

Problema 3: Motor de busca FeiraGugou

Descrição do Problema:

O Google Search é um serviço da empresa Google através do qual é possível fazer pesquisas na Internet sobre qualquer tipo de assunto ou conteúdo. Lançado pela Google Inc em setembro de 1997, é atualmente o serviço de busca on line mais utilizado em todo mundo. A popularidade do serviço é tão alta que é comum que Internautas utilizem a expressão "googlar" em conversas informais, um neologismo que significa executar uma pesquisa na Internet pelo motor de busca Google.

Neste problema, você e sua dupla têm como objetivo criar um mecanismo de pesquisa para uma parte limitada da "web", com alguns dos recursos dos mecanismos de pesquisa comuns, como o Google Search. Um dos passos é poder analisar as páginas da "web" com base em seu conteúdo. Isso nos permitirá avaliar a relevância de uma determinada página para uma determinada solicitação do usuário.

A palavra "web" está escrita entre aspas, pois as páginas a serem identificadas não estarão, de fato, disponíveis na world wide web. Uma página será representada por um arquivo armazenado em disco e seu conteúdo será textual.

O conteúdo de uma página está obviamente correlacionado com as palavras que ela contém e uma página é listada como resultado de busca iniciada pelo usuário a partir de uma palavra-chave, existente naquela página. Para a tarefa de encontrar as páginas a partir da palavra-chave inserida pelo usuário, vocês deverão utilizar uma estrutura de dados com complexidade de $O(\log n)$ para as operações de busca, inserção e remoção.

Na tabela 1 (UserStories), são apresentados os requisitos do sistema de busca que deverão desenvolver.

Tabela 1 - UserStories:

UserStory	Título	Breve Descrição
1	Buscar Páginas	A partir de palavra-chave inserida pelo usuário, o sistema deve apresentar todas as páginas que possuem tal palavra em seu conteúdo. Por padrão, as páginas devem ser apresentadas da mais relevante para a menos relevante (considere como mais relevante a página que contiver a maior frequência da palavra-chave).
2	Ordenar Resultados da Busca	É possível ordenar as páginas encontradas (por relevância crescente ou decrescente).
3	Top-K palavras mais buscadas	O sistema deve apresentar as k palavras mais buscadas pelos usuários.
4	Top-K palavras menos buscadas	O sistema deve apresentar as k palavras menos buscadas pelos usuários.
5	Top-K páginas mais visitadas	O sistema deve apresentar as k páginas mais visitadas pelos usuários.
6	Top-K páginas menos visitadas	O sistema deve apresentar as k páginas mais visitadas pelos usuários.

Como resultado das buscas, as páginas (arquivos em discos), são representadas pelos seus nomes. Caso o usuário deseje visitar a página para conhecer o seu conteúdo, é possível – basta indicar ao sistema, que deve retornar o conteúdo da página selecionada. Outra informação importante: páginas podem ser adicionadas, editadas ou removidas durante a execução do programa.

É fundamental que as duplas criem o repositório de páginas, considerando, pelo menos, que o usuário busque por quinze palavras-chave diferentes e, para cada uma delas, existam pelo menos cinco arquivos que contenham tal palavra em seu conteúdo.

É possível que vocês estejam ansiosos para desenvolverem, além das user stories solicitadas, uma interface gráfica para o sistema de busca FeiraGugou. E, então, vamos aproveitar esta motivação possibilitando uma pontuação bônus de até 2,0 pontos (a serem acrescidos à nota do código) para aqueles que criarem uma interface gráfica de boa usabilidade. Só não vale utilizar GUI Builders - serão aceitas interfaces desenvolvidas exclusivamente pelos estudantes.

Produto:

Você deve enviar um e-mail com o produto final para o seu tutor até 12 horas do dia 29/06 (segunda-feira), anexando o arquivo compactado com o código-fonte e os testes de unidade. Deve ser desenvolvido tratamento das possíveis exceções. O código fonte deve estar todo documentado utilizando o padrão javadoc. Também deverá ser entregue um relatório técnico Individual escrito no formato de artigo da SBC. Deverá ter entre quatro e seis páginas até o dia 29/06 (até 12h), contendo:

- Uma introdução com a apresentação resumida da motivação, contexto e problema;
- A fundamentação teórica;
- A metodologia utilizada;
- Resultados e discussão (com a descrição detalhada da sua solução, declarando as decisões que você teve de tomar e fundamentando-as, sempre que possível);
- As suas conclusões (com as lições aprendidas e os pontos positivos e negativos deste projeto),

sendo facultativo ao tutor o pedido de relatório Impresso, entregue em sala, até 29/06.

Avaliação:

O código entregue corresponde a 40% da nota da unidade, o relatório corresponde a 30% e o desempenho nos tutoriais corresponde a 30% da nota.

Calendário:

Aula	Data	Atividade
1-2	05/06	Apresentação do Problema 3
3-4	08/06	PBL
5-6	12/06	PBL
7-8	15/06	PBL
9-10	19/06	PBL
11-12	22/06	PBL
13-14	26/06	PBL
	29/06	Entrega do Código do Problema 3 Entrega do Relatório do Problema 3 Apresentação do Problema 4