Peso na nota final:	**5 valores em 20**
Informações:	Moodle
Data de entrega	**2021.04.19/ 00h01'** (atenção à hora)
Publicação de nota	**2021.05.18**
Versão	**1.03 (2021.03.16/21h57')**

Sistemas Operativos / 2S 2020-21

Projeto – script bash *wordStats.sh*

1 - Introdução

Pretende-se que elabore o *script* **`wordStats.sh`**. O *script* destina-se a ser executado através da *shell bash*, no ambiente Linux da máquina virtual disponibilizada para a UC de Sistemas Operativos.

O *script* **`wordStats.sh`** destina-se a produzir um conteúdo de texto, produzindo análise estatística e resultados gráficos sobre esse mesmo conteúdo. A análise estatística deve incluir:

- Listagem de todas as palavras ordenadas pela respetiva frequência de ocorrência;

- Criação de uma página HTML e de um gráfico de barras referente ao número de ocorrência das N palavras mais frequentes.

2 - Visão geral

O funcionamento do *script* **`wordStats.sh`** é muito simples:

1. Receção do conteúdo a processar (ficheiro de texto ou PDF, sendo que o PDF deve ser convertido pelo *script* em texto através do utilitário `pdftotext`[1]);
2. Processamento do conteúdo e criação de resultados;
3. Apresentação dos resultados.

3 - Stop-words

Quando se efetua a análise de textos é comum retirar as palavras de ligação, como os artigos in/definidos (*um, uns, uma, umas, o, a, os, as*, etc.) e afins. A designação Anglo-Saxónica para essas palavras é *stop-*

[1] `man pdftotext`

words[2], sendo que, obviamente, as *stop-words* variam consoante as línguas. Os ficheiros de *stop-words* para a língua portuguesa e inglesa encontram-se no Moodle.

O *script* `wordStats.sh` deve estar preparado para funcionar com ou sem *stop-words*, cabendo ao utilizador especificar o modo pretendido, conforme descrito na secção seguinte. Os ficheiros de *stop-words* a serem considerados pelo *script* deve estar no subdiretório `StopWords` do diretório onde se localiza o *script* `wordStats.sh`.

4 - Modos de operação e formatos de saída

O `wordStats.sh` tem três modos de operação: `c/C` (count), `p/P` (plot) e `t/T` (top). Os comandos em minúsculas (**c**, **p** ou **t**) indicam que se pretende utilizar o modo de remoção *stop-words*, ao passo que a especificação de um comando com letra maiúscula (**C**, **P** ou **T**) analisa todas as palavras do ficheiro indicado pelo utilizador. Seguidamente, são descritos esses modos de operação, bem como os respetivos formatos de saída.

O modo é especificado através de uma das seguintes letras na linha de comando:

`-Modo c/C`: o *script* efetua a contagem (C=count) de ocorrências de cada palavra, guardando a lista ordenada de forma decrescente num ficheiro de texto. O nome do ficheiro corresponde ao nome do recurso de entrada acrescentado com o prefixo "`result---`". Por exemplo, caso o recurso de entrada seja "`a.txt`", o ficheiro de resultado chamar-se-á "`result---a.txt`". Caso já exista um ficheiro com o mesmo nome do ficheiro de saída, o ficheiro existente é substituído pelo novo ficheiro.

`-Modo p/P`: o *script* efetua a contagem de ocorrências de cada palavra, produzindo um gráfico (P=plot) de barras com as N palavras que ocorrem com maior frequência. O gráfico é guardado num ficheiro PNG, cujo nome segue as regras definidas para o modo **c/C** (prefixado com "`result---`" e reescrito caso já exista). Para apresentação do gráfico (ficheiro PNG), deve ser criado um ficheiro HTML cujo nome obedece às mesmas regras do ficheiro PNG. Neste modo o *script* deve mostrar o ficheiro PNG do gráfico antes de terminar[3]. O valor de N é definido pela variável do ambiente WORD_STATS_TOP. Caso essa variável não se encontre definida ou não corresponda a um número, deve ter assumido para N o valor 10.

`- Modo t/T`: o *script* efetua a contagem de ocorrências de cada palavra mostrando na saída padrão do terminal o "Top" das palavras, isto é, as **N** palavras que ocorrem com maior frequência. De forma similar aos restantes comandos, é criado um ficheiro de texto com os resultados. O ficheiro segue as mesmas regras de nome definida para os outros modos. Similarmente, o valor de **N** segue as mesmas regras do que o indicado para o modo **p/P**.

