Recuperação de Informação na Web 2017/2

Máquina de Busca para a Wikipédia (Português)

Contexto

Você foi contratado para construir uma máquina de busca destinada a Wikipedia em português. A Wikipedia vem recebendo diversas reclamações de seus usuários do Brasil e de Portugual que a busca interna do site é limitada e de baixa qualidade. Dessa forma, muito deles realizam a busca via máquinas de busca externas, tais como o Google, que também coletam e indexam as páginas da Wikipédia. O risco neste caso é que os usuários da Wikipédia tenham acesso a outra fonte de informação via estas máquinas de busca e, ao longo do tempo, a Wikipédia perca parte de seus usuários do Brasil e de Portugal. A empresa, dessa forma, tem por objetivo aumentar a fidelidade desses usuários ao site da Wikipédia. E a forma que decidiram começar esse processo é através da melhoria de sua máquina de busca interna.

Sobre a 'Equipe'

O projeto deve ser desenvolvido por grupos de no máximo quatro pessoas pessoas. Caso o grupo seja acima de 4, será retirado 2 pontos por aluno excedente.

Tarefas

A Wikipédia se comprometeu apenas em lhe entregar uma amostra de 61.127 documentos do site em português em HTML. Assim, você não precisa se preocupar com a tarefa de coleta. Porém, eles exigem de você a implementação de todos os demais passos necessários para a recuperação de informação, a saber:

- Extração das informações do Documento HTML
- Pré-processamento do texto
- Indexação
- Modelagem
- Processamento de consultas
- Interface simples de interação com seu sistema

Além disso, eles exigem que você realize uma ampla análise de qualidade do modelo proposto, aplicando as métricas tradicionais de avaliação de qualidade estabelecidas pela área de recuperação de informação. Todas as decisões e resultados devem ser devidamente documentadas, visto que a decisão sobre o uso ou não de sua proposta será tomada considerando-se o seu relatório, bem como a versão final do código. A implementação deverá ser na liguagem de sua escolha.

Entregas

Dado o tamanho do projeto contratado, a Wikipédia determinou que o ideal seria dividir a entrega do projeto em duas partes:

- Primeira parte:
  - Data da entrega: 23 de outubro as 23:59 (tolerância de 6 horas)
  - Valor: 12 pontos
  - Tarefas a serem entregues:
    - Extração do conteúdo das páginas HTML
    - Preprocessamento completo do conteúdo
    - Geração dos índices invertidos
- Segunda parte
  - Data de entrega: 04 de dezembro as 23:59 (tolerância de 6 horas)
  - Valor: 15 pontos
  - Tarefas a serem entregues :
    - Modelagem booleana, vetorial e BM25
    - Estratégia de raking dos resultados
    - Interface de interação e consultas ilustrativas
    - Avaliação de qualidade sobre as consultas ilustrativas

Política de atraso
Será descontado 1 ponto por dia de atraso. Após dia 06 de dezemobro não será mais aceito a segunda parte do trabalho.

Sobre os dados
Esta base contém 61.127 documentos. Cada artigo da Wikipedia tem seu titulo e seu id. A estrutura de diretorios foi criada a partir dos primeiros 3 números do id do artigo. Por exemplo, páginas com ids que começam com "145" estarão no diretório "145" cada página será um arquivo html diferente nomeado com o seu id. Foi feito dessa forma para facilitar a abertura dos diretórios – geralmente, ter um diretório com milhares de arquivos, dificulta a abertura do mesmo.

A base de dados pode ser encontrada em: http://goo.gl/CBtEYc

Sobre cada Tarefa

Extração das Informações

Dado que os artigos da Wikipédia estão em uma estrutura de diretórios, você deverá navegar em todos esses diretórios para assim, considerar todos estes artigos durante a busca.

