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

DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO - UFRJ -

Recuperação da Informação (MAB605)

Modelo Booleano – Implementação

Profa. Giseli Rabello Lopes

Dicas: Scilab - `help`

- `:`
- `'`
- `size`
- `+`
- `convstr`
- `tokens`
- `gsort`
- `unique`
- `members`
- `&`
- `|`
- `%T`
- `%F`
- `input`
- `disp`
- `if`
- `for`
- `while`

Dicas: Python

- Google Colaboratory
  - https://colab.research.google.com/drive/1IG_SkBSFM4Iee3-B9B2geYdDkUdZz9xJ

Exercício

- Implemente o modelo booleano de recuperação de informação, tendo como entradas:
  - Vetor coluna onde cada linha representa o texto de um documento (matriz *Nx1*)
  - Vetor linha de strings (matriz *1xNs*), onde o elemento em cada coluna armazena uma stopword
  - String contendo os termos da consulta (separados por espaços)
  - Vetor linha de caracteres (matriz *1xNc*), onde o elemento em cada coluna representa um separador a ser usado na tokenização dos documentos

*Sugestão use Scilab ou Python

Exercício

- Exemplo de entradas (no Scilab):

```
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

- 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)
  - Responder consultas puramente conjuntivas e disjuntivas:
    - AND entre todos os termos da consulta
    - OR entre todos os termos da consulta

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