5 - Parâmetros da linha de comando

A sintaxe do script `wordStats.sh` é a seguinte:

2 Uma possível tradução para *stop-words* é *palavras irrelevantes*.

[3] Pode fazer uso do utilitário display que já se encontra instalada na VM da UC.

```
wordStats.sh <MODE>  <INPUT>  <ISO3166>
```

Cada elemento deve obedecer às seguintes regras:

- **<MODE>** é um dos possíveis três modos, isto é, **c/C**, **p/P** ou **t/T**. Caso não seja especificado um desses modos, o *script* deve terminar, com uma apropriada mensagem de erro (ver Secção 6 - - Exemplos de uso).
- **<INPUT>** corresponde ao nome de um ficheiro local em formato TXT ou PDF.
- **<ISO3166>** corresponde ao código que indica qual a língua em que se encontram o conteúdo especificado pelo parâmetro INPUT. Os valores suportados são '`pt`' e '`en`', respetivamente para português e inglês. Este parâmetro é opcional, sendo que quando não é especificado, deve assumir-se que o conteúdo está em língua inglesa.

6 - Exemplos de uso

NOTA:

i) Nos exemplos seguintes `ficha01.pdf` corresponde ao ficheiro PDF da 1ª ficha prática da UC;

ii) Os resultados (top, contagem de palavras, etc.) podem diferir consoante a metodologia seguida (por exemplo, incluir ou não, um número como 2021 na listagem de palavras, etc.). Assim, caso os resultados uma implementação sejam (ligeiramente) diferentes dos aqui apresentados, não significa que a implementação esteja errada.

Exemplo 1 – modo "c" (com remoção de *stop-words*)

```
./word_stats.sh c ficha01.pdf pt
'ficha01.pdf': PDF file
[INFO] Processing 'ficha01.pdf.txt'
STOP WORDS will be filtered out
StopWords file 'pt': 'StopWords/pt.stop_words.txt' (205 words)
COUNT MODE
     1  73      diretoria
     2  64      ficheiro
     3  61      comando
     4  37      utilizador
     5  34      ficheiros
     6  33      comandos
     7  32      sistema
(…)
RESULTS: 'result---ficha01.pdf.txt'
-rw-rw-r-- 1 user user 15479 Mar 10 23:14 result---ficha01.pdf.txt
912 distinct words
```

Exemplo 2 – modo "C" (sem remoção de *stop words)*

```
./word_stats.sh C ficha01.pdf
'ficha01.pdf': PDF file
[INFO] Processing 'ficha01.pdf.txt'
STOP WORDS will be counted
COUNT MODE
     1  196     o
     2  162     de
     3  162     a
     4  102     para
     5  101     do
     6  85      -
     7  71      diretoria
(…)
RESULTS: 'result---ficha01.pdf.txt'
```

```
-rw-rw-r-- 1 user user 18537 Mar 10 23:15 result---ficha01.pdf.txt
1097 distinct words
```

Exemplo 3 – modo "p" (com remoção de *stop words*)

```
./word_stats.sh p ficha01.pdf pt
'ficha01.pdf': PDF file
[INFO] Processing 'ficha01.pdf'
STOP WORDS will be filtered out
StopWords file 'pt': 'StopWords/pt.stop_words.txt' (205 words)
-rw-rw-r-- 1 user user 154 Mar 10 23:19 result---ficha01.pdf.txt.dat
-rw-rw-r-- 1 user user 11452 Mar 10 23:19 result---ficha01.pdf.txt.png
-rw-rw-r-- 1 user user 525 Mar 10 23:19 result---ficha01.pdf.txt.html
```

Descrição: Execução em modo *plot* / remover *stop words* (língua portuguesa) para análise ao ficheiro "ficha01.pdf"
Ficheiros produzidos: `result---ficha01.pdf.txt.png` e `result---ficha01.pdf.txt.html` (ver Figura 1).

