Encontro de Saberes 2017 - XXV Seminário de Iniciação Científica

Utilização de técnicas de aprendizado de máquina para otimização no custo de requisições web voltadas para webscraping e webcrawler

Gabriel Dornelas Tassar de Almeida (Autor), Felipe Leandro Andrade da Conceição (Orientador)

Instituição de Ensino - Centro Universitário de Belo Horizonte

Palavras Chaves:

Otimização, Web crawler, Web scraping, Requisição, Proxy

Resumo:

Quando se navega através da web, pode-se notar que uma grande massa de dados circula todos os dias, sejam notícias, compras, conversas, pesquisas, entre muitos outros tipos de dados. Segundo pesquisas, dado esse grande volume de dados existentes na web, um processo automático e metódico de indexação de conteúdo se faz extremamente necessário. Esse método pode ser chamado de Web Crawler ou Web Scraping(quando se trata de captura de dados específicos), que busca, filtra e muitas vezes persiste conteúdo das requisições feitas para a captura de dados. Este trabalho busca otimizar as requisições rest para essas técnicas. Dada uma requisição, a mesma pode ter um custo de tráfego, ter um uso de proxy e tempo. Para sistemas que utilizam dessas técnicas, os mesmos utilizam os serviços de proxies para mascarar as requisições, contudo muitos deles podem ser bloqueados e retornarem a resposta indesejada para os usuários desses serviços, assim como o número de tentativas para obter essa resposta pode ser grande com isso aumentando o tempo de espera pela resposta. Para otimização dessas requisições se faz uma escolha inteligente do uso de proxy, identificando o melhor tempo de espera e se a resposta é o que o usuário deseja, assim retornando a resposta desejada com um custo e tempo menor. Ao aplicar o projeto notou-se uma evolução no uso de proxies, assim diminuindo o número de requisições consideradas inválidas, ou seja bloqueio e redirecionamento para teste de robô. Foram usadas 2 técnicas para atingir o objetivo de escolha de proxies, sendo elas, a média móvel exponencial e a árvore de decisão j48, ambas retornaram resultados similares, porém a longo prazo a técnica de árvore de decisão se tornou pouco eficaz, considerando que a lógica de bloqueio dos sites acessados mudam e essa árvore teria que ser treinada novamente de tempos em tempos. Gostaria de agradecer aos meus Orientadores Felipe Leandro e Fabrício Massula que me ajudaram no desenvolvimento do projeto.

Publicado em:

- Evento:Encontro de Saberes 2017
- Área:CIÊNCIAS EXATAS E DA TERRA
- Subárea:CIÊNCIA DA COMPUTAÇÃO