

Solucion_linea_comando

Kevin García Mundo

30 de septiembre de 2024

1. Problema.

Utilizando la línea de comandos y los comandos ya incluidos en la distribución de Linux que instalaste en tu máquina virtual, encuentra las 10 películas peor calificadas en promedio y las 10 películas mejor calificadas en promedio del directorio **train**.

No es necesario encontrar estas respuestas con un solo comando, puedes utilizar varios para lograr el cometido. Puede resultar difícil obtener programáticamente los nombres de las películas, es suficiente poder identificarlas utilizando las direcciones URL presentes en los archivos *urls_neg.txt* y *urls_pos.txt*.

Sugerencia: Investiga sobre los comandos `grep`, `sed` y `awk`, son más complicados que los comandos discutidos en clases.

Solución. Vamos a suponer que ya tenemos descomprimido el archivo y ya accedimos a él, por lo que nuestro directorio actual es **acIImdb**, accedemos al directorio **train** para estar en el lugar adecuado para comenzar.

1.1. Entendimiento del problema.

Debemos observar los archivos del directorio **train** para trabajar. Con el comando

```
ls
```

podemos ver que hay un archivo llamado *README*, para acceder a él podemos ejecutar el comando

```
cat README
```

esto nos muestra el contenido del archivo.

Dicho archivo nos dice que hay directorios llamados **pos** y **neg**, los cuales contienen archivos *txt* los cuales siguen el formato *id - cal*, donde *id* son enteros no negativos, más adelante retomamos esto, y *cal* la calificación de la película.

En el directorio actual tenemos un archivo llamado *urls_pos.txt* y *urls_neg.txt*, estos contienen links de comentarios a películas en cada línea, la manera en que se relacionan con los archivos del tipo *id-cal* es el comentario en la línea *i* se identifica con $id = i - 1$.

Por ultimo, hay que tener en cuenta que los links se repiten, pues son comentarios con distinto *id* y calificación pero misma película.

1.2. Resolviendo el problema (peores).

Comenzamos a trabajar sobre las 10 películas peor calificadas, para esto seguiremos los siguientes pasos:

- 1) Accedemos al directorio **neg**, lo que hacemos con el siguiente comando es listar los nombres de los archivos de nuestro directorio, solamente las que tengan formato de comenzar en numero, tenga un guión bajo, un punto y termine con *txt*. Al resultante vamos a quitar los separadores `_` y `.`, imprimimos lo que resulte de la columna 1 y 2, ordenamos y el resultante lo ponemos en un archivo, llamado **neg_id_cal.txt**, que se creará en **train**.

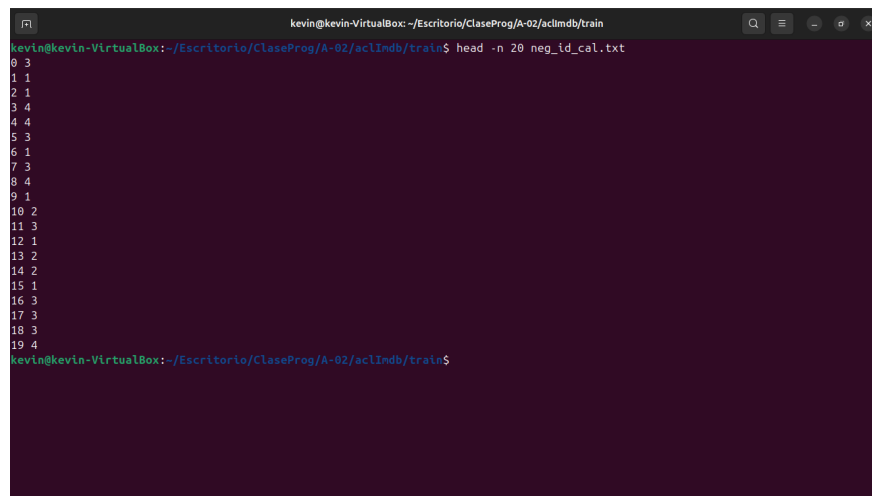
```
ls | grep -Eo '[0-9]+\_[0-9]+\.' |
awk -F'[_.]' '{print $1, $2}' | sort -n > ../neg_id_cal.txt
```

