

Esta 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), pois seu IP pode ser banido de acessar o site.

Data e forma de apresentação do trabalho

O trabalho do grupo deve ser **apresentado em laboratório**, durante o período de aula, até o dia **27/03/2024**. Não será possível apresentar o trabalho fora do horário de aula e nem depois da data limite.

A apresentação consiste em apresentar o código e o programa em funcionamento para o professor. O professor poderá realizar perguntas relativas ao trabalho para qualquer integrante do grupo e as respostas também são consideradas na nota do trabalho.

Relatório individual

Além da apresentação, cada integrante deve entregar um **relatório individual** descrevendo qual sua participação no trabalho e qual sua percepção da participação dos colegas no desenvolvimento do trabalho.

A entrega do relatório é OBRIGATÓRIA, alunos que não entregarem o relatório não terão a nota da apresentação considerada.

Data limite para entrega do relatório é **31/03/2024**, via moodle.

Tarefa 1 – Web Scraping em Ambiente Controlado (5 Pontos)

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

- 1) **(1 Ponto)** Faça um crawler capaz de navegar por todas as páginas de países e baixar seus HTMLS.
- 2) **(2 Pontos)** 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. Área total do país em quilômetros quadrados (campo area).
 - d. Nome de todos os países vizinhos (campo neighbours - Atenção, é o NOME e não a sigla!)

Salvar uma coluna extra no csv contendo um timestamp do momento no qual os dados foram obtidos.

- 3) **(2 Pontos)** 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 (3 Pontos)

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

- 1) **(1 Ponto)** Faça scraping para obter as 250 séries mais populares segundo o IMDB. Devem ser obtidos: Título, Ano de estreia, Número Total de Episódios, Nota do IMDB e o link para página da série.
- 2) **(1 Ponto)** Faça scraping das páginas específicas das 250 séries obtidas no item anterior. Obtenha dessa página uma lista de gêneros, popularidade e lista dos atores presentes no elenco principal (Não é a lista completa de atores!).
- 3) **(1 Ponto)** Salve as informações obtidas em um arquivo de tipo JSON.

Tarefa 3 – Relatório Individual (2 Pontos)

Cada aluno deve elaborar um relatório descrevendo a sua participação no desenvolvimento do trabalho, as tecnologias utilizadas para resolver as tarefas e as facilidades e dificuldades encontradas. Além disso, o aluno também deve descrever a sua percepção sobre a participação dos colegas no trabalho. A entrega deve ser em formato de relatório, contendo título e nome do aluno.