arquivo 8880-8781 pdf e txt gerado e salvo!',
 'Arquivo 8881-8782 pdf e txt gerado e salvo!',
 'Arquivo 8882-8783 pdf e txt gerado e salvo!',
 'Arquivo 8883-8784 pdf e txt gerado e salvo!',
 'Arquivo 8884-8785 pdf e txt gerado e salvo!',
 'Arquivo 8885-8786 pdf e txt gerado e salvo!',
 'Arquivo 8886-8787 pdf e txt gerado e salvo!',
 'Arquivo 8887-8788 pdf e txt gerado e salvo!',
 'Arquivo 8888-8789 pdf e txt gerado e salvo!',
 'Arquivo 8889-8790 pdf e txt gerado e salvo!',
 'Arquivo 8890-8791 pdf e txt gerado e salvo!',
 'Arquivo 8891-8792 pdf e txt gerado e salvo!',
 'Arquivo 8892-8793 pdf e txt gerado e salvo!',
 'Arquivo 8893-8794 pdf e txt gerado e salvo!',
```

```
'Arquivo 8894-8795 pdf e txt gerado e salvo!',
'Arquivo 8895-8796 pdf e txt gerado e salvo!',
'Arquivo 8896-8797 pdf e txt gerado e salvo!',
'Arquivo 8897-8798 pdf e txt gerado e salvo!',
'Arquivo 8898-8799 pdf e txt gerado e salvo!',
'Arquivo 8899-8800 pdf e txt gerado e salvo!',
'Arquivo 8900-8801 pdf e txt gerado e salvo!',
'Arquivo 8901-8802 pdf e txt gerado e salvo!',
'Arquivo 8902-8803 pdf e txt gerado e salvo!',
'Arquivo 8903-8804 pdf e txt gerado e salvo!',
'Arquivo 8904-8805 pdf e txt gerado e salvo!',
'Arquivo 8905-8806 pdf e txt gerado e salvo!',
'Arquivo 8906-8807 pdf e txt gerado e salvo!',
'Arquivo 8907-8808 pdf e txt gerado e salvo!',
'Arquivo 8908-8809 pdf e txt gerado e salvo!',
'Arquivo 8909-8810 pdf e txt gerado e salvo!',
'Arquivo 8910-8811 pdf e txt gerado e salvo!',
'Arquivo 8911-8812 pdf e txt gerado e salvo!',
'Arquivo 8912-8813 pdf e txt gerado e salvo!',
'Arquivo 8913-8814 pdf e txt gerado e salvo!',
'Arquivo 8914-8815 pdf e txt gerado e salvo!',
'Arquivo 8915-8816 pdf e txt gerado e salvo!',
'Arquivo 8916-8817 pdf e txt gerado e salvo!',
'Arquivo 8917-8818 pdf e txt gerado e salvo!',
'Arquivo 8918-8819 pdf e txt gerado e salvo!',
'Arquivo 8919-8820 pdf e txt gerado e salvo!',
'Arquivo 8920-8821 pdf e txt gerado e salvo!',
'Arquivo 8921-8822 pdf e txt gerado e salvo!',
'Arquivo 8922-8823 pdf e txt gerado e salvo!',
'Arquivo 8923-8824 pdf e txt gerado e salvo!',
'Arquivo 8924-8825 pdf e txt gerado e salvo!',
'Arquivo 8925-8826 pdf e txt gerado e salvo!',
'Arquivo 8926-8827 pdf e txt gerado e salvo!',
'Arquivo 8927-8828 pdf e txt gerado e salvo!',
'Arquivo 8928-8829 pdf e txt gerado e salvo!',
'Arquivo 8929-8830 pdf e txt gerado e salvo!',
'Arquivo 8930-8831 pdf e txt gerado e salvo!',
'Arquivo 8931-8832 pdf e txt gerado e salvo!',
'Arquivo 8932-8833 pdf e txt gerado e salvo!',
'Arquivo 8933-8834 pdf e txt gerado e salvo!',
'Arquivo 8934-8835 pdf e txt gerado e salvo!',
'Arquivo 8935-8836 pdf e txt gerado e salvo!',
'Arquivo 8936-8837 pdf e txt gerado e salvo!',
'Arquivo 8937-8838 pdf e txt gerado e salvo!',
'Arquivo 8938-8839 pdf e txt gerado e salvo!',
'Arquivo 8939-8840 pdf e txt gerado e salvo!',
'Arquivo 8940-8841 pdf e txt gerado e salvo!',
```

```
'Arquivo 8941-8842 pdf e txt gerado e salvo!',
'Arquivo 8942-8843 pdf e txt gerado e salvo!',
'Arquivo 8943-8844 pdf e txt gerado e salvo!',
'Arquivo 8944-8845 pdf e txt gerado e salvo!']
```