- 2) Ejecutamos el comando

```
cd ..
```

Para regresar al directorio **train**. Si queremos ver el resultado del paso anterior podemos ejecutar el siguiente comando:

```
head -n 20 neg_id_cal.txt
```

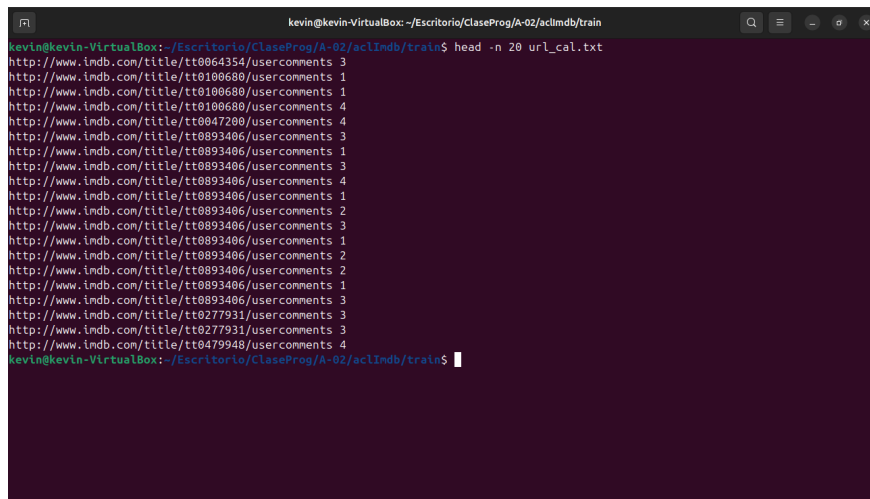


```
kevin@kevin-VirtualBox: ~/Escritorio/ClaseProg/A-02/aclmdb/train
kevin@kevin-VirtualBox:~/Escritorio/ClaseProg/A-02/aclmdb/train$ head -n 20 neg_id_cal.txt
0 3
1 1
2 1
3 4
4 4
5 3
6 1
7 3
8 4
9 1
10 2
11 3
12 1
13 2
14 2
15 1
16 3
17 3
18 3
19 4
kevin@kevin-VirtualBox:~/Escritorio/ClaseProg/A-02/aclmdb/train$
```

- 3) Necesitamos relacionar el archivo que contiene las url de los comentarios negativos con la calificación, una manera de hacerlo es combinando las líneas de los archivos, dado que solo queremos el url y la calificación ya resulta irrelevante el *id*, Empleamos el siguiente comando el cual combina las líneas, separando con un espacio, del archivo *urls_neg.txt* con las de *neg_id_cal.txt*, posteriormente solo imprime las columnas 1 y 3, para solo tener el url y la calificación, el resultante lo asignamos a un archivo llamado *url_cal.txt*.

```
paste -d' ' <(tail -n +1 urls_neg.txt) neg_id_cal.txt  
| awk '{print $1, $3}' > url_cal.txt
```

Revisamos que todo esté bien.



```
kevin@kevin-VirtualBox: ~/Escritorio/ClaseProg/A-02/aclIndb/train$ head -n 20 url_cal.txt  
http://www.imdb.com/title/tt0064354/usercomments 3  
http://www.imdb.com/title/tt0100680/usercomments 1  
http://www.imdb.com/title/tt0100680/usercomments 1  
http://www.imdb.com/title/tt0100680/usercomments 4  
http://www.imdb.com/title/tt0047280/usercomments 4  
http://www.imdb.com/title/tt0893406/usercomments 3  
http://www.imdb.com/title/tt0893406/usercomments 1  
http://www.imdb.com/title/tt0893406/usercomments 3  
http://www.imdb.com/title/tt0893406/usercomments 4  
http://www.imdb.com/title/tt0893406/usercomments 1  
http://www.imdb.com/title/tt0893406/usercomments 2  
http://www.imdb.com/title/tt0893406/usercomments 3  
http://www.imdb.com/title/tt0893406/usercomments 1  
http://www.imdb.com/title/tt0893406/usercomments 2  
http://www.imdb.com/title/tt0893406/usercomments 2  
http://www.imdb.com/title/tt0893406/usercomments 1  
http://www.imdb.com/title/tt0893406/usercomments 3  
http://www.imdb.com/title/tt0277931/usercomments 3  
http://www.imdb.com/title/tt0277931/usercomments 3  
http://www.imdb.com/title/tt0479948/usercomments 4  
kevin@kevin-VirtualBox: ~/Escritorio/ClaseProg/A-02/aclIndb/train$
```

