Alunas: Caroline Galiza e Marcela Carvalho

3) Quais cuidados devem ser observados ao capturar dados de um site?

Ao capturar informações de um site que não possui dados abertos e uma API para consumo é necessário estar atento às questões legais dessas informações, evitando sempre a violação de determinações jurídicas traçadas pelas empresas, como os termos de uso e políticas de direitos autorais, bem como estar atento aos dados que são coletados para que não seja capturados informações sensíveis. Também é necessário fazer requisições sem prejudicar o funcionamento dos sites e respeitar as normas internacionais que regem à internet em aspectos jurídicos e de serviços em geral como a Lei de Fraude e Abuso de Computadores (CFAA)

4) Quais ameaças capturas automáticas proporcionam para sistemas web?

No quesito ameaça à sistemas web, acredito que há uma referência à queda no desempenho do site ou serviço, encontro de falhas que permitem o acesso à informações valiosas que estão pouco protegidas e até à derrubada de sistemas por sobrecarga no servidor.

5) Você diria que bots ou crawlers são programas facilmente paralelizáveis? Se sim, explique como isso seria implementado dando um exemplo.

Os crawlers são usados em mecanismos de busca, como o google e o yahoo, para varrer a web criando uma cópia das páginas visitadas. Estas são posteriormente processadas e utilizadas para gerar resultados mais rápidos quando uma busca é solicitada.

A varredura das páginas gera uma lista de novos hyperlinks, estes são identificados e adicionados à uma lista de endereços que devem ser visitados. Para conseguir varrer todas essas páginas relacionadas e capturar esse volume grande de dados faz-se necessário que os crawlers apliquem a política de Paralelização.

Paralelização é dividir os processos realizados para que seja possível rodar múltiplos processos ao mesmo tempo, esta atividade possui o objetivo de maximizar a taxa de download das páginas.