Preprocessamento
Especificamente, a Wikipédia especificou os seguintes passos de pré-processamento como obrigatórios:
1. Limpeza dos Dados
2. Análise Léxica/Transformações
3. Eliminação de Stopwords
4. Stemming

Para as tarefas acima, a equipe pode usar ferramentas externas desde que, na documentação, explique detalhadamente o funcionamento das mesmas (por exemplo, ao usar uma ferramenta para fazer stemming, qual foi a técnica utilizada para fazer stemming através dessa ferramenta?). Lembrando algumas ferramentas de stemming, que podem ser utilizadas nessa tarefa:

PTStemer
Stemka
snowball

Indexação

Considere um conjunto de documentos. A cada documento é atribuído um conjunto de palavras-chave ou atributos. Um arquivo invertido é constituído de uma lista ordenada de palavras-chave, onde cada palavra-chave tem uma lista de apontadores para os documentos que contêm aquela palavra-chave. Este é o tipo de índice utilizado pela maioria dos sistemas para recuperação em arquivos constituídos de texto. A estrutura de dados a ser implementada deverá ser constituída do vocabulário do texto, incluindo o número de documentos associados com cada palavra-chave e uma lista de ocorrências da palavra na coleção de documentos. Cada entrada da lista indica o número do documento onde a palavra ocorreu e o número de ocorrências. Após a criação do indice, você deverá armazena-lo em arquivo texto ou binário para posterior utilização nas consultas. Não é permitido o uso de banco de dados para a tarefa nem utilização de bibliotecas externas que indexam o documento.

As listas invertidas estarão sempre em memória principal. Há um guia no moodle para sugestões de implementações da estrutura do índice. Inclusive dicas para estruturas desse índice quando o computador não tiver muita memória. Você ira fazer uma pratica avaliativa para a criaçao dos indices invertidos.

Neste trabalho não é permitido uso de banco de dados para armazenar/coletar o indice. O indice deve ser criado em memoria principal e gravado em arquivo texto ou binario.

Modelagem de dados

Nesta tarefa deverão ser implementadas as versões mais robustas de duas modelagens discutidas em sala: booleana, vetorial e BM25. É importante que cada modelagem seja implementada com todas as correspondentes otimizações discutidas (e..g., soluções para evitar que documentos grandes sejam beneficiados). No caso do modelo vetorial e o BM25, o resultado da busca deve ser ordenado de acordo com a medida de distância implementada. Como o modelo booleano não permite ordenação (ranking), vocês estão livres para apresentar o resultado na ordem que acharem mais conveniente.

Interface de interação e consultas ilustrativas

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.

Avaliação de qualidade das buscas

Por fim, é necessário que vocês avaliem a qualidade dos resultados retornados pelas implementações de vocês para a modelagem vetorial e BM25. Especificamente, são necessárias as seguintes métricas de qualidade:
- Precision: @5, @10, @25, @50
- Recall: @5, @10, @25, @50

Reparem que o processo de avaliação requer primeiramete a consolidação das chamadas coleções de referência (i.e., conjunto pequeno de consultas para as quais sabemos todos os documentos relevantes). Informações sobre como gerar tais coleções de referência estão no Moodle

Sobre a avaliação do projeto
O projeto será avaliado a partir das listagens dos programas, da documentação entregue e do resultado da execução. Apresente uma boa documentação do trabalho, contendo pelo menos os seguintes itens: saída legível mostrando o funcionamento do código, comentários explicativos sobre os algoritmos e estruturas de dados, discussão sobre cada decisão de implementação adotada, explicação das ferramentas utilizadas, resultados experimentais medindo tempos de cada tarefa e análise dos resultados. Será disponibilizado um guia no Moodle com as perguntas que devem ser respondidas no relatório de cada tarefa.

Formas de Entrega
Os trabalhos devem ser entregues no Moodle. A entrega consiste em um único arquivo compactado que contém o relatório, códigos, resultados e avaliações referentes a cada tarefa do projeto. Não colocar o código dentro do relatório e sim em um arquivo separado. Além disso, gere o relatório em formato pdf. Trabalhos fora desta especificação perderão 2 pontos.