PCC547 – Recuperação de Informação

Professor: Denilson Alves Pereira

Trabalho Prático 1

- Trabalho individual.
- Deve ser entregue em versão eletrônica pelo Campus Virtual (https://campusvirtual.ufla.br/). Envie somente arquivos texto sem formatação e pdf (não enviar .doc, .docx, .odt etc.). Arquivos compactados, somente .zip e .tar.gz (não enviar .rar, .z etc.). Não use acentos e nem "ç" nos nomes de arquivo.
- **Data limite de entrega: 03/10/2023 Apresentação: 05/10/2023 (15 minutos)**
- **Valor: 18 pontos**

O objetivo do trabalho prático é implementar e avaliar modelos de recuperação de informação usando a biblioteca de busca por similaridade FAISS (Facebook AI Similarity Search). A implementação deve ser feita de acordo com as seguintes instruções:

- Para avaliação do resultado, use a coleção de referência CFC (Cystic Fibrosis). Para o gabarito de resultado das consultas dessa coleção, desconsidere o *score* de relevância, apenas considere todos os documentos no resultado como relevantes. Para indexação dos documentos da coleção CFC, crie um único campo, concatenando os textos contidos nos atributos TI, MJ, MN e AB/EX. Faça o pré-processamento do texto, se necessário;
- Para vetorização dos documentos (e consultas), use duas codificações diferentes de embeddings: TF-IDF e Sentence Transformers (Sentence BERT, SBERT);
- Para indexação dos documentos, experimente o índice Flat e um outro a sua escolha, dentre aqueles disponíveis no FAISS;
- Para efetuar as consultas e obter seus resultados, use a similaridade do cosseno;
- Para avaliação do resultado, crie um módulo para retornar as métricas seguintes. O módulo deve receber como entrada a identificação das consultas com as respectivas identificações dos documentos de seu resultado, informados pelo gabarito da coleção de referência e pelo seu algoritmo, ordenados por relevância, e retornar os valores para as métricas Precisão e Revocação. O módulo deve gerar a tabela de Precisão x Revocação para 11 níveis de revocação, para que seja montado um único gráfico com os valores médios entre todas as consultas da coleção. Além disso, calcule também os valores para as métricas P@5 e P@10 médios, MRR(Q) considerando o *threshold* S_h = 5 e trace o histograma da precisão-R para as 20 primeiras consultas. Apresente também o tempo de execução do conjunto de consultas;
- Avalie cada codificação de embeddings separadamente em cada um dos tipos de índices escolhidos, apresente os seus respectivos gráficos de Precisão x Revocação e os resultados das demais métricas descritas acima.

O que deve ser entregue:

- O código fonte dos programas, devidamente comentados;
- Um relatório técnico contendo introdução, referencial teórico, descrição do trabalho com suas estratégias de solução, experimentos executados e resultados obtidos, conclusão e referências bibliográficas. O relatório deve ter de 6 a 8 páginas, de acordo com o *template* de artigos da SBC;
- Slides para apresentação em sala de aula.