Trabalho Final de ICC (2017/02)

Pablo Cecilio, Marco Antônio, Lucas Souza

1 Introdução

O trabalho consiste no processamento e manuseio entre bases de dados usando a programação em bash. Sendo esse processado através do uso de um script contendo os comandos necessários para executar diversas ações especificas entre as bases de dados originais, gerando a saída requisitada pelos itens abaixo descritos.

2 Pré-processamento

Para realizar as extrações solicitadas os dois arquivos "title.basics.tsv" e "title.ratings.tsv" foram processados utilizando o comando "join", que concatena linhas de dois arquivos através de uma coluna em comum.

```
join -t $'\t' -o 2.1,2.2,2.3,2.4,2.5,2.6,2.7,2.8,2.9,1.2,1.3
↪  title.ratings.tsv title.basics.tsv
```

De modo a conferir se a saída foi gerada corretamente, o comando "wc -l" foi executado ao final do processo para comparação visual de linhas entre os arquivos "titles.ratings.tsv" e o arvivo gerado "titles.tsv".

```
wc -l title.ratings.tsv
wc -l titles.tsv
```

Finalizando o pre-processamento, o comando "sed" foi utilizado para remover a primeira linha do arquivo "titles.tsv", e gerar um novo arquivo "titles.all.tsv". Novamente a saída foi verificada com a contagem das respectivas linhas através do comando "wc -l".

```
sed '1d' titles.tsv > titles.all.tsv
wc -l titles.all.tsv
```

3 Extrações

A extração dos itens 1, 2, 3, 4, 5, 6 e 11 foram realizdas.

Item 1

Liste os tipos de titulo (titleType) únicos existentes e imprima em ordem lexicográfica.

Para isso foi utilado o comando "cut" para selecionar a coluna 2 em "title.all.tsv", após esse comando foi realizado um pipe para um "sort" afim de ordenar a seleção e em seguida outro pipe para "uniq", para que com isso se pudesse listar as entradas únicas da seleção como saída para o arquivo.

```
cut -f 2 titles.all.tsv | sort | uniq > out1.txt
```

Entrada	cut -f 2	sort	uniq
F A X	A	A	A
G B Y	B	A	B
H A Z	A	B	C
K C U	C	C	

Item 2

Quantos títulos tem o primaryTitle e o originalTitle iguais?

Foi utilizado o comando "awk' com uma variável como contador, essa por sua vez foi condicionada a aumentar seu valor caso as colunas 3 e 4 fossem iguais. A saída foi impressa em arquivo com o resultado da variável.

```
awk -F"\t" '{if ($3 == $4) s += 1}END{print s}' titles.all.tsv
↪  > out2.txt
```

Item 3

Qual a media das avaliações feitas entre os anos 1970 e 2000 (incluindo 1970 e 2000)? Imprima o resultado com 4 casas decimais.

Atraves do comando awk compara se a coluna 6 está entre os valores (1970 a 2000), recebe em s a nota do filme, e em t a quantidade de filmes, imprime em out3.txt a media das notas.

Erro nos itens 3 e 4 por ter que colocar duas condições para uma mesma variável separadas.

```
awk -F"\t" '{if (1969 < $6  && $6 < 2001){s += $10; t+=
↪  1}}END{print s/t}' titles.all.tsv > out3.txt
```

Item 4

Compara se a coluna 6 está entre os valores (2000 a 2016), recebe em s a nota do filme, e em t a quantidade de filmes, imprime em out4.txt a media das notas.

Erro nos itens 3 e 4 por ter que colocar duas condições para uma mesma variável separadas.

```
awk -F"\t" '{if (1999 < $6  && $6 < 2017){s += $10; t+=
↪  1}}END{print s/t}' titles.all.tsv > out4.txt
```

Item 5

Separa a coluna 9 do arquivo, elimina (com grep -v) as linhas que possuem ',' e '\N', organiza (com sort) por ordem alfabética, elimina (com uniq) as entradas repetidas, conta as linhas com wc -l e imprime o numero de linhas no out5.txt.

Erro no item 5 sintaxe para usar o grep -v no \N ter que utilizar duas \.

```
cut -f9 titles.all.tsv | grep -v "," | grep -v '\\N' | sort |
↪  uniq | wc -l > out5.txt
```

Item 6

Separa a coluna 9 do arquivo, separa todas linhas que possuem a palavra Action, conta as linhas e printa o numero em out6.txt

```
cut -f9 titles.all.tsv | grep "Action"| wc -l > out6.txt
```

Item 11

Separa a coluna 9, elimina as linhas que possuem ‘,’ e ‘\N’ sobrando somente os generos unicos e conta as linhas com wc -l

```
cut -f9 titles.all.tsv | grep -v "," | grep -v '\\N' | wc -l
```

4 Organização do trabalho

O processo de colaboração do trabalho foi realizado e documentado utilizando o GitHub como plataforma. `https://github.com/Durfan/ICC_TP/`

A comunicação imediata assim como o planejamento foi feito de forma mais informal através do WhatsApp. Sendo realizado através desse aplicativo a divisão e designação de tarefas por demanda ou por forma voluntaria.

Membro	Atividades
Pablo Cecilio	pré-processamento; script.sh; itens 1, 7, 8; documentação
Marco Antônio	itens 2, 3, 4, 5, 6 e 11
Lucas Souza	—
Arthur Rocha	—

5 Conclusão

A falta de uma documentação bash didática e menos técnica dificultou e atrasou o desenvolvimento(...)

6 Extra