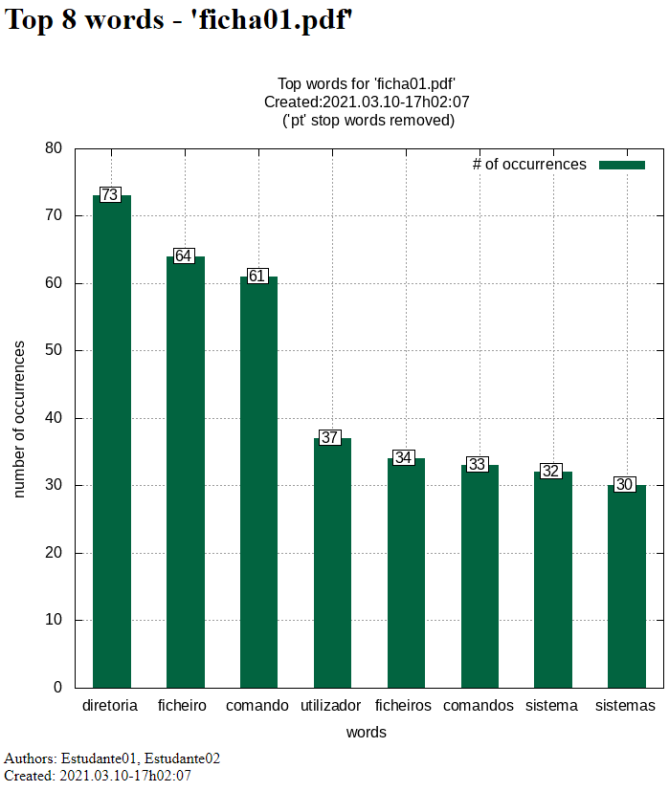

Figura 1: ficheiro "`result---ficha01.pdf.txt.html`"

Exemplo 4 – modo "P" (sem remoção de *stop words*)

```
./word_stats.sh P ficha01.pdf pt
'ficha01.pdf': PDF file
[INFO] Processing 'ficha01.pdf'
STOP WORDS will be counted
-rw-rw-r-- 1 user user 116 Mar 10 23:21 result---ficha01.pdf.txt.dat
-rw-rw-r-- 1 user user 11226 Mar 10 23:21 result---ficha01.pdf.txt.png
-rw-rw-r-- 1 user user 525 Mar 10 23:21 result---ficha01.pdf.txt.html
```

Top 8 words - 'ficha01.pdf'

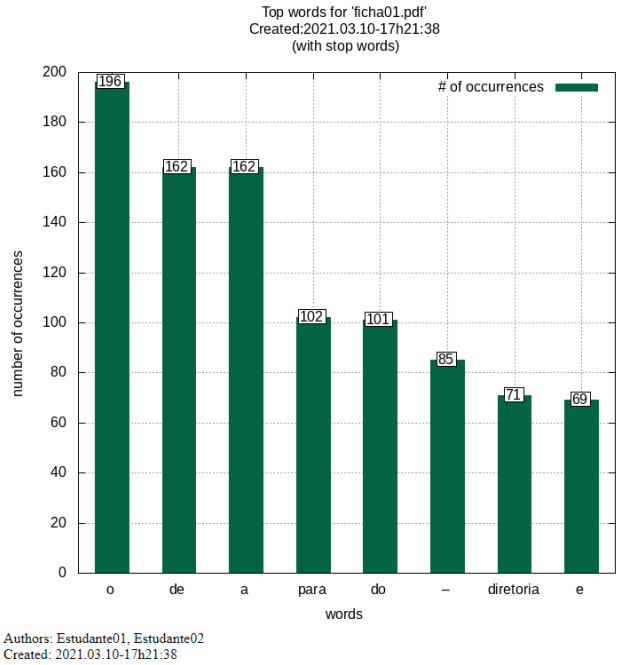

Figura 2: ficheiro "result---ficha01.pdf.txt.html"

Exemplo 5 – modo "t" (com remoção de *stop words*)

```
./word_stats.sh t ficha01.pdf pt
'ficha01.pdf': PDF file
[INFO] Processing 'ficha01.pdf'
STOP WORDS will be filtered out
StopWords file 'pt': 'StopWords/pt.stop_words.txt' (205 words)
Environment variable 'WORD_STATS_TOP' is empty (using default 10)
-rw-rw-r-- 1 user user 190 Mar 10 23:24 result---ficha01.pdf.txt.dat
-------------------------------------
# TOP 10 elements
     1  73      diretoria
     2  64      ficheiro
     3  61      comando
     4  37      utilizador
     5  34      ficheiros
     6  33      comandos
     7  32      sistema
     8  30      sistemas
     9  29      nome
    10  28      operativos
-------------------------------------
```

