Busca e Recuperação de Informação

Exercícios de laboratório:

Unidade II - O problema da Recuperação de Informação

Professor: Fellipe Duarte

UFRRJ
UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO

- O Objetivo desta lista de exercícios:
  - é criar uma lista com os dados da coleção onde as consultas(queries) serão buscadas.
  - Realizar a primeira busca em coleções de dados(datasets) de BRI
- Para cada coleção:
  - você tem que criar uma fila dos documentos da base.
  - o texto de cada documento é o elemento que será buscado.
- Por exemplo, se a coleção for uma lista de receitas:
  - você deverá implementar uma fila
  - o texto de cada receita representa um item da fila.

- **Exercício 1:**
  - Criar uma fila (FIFO) “spa_queue” que deve:
    - conter o texto de todos os documentos da pasta **source** do dataset “Short Plagiarised Answers“
  - Em seguida, responda às seguintes perguntas:
    - 1.1 Quanto tempo levou para abrir, ler o texto e armazenar em spa_queue todos os documentos?
    - 1.2 Qual o tempo médio para realizar a tarefa anterior? (Você deve medir individualmente e calcular a média aritmética)

- **Exercício 2:**
  - Buscar uma consulta da pasta “light”.
  - i. e., cada arquivo da pasta tem um documento de texto que é uma consulta.
  - Por exemplo, o arquivo “g0pB_taskd.txt” tem um texto com a definição do teorema de bayes.
    - Abra este arquivo e encontre na lista “spa_queue” quem tem exatamente o mesmo texto que “g0pB_taskd”.
  - Em seguida responda às seguintes perguntas:
    - 2.1 Quanto tempo levou para abrir, ler o texto e encontrar todos os textos de “spa_queue” ? (lembre-se que este tempo é o tempo para todos os arquivos da pasta “light“)
    - 2.2 Qual o tempo médio para realizar a tarefa anterior? (Você deve medir individualmente e calcular a média aritmética)
    - 2.3 Você encontrou algum resultado em todas as consultas?

- **Exercício 3:**
  - Ao realizar busca do exercício 2 guarde uma lista com os índices das posições de todos os textos encontrados em "spa_queue".
  - Por exemplo:
    - se o texto de uma consulta é igual aos textos na posição 1, 4 e 7 de "spa_queue"
    - no final da busca você tem que ter uma lista [1,4,7] para aquela consulta.
  - Em seguida responda às seguintes perguntas:
    - 3.1 Para a consulta executada algum dos resultados encontrados está correto? (olhar no arquivo "query_answer_json")

- **Exercício 4:**
  - Vamos tentar melhorar o resultado da busca do exercício 3
  - Para tanto: 'vamos considerar que estamos aceitando como resposta qualquer texto que tenha pelo menos uma palavra da consulta.
  - Por exemplo:
    - se o texto a ser buscado é:
    - "A document is represented as a vector and each dimension corresponds to a separate term."
    - vale qualquer texto que tenha uma palavra "A", "document", "is", "represented", "as", "a", "vector", "and", "each", "dimension", "corresponds", "to", "separate" ou "term".
  - **DICA: use o operador IN**
  - Em seguida responda as perguntas 2.1 a 3.1 novamente.

- **Exercício 5 a 8** :
  - realizar os exercícios 1 a 4 na coleção Cranfield.
  - Diferenças importantes:
    - A fila que será criada, para a busca, agora terá nome “cran_queue”
    - Os textos que vão ser armazenados na fila:
      - agora estão em um arquivo só: O arquivo “cran.all.1400”
      - a sua lista tem que ter 1400 textos ao final do exercício 1
  - cada documento é identificado por uma linha “.I numero_documento” isto é “.I 001” é o primeiro, “.I 002” é o segundo e “.I 1400” é o último. (dica use expressão regular para encontrar o identificador e separar os textos)
  - A consultas estão no arquivo “cran.qry” e são 365 no total.
  - O Gabarito do resultado está no arquivo “cranqrel”

- **Exercício 9 a 12** :
  - realizar os exercícios 1 a 4 na coleção CF (do termo “Cystic Fibrosis”).

  - Diferenças importantes:
    - A fila que será criada, para a busca, agora terá nome “cf_queue”
    - Os textos que vão ser armazenados na fila agora estão nos arquivos “cf74-corrigido”, “cf75-corrigido”, “cf76-corrigido”, “cf77-corrigido”, “cf78-corrigido” e “cf79-corrigido” da pasta “cfc-xml”.
    - Os arquivos contém vários textos estruturados em xml usem apenas o texto que está na tag Abstract **(dica: usem algum leitor de xml com DOM)**
    - A consultas e o gabarito do resultado estão no arquivo “cfquery-corrigido”