Universidade Federal de Uberlândia - UFU
GSI521 – Organização e Recuperação da Informação – 2017/01
Prof. Rodrigo Sanches Miani
Data de entrega – 04/07/2017 – Enviar relatório+código para o email orgriufu@gmail.com

Desenvolvimento do projeto final - Etapa 4 (Probabilístico+Coleção real)

1) Reutilize o código feito anteriormente para implementar um modelo simplificado de RI usando como base o modelo probabilístico sem conhecimento prévio dos relevantes ($R=r_i = 0$). Lembre-se que quando $n_i > N/2$ a fórmula original do probabilístico apresenta problemas. Adapte o seu código para lidar com essa situação. Teste o modelo probabilístico usando as coleções "TO DO" (livro) e também a base de hinos.

2) Anteriormente utilizamos uma coleção composta de quatro documentos (“TO DO”) para testar a implementação do ranking no modelo vetorial. A presente atividade de desenvolvimento de projeto consiste em construir uma coleção a partir de uma coleção de referência real.

A coleção de documentos possui o seguinte conteúdo:

.I 1
.T
experimental investigation of the aerodynamics of a
wing in a slipstream .
.A
brenckman,m.
.B
j. ae. scs. 25, 1958, 324.
.W
experimental investigation of the aerodynamics of a
wing in a slipstream .
an experimental study of a wing in a propeller slipstream was
made in order to determine the spanwise distribution of the lift
increase due to slipstream at different angles of attack of the wing
and at different free stream to slipstream velocity ratios . the
results were intended in part as an evaluation basis for different
theoretical treatments of this problem .
the comparative span loading curves, together with
supporting evidence, showed that a substantial part of the lift increment
produced by the slipstream was due to a /destalling/ or
boundary-layer-control effect . the integrated remaining lift
increment, after subtracting this destalling lift, was found to agree
well with a potential flow theory .
an empirical evaluation of the destalling effects was made for
the specific configuration of the experiment .

O arquivo que contém a coleção de documentos é o *cran.all.1400* dentro de *cranfield.zip*.

3) Criar um algoritmo para processar somente os documentos listados no arquivo "*documentos-selecionados.txt*" (64 documentos). O campo "**.I**" fornece o número (ID) do documento, enquanto que os termos de indexação serão extraídos à partir da leitura do campo "**.W**". É importante que:
   a) Cada documento receba a identificação conforme o valor presente no campo "**.I**". Por exemplo, o documento acima seria o documento 1.;
   b) Cada palavra obtida seja convertida em minúscula;
   c) Qualquer caractere que não seja uma letra deve ser removido;
   d) Somente palavras com um número maior ou igual a três caracteres serão indexadas.

4) Com base nos termos encontrados nos documentos, elabore 10 consultas e mostre o ranking (usando o modelo vetorial e o probabilístico) de cada uma delas. O ranking gerado é coerente com a consulta? Justifique. Caso existam, aponte as diferenças entre os rankings gerados pelos modelos vetorial e probabilístico.

Instruções para o relatório. O relatório deverá conter:

a) 10 consultas;
b) ranking de cada uma das consultas para o modelo vetorial;
c) ranking de cada uma das consultas para o modelo probabilístico;
d) Apontar as diferenças de ranking (vetorial x probabilístico) para cada uma das 10 consultas.