Exemplo 6 – modo "T" (sem remoção de *stop words*)

```
export WORD_STATS_TOP=5
./word_stats.sh T ficha01.pdf pt
'ficha01.pdf': PDF file
[INFO] Processing 'ficha01.pdf'
STOP WORDS will be counted
WORD_STATS_TOP=5
-rw-rw-r-- 1 user user 70 Mar 10 23:38 result---ficha01.pdf.txt.dat
-------------------------------------
# TOP 5 elements
     1  196     o
     2  162     de
     3  162     a
```

```
     4  102     para
     5  101     do
-------------------------------------
```

De seguida são apresentadas algumas situações de erro.

Exemplo 7 – ausência de parâmetros

```
./word_stats.sh
[ERROR] insufficient parameters
./word_stats.sh Cc|Pp|Tt INPUT [iso3166]
```

Exemplo 8 – ficheiro de entrada inexistente

```
./word_stats.sh T NaoExiste.txt pt
[ERROR] can't find file 'NaoExiste.txt'
```

Exemplo 9 – comando inexistente

```
./word_stats.sh A ficha01.pdf pt
[ERROR] unknown command 'A'
```

Exemplo 10 – variável do ambiente `WORD_STATS_TOP` impropriamente definida (é empregue o valor por omissão)

```
export WORD_STATS_TOP=54ABC
./word_stats.sh T ficha01.pdf pt
'ficha01.pdf': PDF file
[INFO] Processing 'ficha01.pdf.txt'
STOP WORDS will be counted
'54ABC' not a number (using default 10)
-rw-rw-r-- 1 user user 153 Mar 11 00:35 result---ficha01.pdf.txt.dat
-------------------------------------
# TOP 10 elements
     1  196     o
     2  162     de
     3  162     a
     4  102     para
     5  101     do
     6  85      -
     7  71      diretoria
     8  69      e
     9  64      ficheiro
    10  61      comando
-------------------------------------
```

7 - Implementação

A implementação do script `wordStats.sh`:

i) Deve recorrer somente aos utilitários de linha de comando existentes na máquina virtual da UC.

ii) Os gráficos de barras podem ser feitos com a ferramenta `gnuplot` (http://www.gnuplot.info/) que já se encontra instalada na máquina virtual. Exemplo de um (simples) ficheiro `gnuplot` (bar.gnuplot) e de um ficheiro de dados (result---ficha01.pdf.txt.dat) apropriados são mostrados na Listagem 1e Listagem 2, respetivamente.

iii)

```
# simple gnuplot example
set terminal png
set output "out.png"
set boxwidth 0.5
set style fill solid
plot "result---ficha01.pdf.txt.dat" using 1:2:xtic(3) with boxes
```

Listagem 1: Exemplo de ficheiro gnuplot "bar.gnuplot"

```
1  196     o
2  162     de
3  162     a
4  102     para
5  101     do
```

Listagem 2: Exemplo de dados para "result---ficha01.pdf.txt.dat"

O gráfico pode ser gerado da seguinte forma: `gnuplot < bar.gnuplot`, sendo criado o ficheiro de saída `out.png`.

iv) O uso de código obtido da Internet deve ser limitado a pequenos fragmentos e convenientemente documentado com a indicação da fonte (e.g., URL, etc.). O uso de código externo sem a devida indicação da proveniência é penalizado em 25% da nota final, por cada ocorrência.

8 - Avaliação

A avaliação do projeto é distribuída da seguinte forma:

- Funcionamento: 70%
- Relatório: 10%
- Implementação, organização e qualidade do código: 20%
  - Qualidade e pertinência dos comentários
  - Qualidade e pertinência do nome dos identificadores (variáveis, funções, etc.)
  - Organização do código (uso de funções/estruturas de repetição para evitar duplicação de código, etc.)
  - Qualidade das soluções

9 - Relatório

O projeto deve ser acompanhado de um relatório composto por:

- Um máximo de **quatro** páginas. A primeira página identifica os estudantes do grupo com nome completo, número de estudante, fotografia de rosto atualizada e a seguinte declaração:
  - "*Nome_Estudante_1 (numero_estudante_1) e por Nome_Estudante_2 (numero_estudante_2) declaram sob compromisso de honra que o presente trabalho (código, relatórios e afins) foi integralmente realizado por nós, sendo que as contribuições externas se encontram claramente e inequivocamente identificadas no próprio código. Mais se declara que os estudantes acima identificados não disponibilizaram o código ou partes dele a terceiros*".

