Especialização em Inteligência Artificial (EiA – IFMG-OP)
Recuperação de Informação
Atividade 4 - Workshop sobre Sistemas de RI
Prof. Moisés - Aluno: Fernando dos Santos Alves Fernandes

SIRIE-Sistema de Recuperação de Informação sobre Enfermagem

Contexto

- **Estomaterapia**: especialidade de enfermagem voltada para a assistência de pacientes com estomias, fístulas, tubos, catéteres, drenos, feridas agudas e crônica (diabetes, úlceras, hérnias) e incontinências urinária e anal.
- Objetivo: construir um sistema de recuperação de informação capaz de permitir a busca de informações sobre o tema (conceitos, definições, tipos de procedimentos cirúrgicos associados, tratamentos, links de instituições que oferecem essa especialização, entre outras informações).
- Contexto ampliado para melhor avaliação do sistema: **Enfermagem**.

Coletor

- A estratégia de busca para aprofundamento das páginas encontradas foi a *Breadth-First Search* (Busca em Largura), com limite de profundidade (*max_depth*). Nesse coletor, simples, as *tags* de texto utilizadas foram apenas a <h1> e <h2>. Quanto às requisições, redirecionamentos de páginas foram permitidos, por meio do parâmetro *allow_redirects=True*. O parâmetro *headers* também foi utilizado, para evitar que o coletor fosse bloqueado por determinadas páginas, como as do Governo. Os resultados de algumas coletas de teste (apenas as *URLs* das páginas coletadas) podem ser vistos em arquivos *'.txt'*, que acompanham os arquivos fonte do coletor e do programa principal no *link* da Atividade 3.

Indexador

- O indexador foi implementado utilizando como abordagem uma lista invertida.
- Para a avaliar a força da palavra ou termo, foram considerados os seguintes pesos:
  - f(K): também conhecido como TF (*Term Frequency*, TF(t,d)), corresponde ao número de ocorrências do termo t no documento d;
  - F(K): total de ocorrências do termo t (ou chave K), considerando todos os documentos em que ele é encontrado;
  - n(K): também encontrado na literatura como DF(t), ou *Document Frequency* do termo t, corresponde ao número de documentos em que a chave K ocorre;
  - IDF(t) (*Inverse Document Frequency*, do termo t): peso do termo que considera o número de documentos coletados (N) e o DF do termo.
  - TF-IDF: peso combinando TF(t,d) e IDF(t) e que pode ser obtido pela expressão TF-IDF = f(k) * idf(k) = f(k) * log [N / n(k)].

Buscador

- Para permitir o ranqueamento dos resultados, considerando os diferentes termos de uma busca, foi utilizada a métrica TF-IDF implementada no indexador. Para cada termo da busca, os itens do índice invertido relacionados ao termo são ordenados decrescentemente e adicionados à lista de *links* relevantes.
- Duas estratégias de processamentos de consultas foram implementadas, a **Interseção**, que considera apenas os *links* relevantes que aparecem simultaneamente nos resultados de todos os termos da busca; e a **União**, que consideram todos os links relevantes de todos os termos da busca.
- Na classe do buscador também são implementados os métodos para o cálculo da Precisão (***precision***) e da Revocação (***recall***), que são utilizados para avaliar a qualidade dos resultados do sistema de recuperação de informação.