TRACTAMENT DE DADES AMB SHELL SCRIPTS

1r GED
Fonaments d'Informàtica
30/09/2024
Joel Torrens Espada i Mar Massanas Morató

link al repositori: https://github.com/JoelTorrens/pr-ctica1_fonaments_inform-tica.git

0. awk -F',' '{ if (NF == 16) print $0 }' CAvideos.csv > supervivents.csv

L'ordre llegeix l'arxiu CAvideos.csv, comprova cada línia per veure si té exactament 16 camps separats per comes, i si és així, escriu aquesta línia en un fitxer nou anomenat supervivents.csv.

Tindrem un total de 12620 registres.

```
marmassanas@MacBook-Air-de-Mar Pr1 % wc -l supervivents.csv
   12620 supervivents.csv
```

Exercici 1

Interpretació/Anàlisi:
L'objectiu era seleccionar només les columnes d'interès del fitxer original supervivents.csv, eliminant les columnes 12 i 16.

Resolució:
Hem utilitzat la comanda cut amb el delimitador de coma per seleccionar les columnes 1 a 11 i 13 a 15. El resultat es guarda en un nou fitxer anomenat superviventsModificat1.csv.

Dificultats trobades:
Una possible dificultat podria haver sigut no tenir en compte que al cut s'han de seleccionar les columnes amb les quals et vols quedar i no les que vols eliminar, dit d'una altra manera, fer el "cut -d',' -f12,16".

Exercici 2

Interpretació/Anàlisi:
Aquest pas consistia a filtrar les línies del fitxer per eliminar aquelles on la columna 14 indicava un error (valor "True").

Resolució:
Hem utilitzat la comanda awk per crear un nou fitxer, superviventsModificat2.csv, que conté només les línies que no contenen True. A més, hem comptat el nombre de registres eliminats guardant-ho a una variable.

Dificultats trobades:
Un possible inconvenient podria ser la identificació incorrecta dels errors si hi ha variacions en el format o en l'escriptura del valor "True". Tanmateix, no vam trobar aquest problema en les proves.

Exercici 3

Interpretació/Anàlisi:
En aquest exercici, hem afegit una nova columna "Ranking_Views" a la capçalera del fitxer CSV. Aquesta columna classifica els vídeos segons el nombre de visualitzacions, ajudant a identificar quins són els més populars.

Resolució:
Utilitzant awk, hem afegit "Ranking_Views" a la capçalera i hem creat una classificació basada en els valors de la columna 8 (visualitzacions). Les categories són "bo", "estrella" i "excel·lent" segons el nombre de visualitzacions.

Dificultats trobades:
Una dificultat va ser assegurar-nos que les condicions de classificació eren clares i que el càlcul es feia correctament per a cada registre. Vam haver de revisar els límits per a cada categoria.

Exercici 4

Interpretació/Anàlisi:
L'objectiu d'aquest exercici era afegir dues noves columnes, "Rlikes" i "Rdislikes", que representaven el percentatge de "likes" i "dislikes" en relació amb les visualitzacions, per obtenir una idea més clara sobre la recepció dels vídeos.

Resolució:
Hem utilitzat head i tail per separar la capçalera i la resta de les dades. En un bucle, hem calculat el percentatge de "likes" i "dislikes" línia per línia i ho hem afegit al nou fitxer superviventsModificat4.csv.

Dificultats trobades:
El càlcul del percentatge requeria tenir en compte casos en què el nombre de visualitzacions era zero, per evitar divisions per zero. Això es va solucionar implementant una condició que establia Rlikes i Rdislikes a zero en aquests casos.

Exercici 5

Interpretació/Anàlisi:
L'últim exercici consistia a buscar coincidències en el fitxer resultant superviventsModificat4.csv basades en un paràmetre d'entrada com ara video_id o title.

Resolució:
Hem implementat una cerca amb grep i awk per trobar coincidències i, si s'havien trobat, s'imprimien les dades rellevants en un format llegible.

Dificultats trobades:
Un repte va ser assegurar que el fitxer supervivents Modificat4.csv existís abans d'intentar cercar coincidències. També podria haver-hi confusió si l'usuari introduïa un paràmetre que no es trobava en el fitxer.

Conclusions

- Els exercicis ens han permès aprendre a manipular fitxers de dades de manera efectiva, tot i que s'han trobat algunes dificultats menors.
- És important verificar sempre les condicions en operacions que impliquen càlculs, com el càlcul de percentatges, per evitar errors.