Pontifícia Universidade Católica - PUCRS

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

Exercício: Coleta de Dados

Coleta de Dados na Wikipédia

Administrativo
Este exercício pode ser feito em **trios ou quartetos**, não sendo possível realizar individualmente. Apenas um dos integrantes deve submeter os arquivos do trabalho no moodle.
Data de Entrega: 08/09/2024 (Turma 10) – 15/09/2024 (Turma 30)
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. 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.
- Link para um repositório com os dados obtidos via scraping (páginas html)
- 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 – Desenvolvendo um crawler
Escreva um crawler para descobrir e coletar **páginas sobre pessoas** da Wikipédia em Português. Seu programa deve coletar 1.000 páginas de pessoas diferentes da Wikipédia, à partir da página inicial: `https://pt.Wikipedia.org`. Não devem ser coletadas páginas de outros sites, além de páginas da Wikipédia que não sejam referentes à pessoas.
O crawler deve funcionar da seguinte maneira:

1. Obtenha uma página.
2. Verifique se a página se refere à uma pessoa
3. Salve a página como um arquivo html, chamado `<nome_pessoa>.html` (em caso afirmativo no item anterior)
4. Se a página for da Wikipédia, extraia todos os links que se encontram nessa página (extraía links das páginas que não sejam de pessoas, de modo a otimizar o processo de busca)

5. Filtre os links, removendo os que não se referem à verbetes e dos verbetes que já foram visitados.

6. Guarde esses links em uma lista

7. Escolha um link não visitado para ser a próxima página.

8. Volte ao passo inicial

As páginas de pessoas coletadas deverão ser salvas como arquivos com a extensão `.html`. Lembre-se de tomar cuidado para não estressar o servidor com requisições em excesso!
Ao final do processo de coleta das páginas seu programa deve exibir a proporção de páginas coltadas / páginas visitadas.

Tarefa 2 – 6 Graus de Separação
"A teoria dos seis graus de separação originou-se a partir de um estudo científico desenvolvido pelo psicólogo Stanley Milgram, que criou a teoria de que, no mundo, são necessários no máximo seis laços de amizade para que duas pessoas quaisquer estejam ligadas" (Fonte: Wikipédia).
Vários estudos sobre graus de separação tem sido feitos aplicando esta teoria em redes de colaboração de pesquisadores, laços de amizades nas redes sociais, e na transmissão de doenças. Outro exemplo de ferramenta que relaciona artistas do meio musical pode ser vista em: `https://www.whosampled.com/six-degrees/`.
A segunda tarefa consiste em desenvolver um programa que encontre qual o grau de separação entre duas pessoas que estejam nas páginas que seu grupo coletou na etapa anterior do trabalho. Nesse exercício vamos considerar que duas pessoas estão conectadas caso haja um link na página da primeira pessoa remetendo para a segunda. Seu programa deve receber como entrada o nome das duas pessoas e deve exibir qual o grau de separação entre elas, mostrando todas as pessoas intermediárias entre a primeira e a segunda pessoa.