TRABALHO FINAL
Sugestão de Consultas em Motores de Busca

1 Objetivo

Este trabalho tem por objetivo proporcionar aos alunos a oportunidade de aplicar os conhecimentos adquiridos e as estruturas de dados desenvolvidas em aula na solução de um problema que utilize várias dessas estruturas. O trabalho prático envolve a utilização de estruturas de dados do tipo listas e árvores.

2 Especificação da Aplicação

A funcionalidade de autocompletar é muito usada em motores de busca na Web, serviços de mensagem de celular e até mesmo em editores de código. Ele consiste em tentar prever quais serão os próximos caracteres ou palavras que o usuário irá digitar, à medida que o ele digita. Em motores de busca, essas sugestões são feitas com base em uma série de indicadores, por exemplo: frequência em que a consulta é realizada, taxa de coocorrência das palavras em um corpus, aspectos de localidade, etc.

A tarefa a ser desenvolvida como trabalho final da disciplina simula, de forma bastante simplificada, o funcionamento de ***um mecanismo de sugestão de palavras para consultas em motores de busca***. A aplicação desenvolvida deverá ser composta por dois módulos: (i) *geração de estatísticas* e (ii) *consulta*. Durante a fase de geração de estatísticas, as palavras do texto dado como entrada serão carregadas em uma estrutura de dados (a ser proposta por você). Durante a fase de consultas, a aplicação irá ler um arquivo com as palavras a serem consultadas, fazer buscas na estrutura de dados e gerar um arquivo de saída com as palavras sugeridas.

3 Definições:

Uma *consulta* é composta por uma série de palavras. Uma *palavra* é uma sequência de letras. Todos os outros caracteres (números, espaço, pontuação, quebra de linha, etc.) deverão ser considerados como separadores de palavras. Diferenças entre letras maiúsculas e minúsculas devem ser desprezadas (ex: a = A).

Um *corpus* é uma coleção de textos. O corpus será fornecido para possibilitar o cálculo das estatísticas de associação entre as palavras (isto é, palavras que aparecem consecutivamente). Quanto mais duas palavras aparecerem juntas, mais chance a segunda palavra tem de ser a próxima, caso a primeira seja digitada.

A medida estatística de associação entre palavras a ser usada é:

$$\frac{freq(a,b)}{\sqrt{freq(a) * freq(b)}}$$

Onde *freq(a)* é o número de ocorrências da palavra *a* no corpus, *freq(b)* é o número de ocorrências da palavra *b* no corpus e *freq(a,b)* é o número de vezes que a palavra *a* é seguida pela palavra *b*.

- Fase de geração de estatísticas:
  - Entrada: corpus (arquivo txt)
  - Saída: estatísticas armazenadas na estrutura de dados
- Fase de consultas
  - Entradas: (*i*) arquivo texto com as palavras a serem consultadas (uma palavra por linha) e (*ii*) número de sugestões desejadas

- Saídas: (*i*) arquivo com a palavra consultada e as palavras seguintes sugeridas *em ordem decrescente de associação*; e (*ii*) tempo gasto na busca.

4 Requisitos

- Escolher/propor estruturas de dados adequadas utilizando listas e ou árvores.
- Redigir um relatório que explique como a aplicação proposta funciona e defenda as estruturas de dados propostas.
- A aplicação deve ser chamada **a partir da linha de comando** (passando parâmetros para o main). Por exemplo, o comando

  `C:\minhaaplicacao texto1.txt consulta.txt saida.txt 5`

  significa que é necessário gerar as estatísticas a partir do arquivo com o corpus de nome texto1.txt e a seguir processar as consultas do arquivo consulta.txt sugerindo 5 possibilidades. O resultado será armazenado no arquivo saida.txt.
- Não há limites para o tamanho do texto e para o número de consultas.
- O trabalho deve ser feito, preferencialmente, em duplas. A linguagem de programação aceita é C (Não é C++ nem C#).

5. Exemplo de funcionamento

Entrada (corpus): teste.txt

```
A Lua encontra-se em rotação sincronizada com a Terra, mostrando sempre a mesma face
visível, marcada por mares vulcânicos escuros entre montanhas cristalinas e
proeminentes crateras de impacto. É o mais brilhante objeto no céu a seguir ao Sol,
embora a sua superfície seja na realidade escura, com uma refletância pouco acima da
do asfalto. A sua proeminência no céu e o seu ciclo regular de fases tornaram a Lua,
desde a antiguidade, uma importante referência cultural na língua, em calendários, na
arte e na mitologia. A influência da gravidade da Lua está na origem das marés
oceânicas e ao aumento do dia sideral da Terra. A sua atual distância orbital, cerca
de trinta vezes o diâmetro da Terra, faz com que no céu o satélite pareça ter o mesmo
tamanho do Sol, permitindo-lhe cobri-lo por completo durante um eclipse solar total.
```

Entrada: consulta.txt

```
a
da
```

Comando: `C:\minhaaplicacao teste.txt consulta.txt saida.txt 2`

Saída: `saida.txt`

```
Consulta: a
Sugestão: sua           (0.547722558)
Sugestão: lua           (0.365148372)

Consulta: da
Sugestão: terra         (0.516397779)
Sugestão: gravidade     (0.447213595)

Tempo: 0.01ms
```

6. Entrega e Apresentação

- **1 de dezembro de 2015** apresentação (horário da aula) e entrega pelo Moodle
  - 10% de bônus: ($Nota_{Trabalho} = Nota_{Trabalho} + Nota_{Trabalho} * 10/100$)
- **8 de dezembro de 2015** apresentação (horário da aula) e entrega pelo Moodle

7. Critérios de Avaliação

O trabalho deve ser realizado em duplas e deverá ser apresentado e defendido na data prevista. Para a avaliação serão adotados diversos critérios:

- funcionamento,
- organização e documentação do código.
- tempo gasto no processamento das consultas;
- justificativa para escolha das estruturas de dados envolvidas; e
- relatório;

A escolha das estruturas de dados deve demonstrar conhecimento teórico e prático buscando a melhor combinação que atinja os resultados satisfatoriamente. Esse trabalho não avalia apenas o desempenho, mas a capacidade do aluno de criar estruturas elegantes e fáceis de serem mantidas. Para avaliar esse critério, é muito importante que o aluno DESCREVA COM RIQUEZA DE DETALHES as estruturas utilizadas no programa.

Importante:

Este trabalho deverá representar a solução da dupla para o problema proposto. O plágio é terminantemente proibido e a sua detecção incorrerá na divisão da nota obtida pelo número de alunos envolvidos. Para detectar o plágio, usaremos o software MOSS (http://theory.stanford.edu/~aiken/moss/).