Exa844 – Programação para Redes
Professor: João B. Rocha-Junior

Atividade 1: *Parser HTML*

Descrição:

Desenvolva um Parser HTML. O parser dever ser capaz de baixar uma URL e fazer o parser na página. O Parser deve extrair as seguintes informações: 1) título da página, 2) texto (após extrair tags HTML), número total de termos (termos no título mais os termos no texto), número de termos distintos, tamanho da página em bytes, data da última atualização e centróide. O centróide é um objeto que armazena informações sobre os termos relevantes. Portanto, cada termo da página (presente no título ou no texto) deve ser processado, primeiro removendo a acentuação e transformando todos os caracteres em minúsculo, e depois verificando se ele é relevante ou não. No primeiro passo, o termo 'Canção' vai ser transformado em 'cancao'. No segundo passo, o termo vai verificar se ele pertence a uma lista de termos irrelevantes, chamada de stopList. Caso o termo esteja presente na stopList, ele deve ser desconsiderado (ou seja, não é relevante).

Para cada termo relevante, armazenado no centroide, deve se armazenar o seu peso e número de ocorrências. O peso de um termo deve ser calculado de acordo com o local em que aparecem na página de acordo com a tabela a seguir.

title	10	h6	4	u	3	sup	2
h1	7	a	5	strong	3	font	2
h2	6	big	3	strike	3	address	2
h3	5	b	3	center	3	meta	2
h4	4	em	3	small	2	OUTROS	1
h5	4	i	3	sub	2		

Exemplo:

```
<html>
<title>UEFS – Universidade Estadual de Feira de Santana</title>
<body>
  A UEFS possui diversos cursos como:<b> Computação, Medicina e Direito</b>
</body>
</html>
```

Centróide: (UEFS, 11, 2), (Universidade, 10, 1), (Estadual, 10, 1), (Feira, 10, 1), (Santana, 10, 1), (possui, 1, 1), (diversos, 1, 1), (cursos, 1, 1), (computacao, 3, 1), (medicina, 3, 1), (direito, 3, 1)

Produto:

Você deve implementar o programa, seguindo os conceitos da orientação a objetos. Sendo assim, todos os dados poderão ser consultados, manipulando os objetos individualmente. Portanto, o Parser deve possuir um método chamado getTitle() que retorna o título da página, getCentroide() que retorna um Centróide, dentro do centróide deve ser possível listar todos os termos (sugestão:

iterator) e ao acessar o termo, pode pegar dados do mesmo como número de ocorrências. Deve ser possível executar o programa da seguinte forma: java Parser http://www.uefs.br e o programa imprime todas as informações da página: título, texto, centroide, etc...

O programa deve ser enviado por email, em um único arquivo zip, até a data prevista no cronograma.