Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Departamento de Ciências de Computação
SCC0282 - Recuperação de Informação

Trabalho Prático

Professor: Dr. Marcelo Garcia Manzato (mmanzato@icmc.usp.br)
Estagiário PAE: Luan S. de Souza (luanssouza@usp.br)
Divulgação: 17/05/2021
Data da Entrega: 07/06/2021 (parcial) - 26/07/2021 (final)

Objetivo

O objetivo geral do projeto é desenvolver um sistema de recuperação textual utilizando diferentes mecanismos de indexação apresentados em aula. Além disso, o projeto consiste em comparar as diferentes estratégias desenvolvidas por meio de métricas de avaliação também apresentadas na aula.

Metodologia

Será permitido utilizar APIs específicas para tratamento de texto, indexação, pesagem, etc., embora será de fundamental importância estar familiarizado com suas funcionalidades, de modo a ser possível implementar o sistema conforme os requisitos exigidos nesta especificação.

Os grupos terão liberdade para escolher o conjunto de dados para rodar os experimentos, sendo necessário, entretanto, que o *dataset* escolhido possua já um conjunto de *queries* bem definidas (*information needs*) e resultados relevantes (*ground truth*) a fim de se poder avaliar o sistema. Como exemplo de conjunto de dados que poderá ser utilizado, destacam-se os arquivos disponibilizados pelo *Forum for Information Retrieval Evaluation (FIRE)*, sendo que uma versão está disponível para download no Tidia.

Neste conjunto:

- **en.doc.2010** é uma pasta que contém todos os documentos recuperáveis, organizados em diferentes pastas. Será necessário indexar todos os arquivos.
- **en.topics.76-125.2010.txt** contém as *queries* e respectivas descrições das necessidades de informação que serão usadas para consultas ao sistema. Você pode usar quaisquer informações desse arquivo para representar sua consulta a ser submetida ao sistema.
- **en.qrel.76-125.2010.txt** contém os resultados relevantes para cada *query* (*ground truth*). Neste arquivo, repare que a primeira coluna representa o ID da consulta contida no arquivo de *queries*, a terceira coluna indica um dos documentos contidos na pasta **en.doc.2010**, e a última coluna indica se este documento é relevante (1) ou não (0). Para os experimentos, o que será de fato usado por você será apenas as relações (ID da *query*, documento, 1).

Requisitos

O projeto consiste em analisar os seguintes casos:

1. Comparação entre realizar indexação com e sem eliminação de *stopwords*;
2. Comparação entre realizar indexação utilizando ou não radicalização (*stemming*);
3. Comparação entre realizar e não realizar expansão das consultas;
4. Comparação entre os modelos vetorial e probabilístico.

Para comparar os resultados, os grupos deverão utilizar as métricas de avaliação dadas em aula: precisão (*precision*), revocação (*recall*), F-1, MAP (*Mean Average Precision*), etc. Não é obrigatória a utilização de todas as métricas, mas lembre-se: i) visualização por gráficos é mais fácil de se avaliar, tanto pelo autor quanto pelo leitor; e ii) as diferentes métricas avaliam diferentes aspectos do sistema, sendo que a análise crítica será mais completa se boa parte dos aspectos for considerada.

Entrega

Deverão ser entregues (via escaninho, no Tidia):

1. O código-fonte (link do Colab ou via .ipynb);
2. Um arquivo texto contendo as seguintes informações: i) integrantes do grupo; ii) *link* para base de dados utilizada; e, caso a entrega do código-fonte seja via Colab, iii) *link* para o código-fonte;
3. Um relatório de desempenho de até 5 páginas ou via Google Colabority (em Português ou Inglês), contendo as seguintes seções:
   (a) Título/autores/filiação/email (cabeçalho)
   (b) Introdução (contextualização, motivação e objetivo)
   (c) Técnicas Utilizadas (descrição das estratégias de recuperação utilizadas)
   (d) Avaliação (metodologia de avaliação, métricas utilizadas, base de dados utilizada, etc.)
   (e) Avaliação (metodologia de avaliação, métricas utilizadas, base de dados utilizada, etc.)
   (f) Resultados Obtidos (gráficos, análises, tabelas comparativas, etc.)
   (g) Considerações finais (conclusões sobre o projeto, sobre a disciplina, etc.)

   A entrega parcial irá compor 30% da nota do trabalho, onde serão avaliados:

   - Processamento da base e indexação (com/sem stopwords, com/sem radicalização)
   - Implementação do modelo vetorial e probabilístico
   - Expansão de consultas

   A entrega final irá compor 70% da nota do trabalho, onde serão avaliados:

   - Módulo de avaliação com métricas, graficos, etc.
   - Análises comparativas
   - Relatório

Observações finais

Os seguintes critérios de avaliação serão considerados durante a correção dos trabalhos:

1. Os alunos implementaram corretamente todas as funcionalidades exigidas na especificação?
2. O artigo entregue possui boa apresentação (escrita, formatação, organização e elaboração das seções, limite de páginas, etc.)?
3. O artigo entregue possui boa qualidade técnica (descrição correta das técnicas, argumentação válida, avaliação desenvolvida corretamente, etc.)?
4. Os alunos fizeram uma análise crítica dos casos estipulados? (**análise crítica**: não é só dizer que o caso A teve x% de melhoria em relação ao caso B; é necessário refletir e argumentar sobre o por quê isso ocorreu!).

Dúvidas durante o desenvolvimento do trabalho podem ser sanadas com o professor ou com o estagiário PAE por email.