Exa844 – Programação para Redes
Professor: João Batista da Rocha Junior

Projeto 1: *Robô*

Descrição:

Todo engenho de busca necessita de um "robô" para coletar páginas automaticamente na WEB. O Robô inicia a varredura em um conjunto de *links* iniciais chamados de sementes. As sementes podem ser representadas em um arquivo .xml como descrito através da seguinte DTD:

```
<!DOCTYPE seeds [
   <!ELEMENT seeds (seed+)>
   <!ELEMENT seed (url, visited?)>
   <!ELEMENT url (#PCDATA)>
   <!ELEMENT visited (#PCDATA)>
]>
```

Ao visitar uma página, o robô deve realizar o parser na página HTML. O parser dever ser capaz de ler um arquivo HTML e extrair o seu título, texto, número de termos diferentes, quantidade de termos e centroide. O centroide corresponde a um conjunto de termos relevantes com seus respectivos pesos e ocorrência. No centroide o termo Canção e cancao devem ser considerados idênticos. Para definir os termos relevantes em português você precisará verificar se ele não pertence a um arquivo de termos irrelevantes para o português chamado de stoplist.txt, ex: o, a, um, uma, após, antes, etc. O peso de um termo deve ser calculado de acordo com as posições (tags) que ele aparece em um documento HTML, sugestão de pesos por tag.

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>
```

Centroide: (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)

Os dados extraídos pelo parser devem ser armazenados em um arquivo XML. Os links extraídos das paginas devem ser adicionados ao arquivo .xml que contém as sementes e servirão de matéria prima para novas coletas. O arquivo contendo as sementes não deve possuir *URLs* repetidas.

Produto:

Construir um robô capaz de ler um arquivo XML contendo as sementes, coletar informações sobre essas URLs e armazenar em arquivos XML, modificar o arquivo contendo as sementes atribuindo a data de visitação e adicionando os novos *links* encontrados. Crie métodos para atribuir os parâmetros de sua classe principal, como: endereço das sementes, diretório onde os XML representando as páginas ficaram armazenados, etc.

Você deverá produzir quantas classes ache necessário, sempre observando os princípios da orientação a objetos. Todos os itens poderão ser extraídos e manipulados individualmente, sendo assim o Parser deve possuir um método chamado getTitle() que retorna o título da página, getCentroide() que retorna um Centroide, dentro do centroide deve ser possível pegar um termo através de um método chamado getTermo(), através do Termo deve ser possível pegar o número de ocorrências e assim por diante.