Mestrado em Engenharia Informática
Ciência de Dados
Ano Letivo 2023/2024

Mini Projeto II

Trabalho Individual
Data Acquisition (web scraping)

© Ricardo Campos
ricardo.campos@ubi.pt

O trabalho prático é obrigatório para a obtenção de aprovação na unidade curricular. A não entrega durante o prazo previsto implica a automática reprovação dos alunos.

Objetivo: Familiarização com o processo de aquisição de dados com recurso a web scraping.

Entrega: Os trabalhos (em formato notebook – devidamente documentados) devem ser inseridos na plataforma de e-learning (moodle) até 02/04/2024, 23h59.

Realização do trabalho: Os trabalhos devem ser realizados individualmente.

Tarefa 1: Familiarização com *Web Scraping*

1. Extraia informação de uma página web estática à sua escolha (e.g., extrair informações dos destaques listados na página web do https://ticketline.sapo.pt/. Nota: Não serão admitidos trabalhos que obtenham informação a partir da mesma página web. Nesse sentido, solicita-se que indique a sua preferência (*first come – first serve*) no ficheiro abaixo: https://1drv.ms/x/s!AqbUf6ry5g9tlEG8UZfAp0ENctR5?e=DkxuHQ.

2. Extraia informação a partir de uma página dinâmica à sua escolha (recorrentemente atualizada, por exemplo uma página que reúna notícias ou eventos) com recurso ao *Selenium*. Defina um *job scheduler* (pode recorrer a um *time.sleep*) que programe a obtenção dos dados a cada hora durante o espaço de 2 dias consecutivos. Guarde os resultados num ficheiro JSON. Registe o desenrolar do processo de obtenção de

dados num ficheiro de *logs* (e.g., via biblioteca *logging*). Nota: pode recorrer ao https://www.pythonanywhere.com/ como alternativa à sua máquina local para obter os dados durante o período acima referido. Não serão admitidos trabalhos que obtenham informação a partir da mesma página web. Nesse sentido, solicita-se que indique a sua preferência (*first come – first serve*) no ficheiro abaixo: https://1drv.ms/x/s!AqbUf6ry5g9tlEG8UZfAp0ENctR5?e=DkxuHQ.