INF01124 - Classificação e Pesquisa de Dados - Exercício 3
Professor João Comba

1 Ordenação de Strings

Neste laboratório iremos implementar um algoritmo para ordenação de strings e usá-lo para responder tarefas de análise de strings em um arquivo contendo obras literárias. As tarefas são descritas abaixo.

1.1 Implementar Radix Sort MSD para strings

Implemente o Radix Sort MSD para ordenação de strings como visto em aula. O algoritmo deve receber como entrada um vetor de strings e um número contendo o tamanho de entrada, e gerar como saída o vetor de strings ordenado em ordem lexicográfica (crescente).

1.2 Ordenar palavras de um arquivo de entrada usando Radix Sort MSD

Esta tarefa irá usar arquivos de entrada.txt disponibilizados no arquivos entradas.zip. no Moodle para validar o algoritmo Radix Sort MSD. Ao testar um dado arquivo .txt, faça a leitura do mesmo e armazene-o em um vetor de strings compatível com o algoritmo Radix Sort MSD implementando na etapa anterior. Inicialmente, teste o algoritmo Radix Sort MSD com os arquivos $test1.txt$, $test2.txt$, $test3.txt$ e $test4.txt$, verificando a correção do algoritmo.

Após validar o algoritmo, teste dois arquivos de duas obras literárias disponibilizadas publicamente, *Frankenstein*[1] e *War and Peace* [2]. Ordene os arquivos de texto $frankenstein.txt$ e $war\_and\_peace.txt$, e gere os arquivos $frankenstein\_sorted.txt$ e $war\_and\_peace\_sorted.txt$, contendo em cada arquivo as palavras ordenadas em ordem lexicográfica, uma palavra por linha. Um exemplo de saída do arquivo $frankenstein\_sorted.txt$ é listado abaixo:

```
...
ACCUSTOMED
ACCUSTOMED
ACCUSTOMED
ACCUSTOMED
ACCUSTOMED
ACCUSTOMED
ACCUSTOMED
ACCUSTOMED
ACCUSTOMED
ACCUSTOMED
ACHIEVE
ACHIEVED
ACHIEVEMENTS
ACHING
ACKNOWLEDGED
ACME
ACORNS
ACORNS
ACQUAINTANCE
ACQUAINTANCES
ACQUAINTED
...
```

[1] http://www.gutenberg.org/files/84/84-0.txt
[2] http://www.gutenberg.org/files/2600/2600-0.txt

1.3 Contar as palavras do arquivo ordenado

Usando os arquivos *frankenstein_sorted.txt* e *war_and_peace_sorted.txt* da etapa anterior, conte quantas vezes cada palavra acontece, e imprima em ordem cada palavra seguida de sua contagem. Gere os arquivos *frankenstein_counted.txt* e *war_and_peace_counted.txt*.

A saída deve seguir o formato:

```
palavra1 #ocorrencias_1
palavra2 #ocorrencias_2
palavra3 #ocorrencias_3
...
```

Por exemplo, para War and Peace seria assim:

```
A 10584
AAH 1
AB 1
ABACK 3
ABACUS 1
ABANDON 25
ABANDONED 54
ABANDONING 26
ABANDONMENT 14
ABANDONS 1
ABASEMENT 1
ABASHED 13
ABATE 2
ABB 19
ABBREVIATIONS 1
ABC 1
ABDICATE 1
ABDOMEN 2
ABDOMENS 2
ABDUCTION 3
ABDUCTORS 1
ABHORRENCE 1
ABIDE 1
...
```

1.4 Top 1000 palavras mais frequentes

Usando os arquivos *frankenstein_counted.txt* e *war_and_peace_counted.txt* da etapa anterior, gere um ranking com as 1000 palavras mais frequentes de cada livro. Este ranking deve ordenar as palavras das mais frequentes para as menos frequentes. Para duas palavras de mesmo número de ocorrências, imprima elas no ranking usando a ordem lexicográfica. Gere os arquivos *frankenstein_ranked.txt* e *war_and_peace_ranked.txt*,

A saída deve seguir o formato:

```
palavra1 #ocorrencias_1
palavra2 #ocorrencias_2
palavra3 #ocorrencias_3
...
```

Por exemplo, para War and Peace seria assim:

```
THE 34725
AND 22307
TO 16755
OF 15008
A 10584
HE 10007
IN 9036
THAT 8206
```

```
HIS 7984
WAS 7360
WITH 5710
IT 5617
HAD 5365
S 5200
HER 4725
NOT 4697
HIM 4637
AT 4547
I 4541
BUT 4056
AS 4035
ON 4022
YOU 3871
FOR 3555
SHE 3489
IS 3347
SAID 2842
ALL 2813
FROM 2709
BY 2458
...
VISIT 64
WOOD 64
ANIMATED 63
BOOTS 63
DEVIL 63
EXCITED 63
INSTANT 63
LEARNED 63
LIFTED 63
ONTO 63
VISITORS 63
WISHING 63
BENT 62
EXPLAIN 62
EXPRESS 62
MARSHAL 62
NEEDED 62
REGARD 62
SKY 62
SUN 62
```

2 Entrega

A solução deve ser enviada pelo Moodle dentro de um arquivo .zip, contendo os seguintes arquivos:

- **integrantes.txt**: coloque o nome dos integrantes do grupo (até 2 pessoas) , com um nome por linha
- $frankenstein\_sorted.txt$: lista ordenada de palavras e suas ocorrências, no exato formato listado acima
- $war\_and\_peace\_sorted.txt$**t**: lista ordenada de palavras e suas ocorrências, no exato formato listado acima
- $frankenstein\_counted.txt$: lista ordenada de palavras e suas ocorrências, no exato formato listado acima
- $war\_and\_peace\_counted.txt$**t**: lista ordenada de palavras e suas ocorrências, no exato formato listado acima
- $frankenstein\_ranked.txt$: lista ordenada de palavras e suas ocorrências, no exato formato listado acima
- $war\_and\_peace\_ranked.txt$**t**: lista ordenada de palavras e suas ocorrências, no exato formato listado acima
- código fonte correspondente a solução