- A segunda página do relatório deve descrever, através de uma tabela, o estado de cada funcionalidade, indicando se está funcional ou não (e.g., modo c: totalmente operacional; modo p: problemas na criação do gráfico, etc.).
- O relatório deve ainda explicar a metodologia de contagem, indicando quais as opções tomadas (se números foram consideradas palavras, qual a abordagem em relação à pontuação, etc.).
- O relatório deve ser entregue em formato PDF, com o nome **`relatório_proj1_fundos_n1-n2.pdf`**, em que **n1** representa o número de estudante do 1º elemento do grupo e **n2** o número de estudante do 2º elemento do grupo.
- O relatório é **obrigatório**.
- A entrega do projeto deve ser feita no Moodle, existindo uma zona para cada turno prático. O projeto deve ser entregue na entrada correspondente ao turno prático **do estudante que tiver o menor número de estudante**.

10 - Regras

1 - Na realização do projeto podem ser empregues todos os recursos disponibilizados pela BASH existentes na máquina virtual da UC.

2 - O projeto prático deve limitar-se aos recursos existentes na máquina virtual da UC: o script deve executar sem ser necessário instalar qualquer software adicional na máquina virtual da UC.

3 - O trabalho será realizado **em grupo** (máximo de **dois** estudantes, que podem ser de turnos práticos distintos), ou individualmente.

4 - O trabalho deve estar claramente identificado, com o **nome completo e** respetivo **número de cada estudante** no ficheiro README.txt a ser entregue juntamente com o script.

5 - O *script* deve executar sem ser necessário qualquer modificação. Caso a execução do *script* seja interrompida por erro imputável ao código do mesmo, o trabalho é avaliado com a nota `0` (zero) valores.

6 - Os comentários e os variados identificadores presentes no código fonte (nome de variáveis, funções, etc.) devem estar em inglês.

7 - Todos os ficheiros do projeto (script, ficheiro README.txt e relatório) devem ser reunidos, através de um utilitário de arquivo e compressão (apenas são aceite: zip, 7Z, tar.gz ou tar.xz), num único ficheiro denominado "**SO.proj2020-2021.n1-n2.ext**" em que **n1** representa o número de estudante do 1º elemento do grupo, **n2** o número de estudante do 2º elemento do grupo e **`ext`** representa a extensão do arquivo que pode ser **.zip**, **.7z**, **.tar.gz** ou **tar.xz** dependente do utilitário empregue.

8 - O ficheiro relativo ao ponto anterior (regra nº 5) deve ser entregue através do mecanismo de entrega disponibilizado no moodle da unidade curricular. Em caso de dúvidas deve consultar os docentes.

9 - Fraudes ou tentativas de fraudes originam uma classificação **nula** no presente trabalho para os prevaricadores, bem como o relato do sucedido às instâncias superiores.

10 - Caso faça uso do correio eletrónico para o esclarecimento de dúvidas, deve sempre iniciar o assunto da mensagem por **[EI_SO][Projeto]** (caso contrário, a mensagem corre o risco de não ser corretamente identificada pelo filtro *anti-spam*). Para além disso, deve identificar-se com o nome, número, regime e turno prático que frequenta.

11 - Após a entrega do projeto, poderá ser necessária uma apresentação oral do mesmo através de teleconferência, sendo esta agendada pelo docente. A apresentação é individual, sendo que a nota percentual na apresentação (de 0% a 100%) é multiplicada pela nota resultante da correção para efeitos de cálculo da nota final do projeto.

Bibliografia

- The Linux Command Line, William E. Shotts, Jr. (Licença creative common - http://linuxcommand.org/tlcl.php), 2019
- *gnuplot home page* (http://www.gnuplot.info/)
- Bash: **1)** https://www.gnu.org/software/bash/manual/; **2)** https://devhints.io/bash; **3)** https://wiki.bash-hackers.org/