Universidade Federal da Paraíba

Centro de Informática - PPGI

Disciplina: Teoria da Computação (2021.2)

Professor: Bruno Petrato Bruck

Projeto 2

Web Scrapper (Expressões Regulares)

Esse tema envolve a criação de um programa que seja capaz de analisar o código fonte da página de um artigo qualquer da **Wikipédia** por meio da utilização de **Expressões Regulares**.

Inicialmente, o programa deve perguntar ao usuário qual o link do artigo que deseja analisar. Utilizando expressões regulares, deve ser verificado se o endereço fornecido é válido e se pertence ou não a uma página do domínio *pt.wikipedia.org*. Após uma validação bem sucedida, o programa deve mostrar ao usuário um menu onde seja possível selecionar qual tipo de informação deseja-se extrair da página. No mínimo, devem ser fornecidas as seguintes funcionalidades:

a) Listar os tópicos do índice do artigo

b) Listas todos os nomes de arquivos de imagens presentes no artigo;

c) Listar todos os links para outros artigos da Wikipédia que são citados no conteúdo do artigo.

É importante enfatizar que, para encontrar as informações acima, devem ser utilizadas expressões regulares criadas pelos integrantes do grupo (1 ou 2 pessoas).