Atividade: Extração de E-mails

Pedro Acácio Rodrigues

Universidade Tecnológica Federal do Paraná - Campo Mourão (UTFPR-CM)

Departamento de Computação - DACOM

pedrorodrigues.2019@alunos.utfpr.edu.br

Abstract. *The activity aims to perform the extraction of links contained in a html page using regex.*

Resumo. *A atividade tem o objetivo de realizar a extração de links contidos em uma página html utilizando regex.*

1. Problema

Escrever uma Expressão Regular em Python ou outras ferramentas que recupere os links contidos em uma página HTML.

2. Base de dados utilizada

A base de dados utilizada contém todo o conteúdo da página HTML do portal da UTFPR, para download foi utilizado o comando:

```
wget "https://portal.utfpr.edu.br/campus/campomourao" -O utfpr-cm.html
```

3. Execução

Para a extração dos links, a base de dados utilizada é submetida a expressão regular para localizar todos os emails contidos:

```
<a\s+(?:[^>]*?\s+)?href="([^"#]+)"
```

Separando por partes a expressão regular, temos: “<a” corresponde literalmente ao caractere "<" seguido da letra "a", seguido de “\s+” que corresponde a um ou mais caracteres de espaço em branco.

“(?:” define que tudo contido dentro dos parênteses não será retornado como parte da correspondência. “[^>]*?\s+” corresponde a zero ou mais caracteres que não são o caractere ">" (ou seja, qualquer caractere dentro da tag <a> que não seja a sua tag de fechamento). O “*?” significa que essa correspondência deve parar o mais cedo possível. Depois disso, corresponde a um ou mais caracteres de espaço em branco.

“)?” fecha o grupo de captura e torna todo o grupo opcional. Isso significa que essa sequência de caracteres pode estar presente ou ausente no texto. “href="” corresponde

literalmente à sequência de caracteres "href=" que a tag utiliza. “([^"#]+)” corresponde a um grupo de captura que contém qualquer caractere que não seja "#" ou """, que aparece uma ou mais vezes. Isso corresponde ao valor do atributo "href" da tag <a>. “":” marca literalmente o caractere de aspas duplas que fecha o valor do atributo "href".

4. Referencias

https://docs.python.org/pt-br/dev/howto/regex.html