Trabalho 1 – Recuperação de Texto

INF-0611 – Recuperação de Informação

Neste trabalho, usaremos uma coleção de artigos da revista TIME. Essa coleção é composta por 425 artigos (documentos). Além dos documentos, também disponibilizamos exemplos de consultas com seus respectivos vetores de *ground truth*.

O objetivo deste trabalho é aprimorar o conhecimento sobre os modelos de Recuperação de Informação tf-idf e bm25. Para isso, faremos a comparação desses modelos utilizando os métodos de Avaliação de Ranking, apresentados na aula inicial da disciplina.

Preparação do Ambiente

Antes de começar o desenvolvimento do trabalho, **leia este documento com atenção.** Revise os códigos das Aulas 1 e 2, pois eles servirão de referência para realizar as tarefas a seguir.

Todos os arquivos necessários estão disponíveis na página da disciplina (Moodle), assim sugerimos que organize-os em uma mesma pasta de seu computador. Abaixo listamos os arquivos disponibilizados e uma breve descrição.

`inf0611_trabalho1.R`: Ao abrir este arquivo, mude o encoding para `UTF-8`. Neste arquivo temos um esboço das tarefas de implementação a serem desenvolvidas. Você **deve** fazer o seu trabalho seguindo esse esboço. Algumas tarefas do trabalho pedem implementações, e nesse arquivo temos a assinatura das funções, que devem ser usadas nessas implementações. Outras tarefas precisam de uma resposta discursiva, que também deverá estar nesse arquivo em formato de comentário da linguagem R e no local indicado nesse arquivo. Lembrem-se de colocar os nomes de todos os integrantes do grupo no começo do arquivo.

`trabalho1_base.R`: Neste arquivo, disponibilizamos implementações de algumas funções que facilitarão o desenvolvimento do trabalho.

`ranking_metrics.R`: Implementação das funções de avaliação de ranking.

`time.txt`: Arquivo com os artigos da revista TIME, que formam a nossa coleção de documentos.

`queries.txt`: Exemplos de consultas para buscar documentos em nossa coleção.

`relevance.csv`: Lista de vetores de *ground truth* para as consultas do arquivo `queries.txt`.

Sobre a Submissão do Trabalho

Prazo de entrega: 06 de Agosto de 2023 (domingo), até às 23h55.

Forma de entrega: via sistema Moodle. Apenas um integrante do grupo deve fazer a submissão do trabalho.

Atenção: Submeta um arquivo .zip, contendo o código R e todos os gráficos gerados durante a análise.

Pontuação: Este trabalho será pontuado de 0 a 10, e corresponderá a 30% da nota final.

Questão 1

A função `process_data` disponibilizada no arquivo `trabalho1_base.R` faz um processamento básico dos arquivos, similar ao exemplo visto em sala. O resultado é um `data.frame` com duas colunas: identificador do documento e termo. Use o `data.frame` gerado por essa função para processar a coleção de documentos e as consultas. Em seguida, crie uma matriz Termo-Documento com a coleção e calcule as estatísticas `tf-idf` e `bm25`. As estatísticas deverão ser convertidas para `data.frame`. Esses valores serão usados para computar os rankings. Lembre-se de configurar os parâmetros b e k, usados para o cálculo do `bm25`.

Questão 2

Nesta questão, iremos fazer consultas na coleção de documentos e analisar os rankings gerados por essas consultas.

(a) *Implementações:* No arquivo `inf0611_trabalho1.R`, temos um esboço detalhado de uma função que irá gerar um ranking e apresentará informações sobre esse ranking, para serem analisadas em seguida. A assinatura da função já é fornecida no arquivo, bem como uma lista de parâmetros e suas funcionalidades. A seguir, listamos as tarefas que essa função deve cumprir.

(i) Gerar um ranking usando a função `get_ranking_by_stats`. Essa função computa o ranking de uma consulta baseado nos valores de `tf-idf` ou `bm25`. Ela recebe como parâmetros o nome da estatística, um `data.frame` com as estatísticas e uma lista simples de tokens de uma consulta.

(ii) Calcular a precisão para o ranking gerado, considerando os $k = 20$ elementos do topo do ranking.

(iii) Calcular a revocação para o ranking gerado, considerando os $k = 20$ elementos do topo do ranking.

(iv) Apresentar um gráfico com a Precisão e a Revocação no eixo y, e os valores de 1 à k no eixo x, representando o topo do ranking.

(b) *Análise:* Após a implementação especificada acima, escolha duas consultas do arquivo `queries.txt`. Para cada consulta, crie um ranking usando a estatística `tf-idf` e outro com a estatística `bm25`. Compare os rankings de cada consulta e responda:

(i) Qual dos modelos teve o melhor resultado para as consultas escolhidas? Justifique sua resposta. Nesta questão você pode usar qualquer uma das medidas de avaliação de ranking vistas na Aula 1.

Questão 3

Nesta questão, iremos analisar o impacto das técnicas de processamento de texto.

(a) *Implementações:* Repita os passos da **Questão 1** modificando a chamada da função `process_data` para incluir a remoção de *stopwords*. Em seguida, aplique a função da **Questão 2** gerando novos rankings com as duas consultas escolhidas e as estatísticas `tf-idf` e `bm25`. Analise os novos rankings gerados e responda:

(i) Qual o impacto da remoção de stopwords na precisão tanto de `tf-idf` quanto de `bm25`? E na Revocação?