LAB 1

WEBSCRAPING

Essa atividade de laboratório pode ser realizada utilizando Regex, BeautifulSoup, Selenium ou uma combinação destas ferramentas.

Você pode realizar essa tarefa em grupos de 2 até 4 integrantes. Apenas um dos integrantes deve submeter realizar a entrega no Moodle.

Lembre-se de tomar cuidado para não estressar o servidor com requisições em excesso (principalmente para a Tarefa 2).

Entrega:

- Data de Entrega: **Ver no Moodle**
- Forma de Entrega: **Link no Moodle**

  O que deve ser entregue:

  a) Scripts python ou jupyter notebooks com o código que faz as tarefas abaixo.

  b) Dados obtidos via scraping (csv, json e imagens baixadas)

  c) Orientações sobre como executar os scripts (como comentário no código, arquivo README.txt ou células de texto em jupyter notebook).

  d) Lista de integrantes informando Nome e Matrícula.

  e) (Recomendável) Arquivo *environment.yml* para instalação do ambiente virtual `conda`.

Tarefa 1 – Web Scraping em Ambiente Controlado (Peso 7.0)

Considerando a aplicação web de exemplo vista em aula, faça scraping das seguintes informações:

1) Faça um crawler que navegue pelas páginas de países e baixe os htmls.

2) Faça scraping dos htmls baixados e salve os dados retirados em um arquivo csv. Salvar uma coluna extra no csv contendo um timestamp do momento no qual os dados foram obtidos.

3) Faça um crawler que monitore as páginas de países e procure por atualizações. Caso algum registro tenha sido atualizado esse deve ser atualizado no arquivo CSV, caso contrário manter a versão anterior.

Observação, para realizar modificações no site sugere-se atualizar os dados diretamente no SQLite da página. Um programa util para fazer tal atualização é o SQLiteStudio `https://sqlitestudio.pl/`.

Tarefa 2 – Web Scraping em Ambiente Real (Peso 3.0)

Considerando o site https://www.imdb.com/.

1) Faça scraping para obter os 250 filmes com as maiores avaliações do IMDB. Existe uma página do imdb que possui essa listagem.

2) Faça scraping das páginas específicas dos 250 filmes obtidos no item anterior. Devem ser obtidos: Título, Ano de lançamento, url do poster, imagem do poster, nota imdb, lista de gêneros e lista de diretores. Trate os casos nos quais não existirem essas informações.

3) Salve as informações obtidas em arquivo json.