Peso na nota final:	5 valores em 20.
Informações:	Moodle
Data de entrega	16 de abril de 2020 / 23h59’
Publicação de nota	16 de maio de 2020

Sistemas Operativos

Enunciado IMPAR – destina-se aos grupos nos quais MINIMO(n1,n2) é **ÍMPAR**, sendo n1 e n2, respetivamente, o número de estudante de cada aluno do grupo. Similarmente, no caso de trabalho individual, destina-se a alunos cujo número de estudante é **ÍMPAR**.

Projeto 1 - 2S 2019-2020 – Enunciado ÍMPAR

Script bash *google_scholar.sh*

1 - Introdução

Uma das funcionalidades do *Google Scholar* é o de permitir a cada investigador dispor de uma página de perfil pessoal onde são listadas, de forma automática, as publicações científicas dos mesmos. Para além disso, são ainda registadas o número de citações, assim como outras métricas associadas às citações. Note-se que o número de citações em publicações científicas é frequentemente visto como o reconhecimento da qualidade da publicação científica. Por exemplo, a página https://scholar.google.com/citations?hl=en&user=Wj4ZBFIAAAAJ está associado ao Professor David Patterson, famoso, entre outras coisas, pelo livro “*Computer architecture: a quantitative approach*”. A página indica que foram registadas 98006 citações, das quais 30358 ocorreram nos últimos 5 anos (desde 2015), e que o Professor David Patterson tem um h-index[1] global (considerando todas as publicações) de 107 e de 59 quando apenas são consideradas as publicações dos últimos cinco anos.

Pretende-se que elabore o script BASH google_scholar.sh capaz de extrair as métricas **i) total de**

[1] Um *h-index* de N indica que o autor tem pelo menos N publicações em que cada um tem pelo menos N citações.

citações; **ii) total de citações nos últimos 5 anos**; **iii) h-index** total e **iv) h-index dos últimos cinco anos** e ainda **v) o nome da pessoa detentor do perfil em análise**.

Para esse efeito, pretende-se que elabore o script **google_scholar.sh**. O script destina-se a ser executada através da *shell bash*, no ambiente Linux da máquina virtual disponibilizada para a UC de Sistemas Operativos.

2 - Dados a processar

O script deve poder processar várias páginas de perfil pessoal numa só execução. Os endereços das páginas de perfil pessoal e respetivos nomes são indicados pelo utilizador do script através do ficheiro **`scholar_URLs.txt`**. Caso este ficheiro não exista no diretório corrente, o script deve terminar a execução com a seguinte mensagem de erro:

```
[ERRO] Não foi possivel encontrar 'scholar_URLs.txt'
```

O ficheiro **`scholar_URLs.txt`** contém, em cada linha, o URL e um nome que se pretende usar para salvaguarda do ficheiro HTML que representa o perfil, sendo empregue o símbolo **|** como separador. As linhas iniciadas por **#** correspondem a linhas comentadas e devem ser ignoradas.

A Listagem 1 apresenta um exemplo do ficheiro **scholar_URLs.txt**[2], com a primeira e a última linha comentadas. Na 2ª linha está o URL para o perfil pessoal do Professor David Patterson (1ª secção antes do símbolo **|**). A segunda parte da linha – DavidPatterson.html – representa o nome que o script deve usar para gravar o conteúdo da página HTML do referido perfil pessoal.

O script deve processar cada um dos perfis indicados no ficheiro scholar_URLs.txt. Assim, considerando a Listagem 1, o script deverá processar três perfis (a última linha do ficheiro está comentada com #), gravando o conteúdo de cada um, respetivamente, nos ficheiros `DavidPatterson.html`, `GeoffreyHinton.html` e `YoshuaBengio.html`. Esses ficheiros devem ser gravados no subdiretório **Scholar**.

```
# URL|HTML filename
https://scholar.google.com/citations?hl=en&user=Wj4ZBFIAAAAJ|DavidPatterson.html
https://scholar.google.com/citations?hl=en&user=JicYPdAAAAAJ|GeoffreyHinton.html
https://scholar.google.com/citations?hl=en&user=ZpG_cJwAAAAJ|YoshuaBengio.html
# https://scholar.google.com/citations?hl=en&user=mG4imMEAAAAJ|AndrewNg.html
```

Listagem 1: exemplo de ficheiro scholar_URLs.txt

No que respeita à origem dos dados – ficheiro ou Internet –, o script **`google_scholar.sh`** deve ter o seguinte comportamento:

i) Caso seja indicada a opção **`-i`** na linha de comandos (e.g., `google_scholar.sh -i`), o script deve proceder ao descarregar da página associada a cada perfil, gravando-a para o subdiretório **Scholar** com o nome indicado como 2º elemento da linha (e.g., DavidPatterson.html). O script deve criar o diretório **Scholar** caso esse não exista.

