Projeto Final – Recuperação de Informação em Texto

Definições iniciais

Com o objetivo de permitir a mesma formação do projeto inicial, a implementação do projeto final poderá ser realizada individualmente ou em equipes formadas por até três integrantes. O projeto final possui **peso de 70% na avaliação do curso**. Importante lembrar que a emissão do certificado e a **atribuição das horas complementares está condicionada à obtenção da média 6**. O prazo para entrega deste projeto é dia 10/9/2021, às 17h.

Descrição da Aplicação

O projeto consiste na implementação de uma aplicação de um mecanismo de busca com o objetivo de produzir um ranking dos documentos de acordo com a sua similaridade com uma *string* informada pelo usuário.

Os professores disponibilizarão, juntamente com este enunciado, um *dataset* composto por 20 documentos relacionados com as paralímpiadas. Os documentos são notícias (ou partes de notícias) publicadas na mídia e estarão em formato PDF. Os documentos contêm informações do curso no cabeçalho e a fonte da informação no rodapé. As informações contidas no cabeçalho e rodapé deverão ser ignoradas pela aplicação, ou seja, apenas as informações da notícia devem ser comparadas com a *string* de busca. Este *dataset* será utilizado para a correção do projeto.

A aplicação deverá solicitar que o usuário informe o local onde o *dataset* com os documentos que serão alvo do mecanismo de busca estão armazenados. Além disso, a aplicação deverá solicitar que o usuário informe quais palavras deseja buscar nos documentos para que a aplicação produza o ranking. A entrada deverá seguir as seguintes regras:

a. As palavras que deverão ser localizadas no texto serão informadas em letras minúsculas.
b. Todas as palavras informadas deverão ser usadas como base para a criação do ranking, ou seja, a aplicação deve considerar o uso do operador AND. Importante: o operador não precisa ser informado e as palavras que compõem a *string* podem apenas ser separadas por espaços.

*** Um exemplo de entrada é: desenvolvimento aplicação mineração texto.**

A aplicação deverá implementar as funcionalidades listadas abaixo. É importante frisar que a aula do dia **02/09/2021** explicará o modelo de recuperação de informações que deverá ser implementado.

1. A aplicação deverá ler todos os documentos PDF localizados no local informado pelo usuário e criar o modelo de espaço de vetores (Vector Space).
2. A criação do espaço de vetores deve considerar a remoção de stop words e a redução de cada termo ao seu radical (*stemming*).
3. Para facilitar a criação do ranking, a inclusão do vocabulário de cada documento no espaço de vetores deverá considerar o número de ocorrência de cada termo.
4. A aplicação deverá conter uma opção para salvar em arquivo o espaço de vetores no formato CSV.
5. A aplicação deverá armazenar cada consulta na base SQLite. Os dados armazenados são: data, hora, *string* de busca, ranking dos documentos e valor de similaridade de cada documento.
6. A aplicação deverá exibir o ranking dos documentos em tela, juntamente com a similaridade com a *string* de busca. Caso dois ou mais documentos apresentem o mesmo *score* de similaridade, eles podem ser apresentados em qualquer ordem (considerando apenas a posição dos documentos com mesmo *score*!).

É importante frisar que a as palavras poderão constar no documento em mais de um formato. Por exemplo, a palavra **aplicação** pode constar também como **Aplicação**, **APLICAÇÃO**, **aplicacao**, etc.

EXTRA: a implementação do peso IDF (Inverse Document Frequency) na inclusão do vocabulário de cada documento no modelo de espaço de vetores.

Entrega da aplicação

A entrega do projeto será realizada através do GitHub. Para isso, deverá ser usado a mesma conta do primeiro projeto. Na conta do usuário, o projeto final deverá ser nominado IR-Unisinos. Caso aconteça alguma mudança na equipe ou na conta usada, o aluno/a equipe deverá contatar o professor.

Apesar do Form usado no projeto inicial conter os integrantes da equipe, solicitamos que os créditos também sejam incluídos na aplicação desenvolvida.

No dia 10/9/2021, a partir das 17h, os professores acessarão o projeto IR-Unisinos em cada GitHub e farão um clone das aplicações para avaliação.

Método de avaliação

Os critérios de avaliação do projeto são apresentados a seguir. Também informamos o peso que será considerado para cada item na correção:

1. Geração do espaço de vetores com os vocabulários, termos e pesos corretos (peso: 4 pontos);
2. Implementação de remoção de *stop words* e *stemming* (peso: 2 pontos);
3. Armazenamento do histórico de pesquisa na base SQLite (peso: 1 pontos);
4. Criação do ranking corretamente a partir da similaridade da *string* de busca com os documentos (peso: 3 pontos);
5. Implementação do critério EXTRA considerando a frequência de um termo no corpus (valor extra: 1 ponto na média final, considerando que a nota máxima do curso é 10).