

CURSO DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
RECUPERAÇÃO DE INFORMAÇÃO - MÁQUINAS DE BUSCA NA WEB

Última alteração: 29 de abril de 2009

1º semestre de 2009

Professores: Berthier Ribeiro-Neto e Nivio Ziviani

Monitor: Anísio Mendes Lacerda

Trabalho Prático 1:

30/04/2009

Data de Entrega: 25/05/2009

Penalização por Atraso: 1 ponto até 01/06/2009 mais 1 ponto por dia útil a seguir

Processamento de Consultas

O objetivo deste trabalho é projetar e implementar um sistema de programas para responder eficiente e eficazmente consultas realizadas sobre o índice gerado a partir dos programas do Trabalho Prático 1.

O sistema deverá:

1. Realizar ordenação (*ranking*) usando:
 - (a) o modelo de espaço vetorial.
 - (b) o modelo BM25.
2. Desenvolver uma interface HTML com uma caixa de consultas e colocar a interface disponível para consultas.
3. Retornar as top10 URLs mais relevantes obtidas com cada modelo de ranking para cada uma das 50 consultas disponíveis no endereço:

<http://garnize.latin.dcc.ufmg.br/colecaoRI/consultas-50>.

O formato a ser utilizado é:

```
<CONSULTA>
VETORIAL
<URL1>
<URL2>
...
<URL10>
BM25
<URL1>
<URL2>
...
<URL10>
```

Tudo que está entre < ... > deve ser substituído pelos dados reais. Deve ser gerado um arquivo por consulta.

4. Participar do processo de avaliação de documentos relevantes da coleção que foi indexada no TP1, para as 50 consultas citadas acima. Para tal vamos utilizar o sistema disponível em <http://garnize.latin.dcc.ufmg.br/ri-evaluation>, cujo uso será explicado em sala de aula.

Referências

- [1] B. A. V. Pôssas. Set-Based Vector Space Model: A new Approach for Correlation Based Model. Tese de Doutorado, Departamento de Ciência da Computação, UFMG, 22 de agosto de 2005.
- [2] S. E. Robertson and K. S. Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, May-June 1976.
- [3] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *Proceedings of the Third Text Retrieval Conference (TREC-3)*, pages 109–126. NIST Special Publication 500-225, April 1995.
- [4] Ian H. Witten, Alistair Moffat e Timothy C. Bell: *Managing Gigabytes: Compressing and Indexing Documents and Images*, Morgan Kaufmann Publishers, 1999, second edition.