Instituto Federal de Brasília
Campus Taguatinga
Superior em Computação

Trabalho de Programação

Solução de *Web Crawling/Scraping multithreaded* para a busca de documentos em formato PDF

Sistemas Operacionais - 2023/1
Professor João Victor de A. Oliveira

Data de Entrega: 01/12/2023
(Não serão aceitas entregas fora do prazo).

Grupos: Formados por até 3 estudantes.

Linguagem de programação a ser utilizada: C padrão ou C++.

Sistema Operacional de teste: Linux.

Formato de entrega: via ambiente virtual Google Classroom. Todos os códigos fontes além de um arquivo de texto README.txt deverão ser colocados em uma pasta contendo o primeiro e último nome de todos os componentes do grupo. O arquivo README deve conter o nome completo dos alunos, além de uma descrição sucinta de como compilar o código fonte, de como o programa funciona e de como o sistema foi implementado.

Introdução

Web crawling, também conhecido como spidering ou web scraping, é o processo automatizado de navegar pela World Wide Web e coletar informações de sites da internet. O objetivo principal do web crawling é indexar o conteúdo da web para fins de pesquisa e análise. Os programas de web crawling, chamados de web crawlers ou spiders, seguem links de página em página, baixando o conteúdo das páginas da web e coletando dados específicos, como texto, imagens, links e metadados.

Já o web scraping, também conhecido como raspagem de dados web, é uma técnica de extração de informações de sites da internet de forma automatizada. Ao contrário do web crawling, que se concentra em navegar pela web e indexar conteúdo, o web scraping se concentra na extração

específica de dados de páginas da web para fins de análise, armazenamento ou processamento posterior.

O uso dessas duas técnicas pode ser útil para extrair dados úteis em sites de instituições de ensino públicas, tais como informações de processo seletivo, cursos e eventos ofertados, normativas e outras informações pertinentes. Normalmente nesses tipos de sites, em especial de instituições públicas, grande parte das informações relevantes são armazenadas em PDF.

Neste trabalho, temos como objetivo criar uma solução de web crawling/scrapping que consiga listar todos os pdfs disponíveis no site do Instituto Federal de Brasília. Tal solução usará um método que visa explorar o paralelismo usando a técnica de *multithreading.*

Especificação da solução

A solução a ser desenvolvida deve conter a seguinte estrutura de threads:

- **Uma thread despachante:** responsável por encaminhar urls às threads que estiverem livres.
    - As URLs devem ser unicamente dentro do domínio **ifb.edu.br** e estão armazenadas em uma fila (ou estrutura equivalente);
    - A fila de URLs inicia com apenas uma URL ativa: https://ifb.edu.br/.

- **8 threads operárias:** Ficam inativas até o momento em que a thread despachante envia uma URL a ser analisada.
    - A thread deve obter o código fonte a partir da URL recebida;
    - Dentro do código fonte, a thread deve realizar uma busca de todas as URLs (do portal ifb.edu.br) salvando-as na fila de URLs;
        - Deve-se tomar cuidado de não armazenar uma URL já salva/acessada anteriormente;
    - Dentro do código fonte, a thread deve realizar uma busca de todos as referências a PDF;
        - As referências de arquivos pdf devem ser salvas em um arquivo no formato:
            - Uma linha com a url da página que contém pdfs
            - Um conjunto de linhas (com uma indentação de uma tabulação (\t)), onde cada linha possui o link do arquivo .pdf

O sistema deve funcionar enquanto houver URLs a serem acessadas no domínio ifb.edu.br. Caso a thread despachante identifique que nenhuma thread operária esteja em funcionamento e que a lista de URLs esteja vazia, o sistema deve ser encerrado.

Observações:

- Usar o pacote ***pthreads*** para a implementação das threads;
- A pesquisa sobre funcionalidades referentes à aquisição dos códigos fontes, quais funções de threads e quais mecanismos de comunicação entre threads faz parte do trabalho;
- Caso o programa não compile, ou tenha um funcionamento inesperado, o professor poderá marcar um horário com o grupo, de forma que este apresente seu programa;
- Cuidado ao realizar as comunicações entre threads (garanta a exclusão mútua);
- Cuidado ao acessar as URLs (evite fazer muitos acessos à mesma URL)
- Faz parte do trabalho procurar esclarecer eventuais dúvidas sobre a especificação aqui apresentada.

Bom trabalho!