Pontifícia Universidade Católica - PUCRS

Disciplina: Coleta, Preparação e Análise de Dados
Prof. Lucas Rafael Costella Pessutto

Trabalho 1

Coleta de Dados

Esse trabalho pode ser realizado utilizando Regex, BeautifulSoup, Selenium e/ou uma combinação destas ferramentas.
Você pode fazer o trabalho em **duplas ou trios**, não sendo possível realizar individualmente. Apenas um dos integrantes deve submeter os arquivos do trabalho no moodle.
Lembre-se de tomar cuidado para não estressar o servidor com requisições em excesso (principalmente para a Tarefa 2).
Data de Entrega: 16/09/2024 (Tarefas 1 e 2) – 23/09/2024 (Tarefa 3)
Forma de Entrega: O trabalho deve ser apresentado durante o período de aula. Não haverá uma aula de apresentação para este trabalho. O grupo deve marcar com o professor um horário para realizar a apresentação até o dia da entrega do trabalho. A apresentação consiste em mostrar o código e o programa em funcionamento para o professor. Poderão ser feitas perguntas sobre o funcionamento do trabalho para o grupo.
Além disso, deve ser entregue no moodle:

- Scripts Python ou Jupyter notebooks com o código que realiza as tarefas solicitadas.
- Dados obtidos via scraping (csv, json)
- Orientações sobre como executar os scripts (como comentário no código, arquivo README.txt ou células de texto em jupyter notebook). Incluir comentários sobre a configuração do ambiente de desenvolvimento necessário para rodar os scripts.

Tarefa 1 – Web Scraping em Ambiente Controlado (50%)
Considerando a aplicação web de exemplo vista em aula e disponível no moodle, considere as seguintes tarefas:

1. **(10%)** Faça um crawler capaz de navegar por todas as páginas de países e baixar seus HTMLS.
2. **(20%)** Faça scraping dos htmls baixados e armazene os seguintes dados dos países em um arquivo CSV:
    - Nome do país (campo country)
    - Nome da capital do país (campo capital)

- Nome da moeda do país (campo Currency Name)
- Nome do Continente (Atenção: é o nome do continente e não a sigla!)

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

3. **(20%)** 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 (30%)
Considerando o site `https://www.imdb.com/`.

1. **(10%)** Faça scraping para obter os 250 filmes com as maiores avaliações do IMDB. Devem ser obtidos: Título, Duração, url do poster, imagem do poster e nota imdb.

2. **(10%)** Faça scraping das páginas específicas dos 250 filmes obtidos no item anterior. Obtenha dessa página a popularidade e a listagem do elenco principal (incluindo nome do ator/atriz e da personagem).

3. **(10%)** Salve as informações obtidas em arquivo json.

Tarefa 3 – Relatório Individual (20%)
Cada aluno deve elaborar individualmente um relatório descrevendo a sua participação no desenvolvimento do trabalho e os principais desafios encontrados e habilidades adquiridas com esta atividade. Além disso, cada aluno deve descrever como o grupo se organizou para realizar o trabalho e a sua percepção sobre a participação dos colegas no trabalho. A entrega deve ser feita em um arquivo pdf.