UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO

INSTITUTO DE CIÊNCIAS EXATAS

DEPARTAMENTO DE COMPUTAÇÃO

CURSO DE SISTEMA DE INFORMAÇÃO

Professor Tiago Cruz de França

Para os exemplos a seguir, pode ser necessário pedir credenciais de acesso. Você também pode buscar bibliotecas para facilitar o acesso. Esse exercício deverá ser feito durante o horário da aula do dia 23-09-2020.

- As questões 1, 2 e 4 são exercícios que possuem o objetivo de proporcionar uma experiência prática para cada aluno. Então, mesmo que tenha dificuldade, tente superá-la. Você pode perguntar, interagir, fazer junto, etc. Não deixe de ter essa experiência.
- As questões 3 e 5 são desafios que possuem propósitos semelhantes, porém são exercícios mais complexos. Vocês também podem se juntar em grupo ou toda turma colaborar para responder essas questões. Contudo, elas não precisam ser entregues no nosso prazo final.
- Por fim, a questão 6 é individual. Deve ser respondida por último. Construa sua proposta. Pense! Essa é uma questão aplicada e para reflexão.

A tarefa é desafiadora, mas as revelações serão sensacionais e a satisfação grande. Insista! Vamos fazer isso juntos.

1. Criar ou usar um mecanismo de coleta de dados de mídias sociais usando a API (pode ser feito para Twitter, Instagram ou Facebook). Observe a forma de autenticação ou uso de credenciais. Salve os dados em um arquivo ou no MongoDB/Couchbase.

2. Raspar dados de uma página web identificando 3 tipos de informação que você quer obter da página. Salve essas informações (dados) em um arquivo.

3. (DESAFIO) Tente raspar dados de uma página com mecanismos de detecção de robô. Sugiro tentar a página de apostas do bet365. Seu mecanismo funcionaria em páginas dinâmicas que atualizam/carregam mais dados a partir de eventos?

4. Busque mecanismos de *crawler* na web. Cite 2 deles.

5. (DESAFIO) I. Tente fazer uma busca no Google com a palavra "COVID-19" no período de 01-02-2020 a 31-03-2020 e marque que só deseja resultados em português brasileiro. II. Com o resultado, faça o *scrape* de todo conteúdo dos dois primeiros resultados. III. Crie um mecanismo para automatizar as etapas I e II e que receba por parâmetro a data inicial e final da coleta, o idioma, o termo de busca e a quantidade de resultados que devem ser considerados em cada busca para que seus conteúdos (dados) sejam “raspados”.

6. (FAÇA SOZINHO) Descreva uma estratégia (pense individualmente) para coletar dados sobre a COVID-19 no Brasil para buscar denúncias sobre aglomerações ou abusos ocasionadas por serviços públicos (ônibus, órgãos públicos com redução ou local para idosos sentarem, bancos gerando aglomerações em filas, falta de suporte para gestantes e puérperas/lactantes, etc.).
Obs: Escolha uma situação dentro do cenário da pandemia descreva o que quer saber (qual informação que você quer saber). Seu cenário deve proporcionar esse entendimento. Em seguida, aponte uma mídia social ou página web(pelo menos) onde você vai coletar dados. Descreva sua estratégia: possui API (sim ou não)? vai usar a API caso exista (sim ou não)? vai construir um scrape? vai buscar por termos (palavras, frases ou hashtags) ou por publicações em algum local? vai restringir por idioma, qual o período?
O objetivo é “afiar” um pensamento crítico sobre a coleta de dados pensando em um cenário de análise.