Universidade Federal do Rio de Janeiro (UFRJ)
Departamento de Ciência da Computação (DCC)

DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO - UFRJ -

Recuperação da Informação (MAB605)

Modelo Probabilístico – Implementação

Profa. Giseli Rabello Lopes

Exercício - Modelo Probabilístico

- Partir da implementação desenvolvida nas aulas anteriores e fazer a implementação do modelo BM25 adotando a formulação a seguir:

$$sim_{BM25}(d_j, q) \sim \sum_{k_i[q,d_j]} \mathcal{B}_{i,j} \times \log \left(\frac{N - n_i + 0.5}{n_i + 0.5} \right)$$

$$\mathcal{B}_{i,j} = \frac{(K_1 + 1) f_{i,j}}{K_1 \left[(1 - b) + b \frac{len(d_j)}{avg\_doclen} \right] + f_{i,j}}$$

Exercício - Modelo Probabilístico

- Por fim, gere o *ranking* final dos documentos para uma consulta q especificada (ordem que os documentos seriam ranqueados).
- Sua implementação deve permitir a configuração dos parâmetros b e K_1
  - Utilize K_1=1 e b=0.75 para testar sua solução na coleção de exemplo que vem sendo adotada
  - A estimativa do tamanho de cada documento deve ser realizada pelo somatório das frequências dos termos de indexação presentes nele

Exercício - Relembrando

- Exemplo de entradas:

```
M=['O peã e o caval são pec de xadrez. O caval é
o melhor do jog.';
'A jog envolv a torr, o peã e o rei.';
'O peã lac o boi';
'Caval de rodei!';
'Polic o jog no xadrez.']; //conjunto de
documentos
stopwords=['a', 'o', 'e', 'é', 'de', 'do', 'no',
'são']; //lista de stopwords
q='xadrez peã caval torr'; //consulta
separadores=[' ',',','.','!','?']; //separadores
para tokenizacao
```

Exercício - Relembrando

- Sua implementação deve:
  - Tokenizar os documentos utilizando os separadores adequados
  - Normalizar termos (ex. caixa-baixa) e eliminar stopwords das consultas e documentos
  - Usar uma solução de indexação utilizando uma variação da matriz de incidências (obs.: guarde a frequência de aparecimento dos termos em cada documento)

Universidade Federal do Rio de Janeiro (UFRJ)
Departamento de Ciência da Computação (DCC)

DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO - UFRJ -

Recuperação da Informação (MAB605)

Dúvidas?

Profa. Giseli Rabello Lopes
giseli@dcc.ufrj.br
CCMN - DCC - Sala E-2012