Universidade Federal de Uberlândia - UFU
GSI024 – Organização e Recuperação da Informação – 2017/01
Prof. Rodrigo Sanches Miani
Data de entrega – 25/07/2017

Desenvolvimento do projeto final - Parte 5

Objetivo: Investigar formas de avaliar a recuperação de informação.

1) Antes de iniciar a resolução da última etapa do projeto descompacte o arquivo "arquivos_etapa5.zip". Considere o seguinte **trecho** do arquivo "colecaoinformacao.txt":

.I 001
.W
what similarity laws must be obeyed when constructing aeroelastic models of heated high speed aircraft .
.I 002
.W
what are the structural and aeroelastic problems associated with flight of high speed aircraft .
.I 004
.W
what problems of heat conduction in composite slabs have been solved so far .

A tag ".I" indica a identificação da necessidade de informação enquanto que ".W" indica respectiva a necessidade de informação. Para cada necessidade de informação (I.001, I.002, .I 008 e .I 039) elabore **duas** consultas.

2) Considere o seguinte **trecho** do arquivo " julgamento-de-relevancia.txt":

1 57 2
1 378 2
1 859 2
1 185 3
1 30 3
1 37 3
1 52 4
1 142 4
1 195 4
1 875 2
1 56 3
1 66 3
1 95 3
1 462 4
1 497 3
1 858 3
1 876 3

1 879 3
1 880 3
2 12 1
2 15 2
2 184 2
2 858 2
2 51 3

A primeira coluna indica a identificação da necessidade de informação, a segunda coluna mostra o número do documento e a terceira coluna apresenta um código de relevância com a seguinte escala:

1 - o documento é a resposta completa para a necessidade de informação;
2 - o documento possui um alto nível de relevância para a necessidade de informação;
3 - o documento possui um nível moderado de relevância para a necessidade de informação;
4 - o documento possui um nível baixo de relevância para a necessidade de informação.

Com base nas duas consultas geradas no exercício anterior, encontre os níveis de precisão e revocação dos modelos vetorial e probabilístico (inclusive os gráficos) e a Média das Precisões Médias (MAP) para cada uma das consultas que foi elaborada. Considere que um documento relevante para uma necessidade de informação foi avaliado com qualquer um dos valores de relevância (1, 2, 3 ou 4). Interprete os resultados obtidos. Qual modelo (vetorial ou probabilístico) obteve os melhores resultados?

3) Refaça o exercício 4) mas agora considere que documentos relevantes devem ter códigos de relevância igual a 1 ou 2. Os níveis de precisão e revocação são diferentes? Explique.

4) Use o coeficiente de Spearman (correlação) para comparar o ranking obtido entre cada uma das consultas nos modelos vetorial e probabilístico. Elabore a seguinte tabela e discute a diferença entre os rankings gerados pelo modelo vetorial e probabilístico:

Necessidade de Informação	Consulta	Coeficiente de correlação (vetorial/probabilístico)
I.001	Q1	X1
I.001	Q2	X2
I.002	Q1	X3
I.002	Q3	X4
...	...	...