Trabalho 1 (Individual ou em Dupla)

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 83 exemplos de consultas com seus respectivos vetores de *groudthruth*.

Objetivo Geral do Trabalho

O objetivo deste trabalho é exercitar 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. O trabalho está dividido em duas partes: *Parte 1 – Calculando os* rankings e *Parte 2 – Comparando as técnicas*, que são detalhadas a seguir.

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 deste trabalho.

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 todos os arquivos disponibilizados e uma breve descrição sobre eles.

`inf0611_trabalho1.R`: 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 deverão estar nesse arquivo em formato de comentário da linguagem R e no local indicado nesse arquivo.

`trabalho1_base.R`: Neste arquivo disponibilizamos algumas implementações 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 groundthruth para as 83 consultas do arquivo `queries.txt`.

Parte 1 – Calculando os *rankings*

Questão 1

Use a função `process_data` disponibilizada no arquivo `trabalho1_base.R` para fazer o processamento do texto dos documentos e das consultas. Calcule o tf, idf, tf-idf e o bm25 para a coleção. Faça duas chamadas da função que efetua os cálculos das estatísticas com diferentes valores de k e b e salve os resultados em variáveis diferentes (evitando sobrescrita).

Questão 2

Escolha duas consultas do arquivo `querys.txt`. Faça as consultas com os métodos TF-IDF e BM25, usando todas as versões das estatísticas calculadas na Questão 1. Em seguida, compute a precisão e a revocação de cada método. Gere os gráficos de Precisão × Revocação. Qual dos modelos teve o melhor resultado para as consultas escolhidas? Justifique sua resposta.

Questão 3

Modifique a chamada da função `process_data` e aplique (ou remova) etapas no processamento do texto e repita os experimentos das Questões 1 e 2. Qual o impacto dessas alterações na precisão? E na Revocação?

Parte 2 – Comparando as técnicas

Na parte anterior, escolhidas duas consultas, comparamos os resultados de precisão e revocação dos modelos TF-IDF e BM25. Nesta etapa, esperamos comparar a performance desses modelos de maneira mais consistente, utilizando a média das precisões de **todas** as consultas disponíveis.

Questão 4

Compute a Precisão Média e a Média das Precisões Médias para os *rankings* gerados pelos modelos TF-IDF e BM25. Use as 83 consultas que possuem *groundthruth* para os dois modelos. Qual dos modelos teve o melhor resultado? Justifique sua resposta.

Sobre a Submisão do Trabalho

Prazo de entrega: 08 de Março de 2020 (Domingo), até às 23h55.

Forma de entrega: via sistema Moodle:

- `https://moodle.lab.ic.unicamp.br/moodle/course/view.php?id=394`

Apenas um integrante da dupla deve fazer a submissão do trabalho no Moodle.

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