[2] Um exemplo de ficheiro scholar_URLs.txt está disponível em https://pastebin.com/raw/yz6PZaXb

ii) Caso **não** seja indicada a opção "**-i**" na linha de comando, o script deve procurar no subdiretório **Scholar** os ficheiros HTML indicados como 2º elemento no ficheiro **scholar_URLs.txt**. Caso não encontre um desses ficheiros, o script deve terminar com a seguinte mensagem de erro:

`[ERRO] Não foi possível encontrar o ficheiro 'FILENAME'`

em que que FILENAME representa o nome do ficheiro em falta.

3 - Funcionamento

a) Para cada perfil indicado no ficheiro scholar_URLs.txt, o script deve extrair os seguintes dados**: i) total de citações**; **ii) total de citações nos últimos 5 anos**; **iii) h-index total** e **iv) h-index dos últimos cinco anos** e ainda **v) nome da pessoa do perfil**, sendo que para o nome deve eliminar o espaço entre o nome e o apelido (e.g., DavidPatterson). Quando executado, o script deve produzir a seguinte saída (Listagem 2), para cada um dos perfis existentes no **scholar_URLs.txt**:

```
-----------------------------------------------------
[A processar]: https://scholar.google.com/citations?hl=en&user=Wj4ZBFIAAAAJ
[INFO] A utilizer o ficheiro local 'DavidPatterson.html'
Scholar: 'DavidPatterson'
Citacoes - Total: 98006, ultimos 5 anos: 30358
H-Index - Total: 107, ultimos 5 anos: 59
```

Listagem 2: Saída do script para o perfil de David Patterson

Para além disso, o script deve ainda registar num ficheiro cujo nome corresponde ao nome do perfil acrescido da extensão **.db** (e.g., **DavidPatterson.db**), as métricas acima indicadas. O ficheiro deve ter um cabeçalho conforme indicado na Listagem 3 e ter como última linha a data e hora da última atualização. Acresce-se que todos os ficheiros **.db** devem ser guardados no subdiretório **Scholar**.

```
# Ficheiro: 'DavidPatterson.db'
# Info Scholar: 'DavidPatterson'
# Criado em: 2020.03.23_17h44:27
# Citacoes:Citacoes-5anos:h-index:h-index_5anos
2020.03.23:98006:30358:107:59
# Ultima atualizacao: 2020.03.23_22h08:14
```

Listagem 3: Exemplo de ficheiro .db (DavidPatterson.db)

4 - Opções da linha de comando

O script deve suportar as seguintes opções da linha de comando:

`-h`: mostra ajuda sucinta

`-i`: tenta obter os dados a partir do URL indicado anteriormente

5 - Avaliação

A avaliação do projeto é distribuída da seguinte forma:

- Funcionamento: 80%

- Implementação, organização e qualidade do código: 20%

6 - Relatório

O projeto deve ser acompanhado de um relatório composto por **duas** 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 o estado de cada funcionalidade, indicando se está funcional ou não (e.g., opção -i: totalmente operacional; opção -h: não implementada). O relatório deve ser entregue em formato PDF, com o nome **`relatório_proj1_scholar_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**.

7 - Regras

1 - O trabalho será realizado **individualmente** ou em grupo (máximo de **dois** estudantes, que podem ser de turnos práticos distintos).
2 - 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.
3 - O script deve executar sem ser necessário qualquer modificação. Caso a execução do script seja interrompida por erro imputável ao código do script, o trabalho é avaliado com a nota **`0 (zero)`** valores.
4 - Os comentários e os variados identificadores presentes no código fonte (nome de variáveis, funções, etc.) devem estar em inglês.
5 - Todos os ficheiros do projeto (script, ficheiro README.txt, scholar_URLs.txt e relatório) devem ser reunidos, através de um utilitário de arquivo e compressão (zip, 7Z, tar.gz, ou tar.bz2), num único ficheiro denominado "**SO.proj1_scholar_2019-2020.n1-n2**[3]" em que **n1** representa o número de estudante do 1º elemento do grupo e **n2** o número de estudante do 2º elemento do grupo.
6 - 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.
7 - Fraudes ou tentativas de fraudes originam uma classificação **nula** no presente trabalho para os prevaricadores, bem como o relato do sucedido às instâncias superiores.
8 - Caso faça uso do correio eletrónico para o esclarecimento de dúvidas, deve sempre iniciar o assunto da mensagem por **[EI_SO][Projeto_1]** (caso contrário, a mensagem corre o risco de não ser corretamente classificada pelo filtro anti-spam). Para além disso, deve identificar-se com o nome, número, regime e turno prático que frequenta.
9 - 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.

[3] A extensão do arquivo (.zip, .7z, .tar.gz, tar.bz2) depende do utilitário empregue para a compactação.