Disciplina: Coleta, Preparação e Análise de Dados
Professor: Luan Fonseca Garcia
Trabalho 1

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

Você pode realizar essa tarefa em duplas ou trios, não é possível realizar individualmente. Apenas um dos integrantes deve submeter o arquivo no moodle.

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

Data de Entrega: Verificar no moodle
Forma de Entrega: Moodle
O que deve ser entregue:

a) Scripts python ou jupyter notebooks com o código que faz as tarefas abaixo.
b) Ambiente conda utilizado para executar o código (arquivo environment.yml)
c) Dados obtidos via scraping (csv, json e imagens baixadas)
d) Orientações sobre como executar os scripts (como comentário no código, arquivo README.txt ou células de texto em jupyter notebook).
e) Lista de integrantes informando Nome e Matrícula.

Tarefa 1 – Web Scraping em Ambiente Controlado (8.0)

Considerando a aplicação web de exemplo vista em aula e disponível no moodle, considere as seguintes tarefas:

1) Faça um crawler capaz de navegar por todas as páginas de países e baixar seus HTMLS.

2) Faça scraping dos htmls baixados e armazene os seguintes dados dos países em um arquivo CSV:

    a. Nome do país (campo country)
    b. Nome da capital do país (campo capital)
    c. Nome da moeda do país (campo Currency Name)
    d. População do país (campo population).

    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.

Tarefa 2 – Web Scraping em Ambiente Real (2,0)

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

1) Faça scraping para obter os 250 filmes com as maiores avaliações do IMDB. Devem ser obtidos: Título, Ano, url do poster, imagem do poster e nota imdb.

2) Faça scraping das páginas específicas dos 250 filmes obtidos no item anterior. Obtenha dessa página uma lista de gêneros, popularidade e lista de diretores (trate os casos nos quais a informação não estiver presente).

3) Salve as informações obtidas em arquivo json.