2ª parte do Trabalho Prático Wikipédia (TP3) – Busca e Avaliação do Resultado

Nesta parte do trabalho prático, será feito a modelagem booleana, vetorial, BM25 e, ainda, a visualização da busca e avaliação do resultado através de coleções de referência. Para isso será necessário:

Busca
Para que seja realizada a busca no índice, o vocabulário deverá estar obrigatoriamente em memória primária. A lista de ocorrência pode estar em arquivo. Durante a busca, não será permitido busca através de banco de dados.

Modelagem de Busca
Nesta tarefa deverão ser implementadas três modelagens discutidas em sala: booleana, vetorial e BM25. No caso do modelo vetorial, o resultado da busca deve ser ordenado de acordo com a medida de distância implementada (distância do cosseno). O BM25 deve ser usado com os parâmetros: b = 0.75 e k =1. Como o modelo booleano não permite ordenação (ranking), vocês estão livres para apresentar o resultado na ordem que acharem mais conveniente.

Interface de interação
É necessário a implementação de uma interface simples (desktop) de interação com a máquina de busca implementada. Essa interface deve permitir ao usuário digitar qualquer consulta e selecionar qual modelagem de dados ele deseja utilizar na busca (booleana, vetorial ou BM25). Os resultados da busca devem ser apresentados ao usuário nessa mesma interface. O artigo da Wikipédia será representado no resultado através de seu título. Como você armazenou os ids dos documentos, para resgatar os títulos dos documentos para um determinado id utilize o arquivo "titlePerDoc.dat" disponibilizado no Moodle. Neste aquivo, para cada linha, você poderá resgatar para um id o seu respectivo título.

Lembre-se, para um rápido processamento da consulta, antes do usuário começar a interagir é necessário:
1- Carregamento completo do índice
2- Preprocessamento: (1) de valores necessários para o modelo vetorial e BM25; (2) dos documentos relevantes nas coleções de referência (apresentado a seguir); e (3) dos títulos por documentos, usados para apresentar o artigo para o usuário.

Logo após esses passos, poderá ser possível um usuário realizar consultas. Lembrando que a consulta deve passar pelo mesmo preprocessamento de texto dos artigos da Wikipédia. Durante a consulta, se ela for uma das consultas que possuem coleção de referencia (apresentados a seguir), mostrar o resultado da avaliação de qualidade da resposta à consulta – apresentados a seguir - (P@5, P@10, RECALL@5,..). Ao exibir o resultado, é interessante visualizarmos o título do documento, para isso, o arquivo titlePerDoc.dat deve ajudar. Ele mostra, por linha, o id do documento e seu respectivo título separados por ponto-e-vírgula. Não esqueça de preprocessar isso antes da consulta.

Avaliação
Por fim, é necessário que vocês avaliem a qualidade dos resultados retornados pelas implementações de vocês somente para a modelagem vetorial. Especificamente, são necessárias as seguintes métricas de qualidade:
- Precision (Precisão): @5, @10, @25, @50
- Recall (revocação): @5, @10, @25, @50
A precisão e a revocação será realizada através das coleções de referência especificadas no final deste documento.

Pontos extras
2,0 pontos extras: Utilização de interface web (ou interface grafica) para a busca ao invés da interface de caracteres em desktop. Só ganhará ponto extra o trabalho que estiver sido entregue completo de acordo com esta especificação.

Relatório

No relatório deverá constar a documentação do projeto, decisões de implementações, desafios e soluções propostas. Além disso:

- Explicação das classes principais bem como os métodos: Esse detalhamento é importante para que eu possa entender o código que foi criado. O código também deverá estar devidamente comentado e legível para facilitar o entendimento. Para facilitar o entendimento, uma sugestão é fazer um algoritmo em alto nível de como foi realizado o processamento e quando cada método no código foi chamado.

- Analise a ocorrência de termos no documento fazendo, no mínimo:
  - Quais são os 10 termos com maior e menor IDF da coleção? Com base nos termos de menor IDF, será que podiamos propor stopwords novas? Com base nesses termos, existe algo que você poderia melhorar no processamento?

