INSTITUTO SUPERIOR MIGUEL TORGA

MOMENTO DE AVALIAÇÃO

Curso(s): Licenciatura em Informática

Unidade Curricular: Estrutura de Dados e Algoritmos

Avaliação Continuada

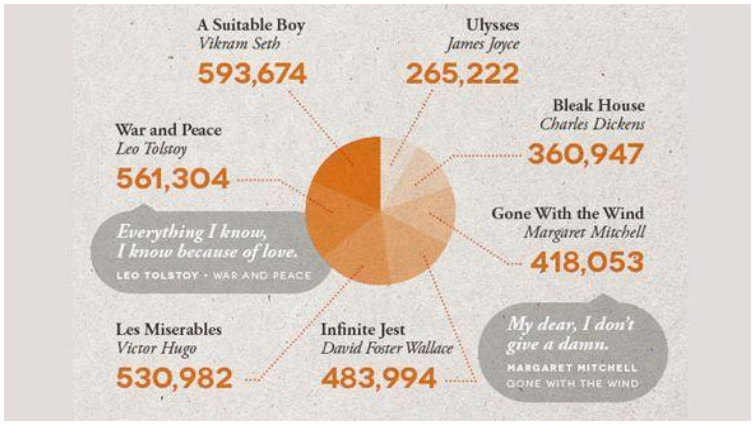
Docente: André Monteiro

Duração: Até 08-01-2020 Data: 14-11-2019

Observações: O objetivo do projeto final é desenvolver e aplicar algumas das técnicas de conceção e análise de algoritmos abordadas nas aulas aplicadas à resolução de um problema real de biométrica.

CONTEXTO

Uma das vertentes mais importantes da escrita comercial é a contagem de palavras. A contagem de palavras é nada mais do que o comprimento de uma peça, seja uma novela, uma história, uma publicação de blog, um artigo de revista ou uma brochura de vendas. O tipo de trabalho é importante porque a contagem de palavras é afetada por ele. Por exemplo, um romance sempre será mais longo (80.000 palavras) do que um artigo de revista (geralmente 1000 palavras).



Da mesma forma, alguns tipos de romances serão mais longos do que outros. Uma história curta geralmente é mais de 1000 palavras, mas também menos de 20.000 palavras. Portanto, há histórias curtas de contagem de palavras variando entre eles. E quanto a aventuras? Existem livros de 50.000 e são basicamente os mais curtos. Existem livros de 80.000, 100.000 e até 120.000. *Harry Potter e a Ordem da Fénix* tem 257.000 palavras.

Existem várias maneiras de analisar o texto, podemos:

- Contar a ocorrência de letras, palavras ou frases específicas, muitas vezes resumidas como nuvens de palavras;
- Categorizar as palavras por temas-chave, tópicos ou pontos comuns (Text Mining)
- Classificar as atitudes, as emoções e as opiniões de uma fonte em relação a algum tópico, chamado Análise de Sentimentos; existem muitas aplicações de análise de sentimentos em negócios, marketing, gestão de clientes, ciências políticas, direito, sociologia, psicologia e comunicações;
- Explorar as relações entre palavras usando o WordNet. As relações podem refletir definições ou outros pontos comuns.

Suponhamos que uma editora quer saber exatamente se os conteúdos dos livros que vende têm relação com os autores e leitores. O objetivo é registar e contabilizar as palavras utilizadas, sendo devolvida uma lista das "n" potenciais palavras da mesma família, para avaliar os autores e o mercado.

O trabalho, a implementar em linguagem Java, consiste na seleção e leitura do ficheiro de texto do livro introduzido, carregamento em memória, estatísticas de contagem e pesquisa dos elementos mais parecidos a uma palavra e devolução dos "top-n matches":

- n=1, top-1: o mais parecido
- n=2, top-2: os dois mais parecidos

Exemplo: palavra introduzida "salty", com n=1 a palavra "salt" e com n=2 a palavra "sal"

Exemplo de implementação:

```
//////////////////////////////////////
// Programa de gestão de palavras     //
// 1 – Ler ficheiro de livro          //
// 2 – Mostrar estatística de contagem//
// 3 – Pesquisar palavra algoritmo1   //
// 4 – Pesquisar palavra algoritmo2   //
// 5 – Ordenar livro algoritmo1       //
// 6 – Ordenar livro algoritmo2       //
// 0 – Sair                           //
// Opção?
> 1
> Livro a ler? Dracula-BranStoker.txt
> Livro selecionado corretamente!
> 2
//////////////////////////////////////
// #  |  Palavra         | Ocorrências//
// 1  |  the             | 332        //
// 2  |  Dracula         | 211        //
...
// 123|  Portugal        |   1        //
//////////////////////////////////////
> 3
> Indique a palavra (ou parte) a pesquisar:
> saltam
> Introduza o top-n desejado:
> 2
> As palavras mais prováveis são:
saltam, Distancia 0, ocorrências 77
salt, Distancia 1, ocorrências 2
sal, Distancia 2, ocorrências 11
Demorou 3600ms
```

PROJETO FINAL

Dentro deste contexto, o objetivo do projeto final é utilizar algoritmos que pesquisem padrões específicos dentro de texto e forneçam as palavras correspondentes e os n-elementos mais parecidos da lista. Também é necessária a ordenação, uma experiência puramente académica pois a ordenação corrompe o conteúdo do livro.

O projeto final será composto de três partes:

1. Elaboração de uma proposta de projeto com uma breve descrição dos algoritmos utilizados;
2. Implementação, teste e validação dos algoritmos;
3. Redação de um relatório sucinto que descreva o problema, algoritmos usados, análise dos algoritmos usados e resultados experimentais incluindo o tempo de execução e gráficos.

Como o propósito principal deste projeto é verificar a eficiência de algoritmos, deverão ser implementados pelo menos dois métodos, sendo um considerado "menos bom", ou seja, com um tempo de execução não adequado à magnitude do problema tratado e um outro método que forneça "o melhor" tempo de execução possível.

DADOS

Para o desenvolvimento do trabalho, são fornecidos vários livros teste (devem ser utilizados os .txt) que também estão disponíveis online em http://www.gutenberg.org/ebooks/search/?sort_order=downloads, como por exemplo

- Dracula, Bram Stoker -> http://www.gutenberg.org/cache/epub/345/pg345.txt
  - Linhas= 15974
- Pride and Prejudice, Jane Austen, -> http://www.gutenberg.org/files/1342/1342-0.txt
  - Linhas= 13428

RECOMENDAÇÕES

O projeto final deverá ser elaborado individualmente. Pode utilizar e reutilizar código, algoritmos, funções, bibliotecas e classes de outras fontes desde que sejam referenciadas. Deve pesquisar um algoritmo que resolva seu problema, analisá-lo (funcionamento e tempo de execução) e então implementá-lo utilizando a linguagem de programação Java.

IMPLEMENTAÇÃO

Uma vez definidos os algoritmos a serem utilizados, deverá implementá-los utilizando a linguagem de programação Java.

O programa deverá ler um livro, em *txt*, de tamanho qualquer de um ficheiro cujo padrão seja similar ao dos ficheiros fornecidos e receber do teclado um **n**, cuja presença se vai verificar as palavras apenas diferem em **n** elementos. Deve elaborar uma estatística das palavras utilizadas, ordenadas por ocorrências. Além disso, deve ter a opção de escolher entre o algoritmo "mau" e o "bom".

No ecrã de saída deve ser indicado:

1. as top-**n** palavras mais parecidas com a introduzida, na pesquisa;
2. tempo de execução, ambos para a pesquisa e ordenação.