Universidade Federal de Lavras
Departamento de Ciência da Computação

PCC518 – Tópicos em Computação I - Recuperação de Informação
GCC204 – Recuperação de Informação

Professor: Denilson Alves Pereira

Trabalho Prático 1

- Trabalho individual
- O trabalho deve ser entregue em versão eletrônica pelo Moodle (http://aluno.dcc.ufla.br). Envie arquivos somente nos formatos txt e pdf (não enviar .doc, .docx, .odt etc.). Arquivos compactados, somente .zip e .tar.gz (não enviar .rar, .z etc.). Não use acentos e nem "ç" nos nomes de arquivo.
- **Data limite de entrega: 06/04/2015**
- **Valor: 15 pontos**

O objetivo do trabalho prático é implementar o Modelo Vetorial para recuperação de informação e avaliá-lo em uma coleção de referência, de acordo com as seguintes instruções:

- A implementação pode ser feita nas linguagens de programação Java, C ou C++;
- Crie uma estrutura de dados usando Índice Invertido, com o vocabulário mantido em uma Tabela Hash, para armazenar os dados necessários para cálculo dos pesos dos termos;
- Use um dos esquemas de peso e fórmula do cosseno propostas no livro texto da disciplina;
- Para avaliação do resultado, use a coleção de referência CFC (Cystic Fibrosis) e trace o gráfico de Precisão x Revocação para 11 níveis de revocação. Considere as médias de precisão e revocação para todas as consultas da coleção;
- Para indexação dos documentos da coleção CFC, utilize os termos contidos nos atributos AU, TI, SO, MJ, MN e AB/EX.

O que deve ser entregue:

- O código fonte dos programas, devidamente comentados;
- Um relatório (em forma de artigo) contendo introdução, referencial teórico, descrição do trabalho com suas estratégias de solução, experimentos executados e resultados obtidos, conclusão e referências bibliográficas. O relatório deve ter de 4 a 6 páginas, seguindo o template da SBC (http://www.sbc.org.br/index.php?option=com_jdownloads&Itemid=195&task=view.download&catid=32&cid=38).

Passo-a-passo: instruções a serem seguidas em cada semana de aula para completar o Trabalho Prático 1:

1. Implementar um módulo para pré- processamento de documentos e um módulo para indexação dos documentos de uma coleção.
   - A coleção de documentos deve estar em um diretório específico, com cada documento em um arquivo texto;
   - O módulo de pré-processamento deve ler cada documento da coleção, tratar codificações de caracteres, converter todo o texto para letras minúsculas, eliminar pontuações, símbolos desnecessários, tags (no caso de HTML, XML etc.) e stopwords, tokenizar cada palavra do texto e enviá-las para o módulo de indexação. O módulo deve ser genérico, funcionando para qualquer coleção, independente do teste com a coleção de referência CFC;
   - No caso específico de teste com a coleção de referência CFC, um outro pré-processamento é necessário para extrair somente os atributos de interesse de cada documento. Isso deve ser feito antes de passar a coleção para o módulo descrito no item anterior;
   - O módulo de indexação de documentos recebe do módulo de pré-processamento cada termo (token, palavra) e a identificação do documento onde ele ocorre, e o insere em uma estrutura de índice invertido. Nesta estrutura, o vocabulário deve ser mantido em uma Tabela Hash e a lista de ocorrências deve armazenar, em cada posição, a identificação do documento onde o termo ocorre e a frequência com que ele ocorre neste documento.
2. Implementar um módulo para processamento de consultas.
   - O módulo recebe como entrada os termos de uma consulta e retorna os documentos da coleção que são relevantes para a consulta, em ordem decrescente de relevância (*ranking*);
   - A função de relevância deve ser implementada usando o Modelo Vetorial com o esquema de pesos TF-IDF;
   - Para teste, use as consultas da coleção de referência CFC.
3. Implementar um módulo de avaliação de resultados.
   - O módulo recebe como entrada a identificação das consultas com as respectivas identificações dos documentos de seu resultado, informados como gabarito e pelo seu algoritmo, ordenados por relevância, e retorna os valores para as métricas Precisão e Revocação;
   - O módulo também gera a tabela de Precisão x Revocação para 11 níveis de revocação, para que seja montado o gráfico.