  - Apresente o gráfico de frequência das palavras. Tais palavras devem ser ordenadas decrescentemente de acordo com a sua frequência. Note que a frequencia da palavra é a frequencia total do termo na coleção (Definido por F(termo) na aula sobre TF-IDF). De forma similar, faça o gráfico do IDF de cada palavra (veja o grafico similar na aula sobre TF-IDF). Mostre também, quais que tipos de palavras (específicas? Erros de processamento?) possuem TF alto, mediano e baixo. Faça o mesmo para IDF.

  - [opcional] Qual é o tamanho do vocabulario (número de palavras) quando usamos stemming e remoção de stopwords? e quando não usamos?

- Caso tenha feito alguma alteração na forma de indexação da primeira fase para esta, favor especificar e justificá-la
- Detalhamento de todas as fórmulas dos modelos de busca usados
- Durante a implementação dos modelos de busca (modelo vetorial e BM25), quais valores da fórmula podem ser preprocessados antes da consulta?
- Faça algumas consultas exemplos, mostrando o tempo de execução e o número de documentos retornados para cada consulta para cada modelo implementado.
- Apresentação e discussão dos gráficos de avaliação do modelo vetorial e do BM25. Indique alguma melhoria que poderia ser feita no modelo com o objetivo de aumentar a precisão. Também mostre o tempo de execução e o número de documentos retornados para cada consulta.
- Ainda analisando a avaliação: Pense na especificidade de cada termo das consultas teste “Belo Horizonte”, “São Paulo” e “Irlanda”. O que se pode falar sobre a relação entre especificidade do termo e a precisão da consulta? Justifique.

Definição de 'coleções de referências' para o projeto

Para o projeto realizaremos uma avaliação bem simples, com o único intuito de simularmos um processo real de avaliação. Para tanto, consideraremos como conjunto de consultas de teste apenas três consultas:

'Irlanda'
'Belo Horizonte'
'São Paulo'

O conjunto de documentos de teste compreenderá todas as páginas da base de dados da Wikipédia PT_BR utilizadas no projeto. Porém você já tem, em alguns arquivos disponibilizado para vocês, o id de documentos relevantes (separados por vírgula) para as consultas teste.

Por exemplo, um documento D será considerado relevante para a consulta 'Belo Horizonte' somente se D o id de D estiver no arquivo “Belo Horizonte.dat”. Não esqueça de armazenar o conteúdo desses arquivos em memória para diminuir o tempo de busca. Feito isso, a coleção de referência para as três consultas estará montada e pode-se realizar os cálculos de avaliação corretamente.

Para os documentos relevantes para a consulta 'Irlanda' (Arquivo Irlanda.dat): Foram considerados relevantes artigos da seguintes categorias:

- Irlanda
- Economia da Irlanda
- História da Irlanda
- Cultura da Irlanda
- Romancistas da Irlanda
- Físicos da Irlanda
- Reis da Irlanda
- Lordes da Irlanda

Categorias relevantes para a consulta 'Belo Horizonte' (Arquivo Belo Horizonte.dat): Foram considerados relevantes artigos das seguintes categorias:

- Bairros de Belo Horizonte
- Bandas de Belo Horizonte
- Belo Horizonte
- Edifícios de Belo Horizonte
- Metrô de Belo Horizonte
- Naturais de Belo Horizonte
- Prefeitos de Belo Horizonte
- Vereadores de Belo Horizonte

Categorias relevantes para a consulta 'São Paulo' (arquivo São Paulo.dat). Foram considerados relevates artigos da seguinte categoria:

- Atrações turísticas da cidade de São Paulo
- Áreas protegidas de São Paulo
- Prefeitos de São Paulo
- São Paulo
- Turismo em São Paulo
- Universidades de São Paulo
- Rodovias de São Paulo
- Museus da cidade de São Paulo
- Governadores de São Paulo
- Municípios de São Paulo