|  |  |  |
| --- | --- | --- |
|  | **CURSO:** | **MBA em BI e Big Data** |
| **DISCIPLINA:** | Captura e Visualização de Dados |
| **GRADE:** | Teoria, Prática e Introdução a Captura e Visualização (Scrapy, Javascript) |
| **PROFESSOR:** | Diénert Vieira (dienertalencar@gmail.com) |
| **ALUNO(S):** |  |

**Avaliação 1**

1. Implemente um programa que entre no site do UOL e imprima apenas a seguinte mensagem: A cotação atual do dólar é: <cotação>, onde <cotação> vai ser o valor capturado do site no momento. Procure uma forma de omitir as mensagens de log na execução do seu programa para aparecer apenas essa mensagem como saída.
2. Implemente um programa que receba um produto como parâmetro e liste o nome e o preço de todos esses produtos no mercado livre, com paginação incluída. Busque uma forma de passar um parâmetro para o seu programa.
3. Quais cuidados devem ser observados ao capturar dados de um site?

A captura de dados na web ainda é um assunto que gera muita discussão, sobretudo nos países onde a legislação ainda não foi devidamente atualizada. Muitos pontos ainda pouco definidos sobre o que é legal ou considerado ilegal na realização de *web scraping* permanecem sendo pontos que geram incertezas nas fronteiras do universo online. O grande cerne da questão é a forma como tais dados serão utilizados após a sua captura. Nos dias atuais a ciência de dados tem papel fundamental no universo corporativo. E a mineração de dados é uma das fontes de informações que são fundamentais no desenvolvimento de estratégias de negócios. Mas existem barreiras legais, além daquelas éticas na realização de captura de informações, quer sejam elas através de scraping ou crawling. É evidente que tais práticas acabam sendo utilizadas de forma indevida, ou efetivamente com fins ilegais para a pratica de crimes virtuais. À medida em que não se respeita os termos de serviço de uso ou não possui uma autorização formal para captura de dados, o operador está cometendo um ilícito. Atualmente nos Estados Unidos existe uma vasta legislação que regula o assunto.

Soluções como validação de usuário, detecção de ataques e sua consequente mitigação, monitoração do ambiente web onde as informações estão disponibilizadas, além de limitação de requisições de um determinado IP e alteração de marcação HTML(randomização de conteúdo), são medidas importantes no combate a este tipo de prática. A existência de uma equipe de TI capacitada e focada no monitoramento do ambiente, além da observância constante sobre a atualização dos softwares de segurança (anti-malwares) também são formas de prevenção das práticas citadas.

1. Quais ameaças capturas automáticas proporcionam para sistemas web?

Inicialmente os sistemas de capturas automáticas podem ser uma forma de identificar vulnerabilidades de segurança em ambiente web, além da quebra de privacidade dos dados de terceiros (roubo de dados). Sob a ótica do business *intelligence*, a espionagem industrial para obtenção de vantagem comercial. Outras ameaças mais comuns são: o *phising* que consiste na distribuição de links maliciosos que podem vir a executar uma vasta quantidade de funções conforme seja o interesse do operador; a quebra de senhas com o objetivo de acessar dados armazenados pela vítima; *Ransomware* cujo objetivo consiste em realização de um bloqueio total por criptografia do equipamento invadido, que somente é desbloqueado mediante pagamento.

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

Os crawlers são programas que com capacidade para rodar inúmeros processos ao mesmo tempo, sendo que as tarefas são totalmente independente entre si, sem que haja qualquer vinculação de seus jobs. O principal objetivo dessa prática é maximizar a taxa de download e minimizar os custos da paralelização. Para otimizar ainda mais o processo, existem práticas/políticas que evitam o repetido download de uma mesma url, situação que compromete a performance do processo sendo elas a atribuição dinâmica e a atribuição estática.

Obs.: As questões com códigos fonte devem apresentar o link da solução no seu repositório no Github