- 4) Por ultimo debemos calcular el promedio de cada película con el siguiente comando:

```
awk '{ puntos[$1] += $2; suma[$1]++; } END { for (url in puntos)  
print url, puntos[url]/suma[url] }' url_cal.txt | sort -k2,2n  
| head -n 10 > promedio_neg.txt
```

El comando anterior trabaja sobre las columnas de nuestro archivo acumulando la suma de los promedios cuando un url se repita, una vez que termina para cada una de las url aplica un promedio dependiendo de cuantas url iguales encontró y cuanta calificación tenía cada una, una vez obtenidos los promedios los ordena de menor a mayor y por ultimo mostramos los primeros diez elementos, si queremos visitar esta lista lo asignamos a un archivo llamado *promedio_neg.txt*.

- 5) El resultante lo visualizamos con

```
cat promedio_neg.txt
```

Obteniendo lo siguiente:

```
kevin@kevin-VirtualBox: ~/Escritorio/ClaseProg/A-02/aclmdb/train$ cat prunedio_neg.txt
http://www.imdb.com/title/tt0004736/usercomments 1
http://www.imdb.com/title/tt0015477/usercomments 1
http://www.imdb.com/title/tt0021149/usercomments 1
http://www.imdb.com/title/tt0021861/usercomments 1
http://www.imdb.com/title/tt0023037/usercomments 1
http://www.imdb.com/title/tt0023669/usercomments 1
http://www.imdb.com/title/tt0024078/usercomments 1
http://www.imdb.com/title/tt0024509/usercomments 1
http://www.imdb.com/title/tt0024554/usercomments 1
http://www.imdb.com/title/tt0024675/usercomments 1
kevin@kevin-VirtualBox: ~/Escritorio/ClaseProg/A-02/aclmdb/train$
```

1.3. Resolviendo el problema (mejores).

Una vez que tenemos la solución para las peores, para las mejores es análogo, en este caso pondré solamente comandos e imágenes para mostrar lo obtenido, con comentarios en lo necesario.

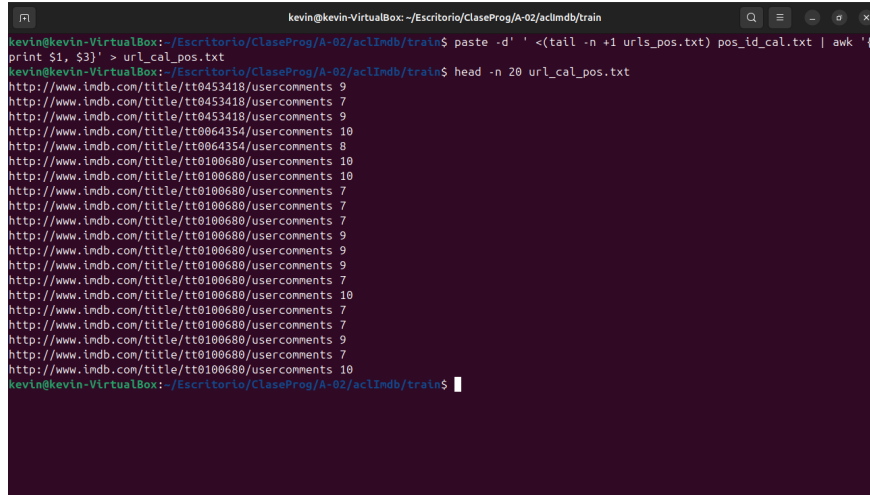
- 1) Accedemos ahora al directorio **pos** y ejecutamos

```
ls | grep -Eo '[0-9]+\_[0-9]+\.[txt]'
| awk -F'[_.]' '{print $1, $2}' | sort -n > ../pos_id_cal.txt
```

```
kevin@kevin-VirtualBox: ~/Escritorio/ClaseProg/A-02/aclmdb/train$ cd pos
kevin@kevin-VirtualBox: ~/Escritorio/ClaseProg/A-02/aclmdb/train/pos$ ls | grep -Eo '[0-9]+\_[0-9]+\.[txt]' | awk -F'[_.]' '{print $1, $2}' | sort -n > ../pos_id_cal.txt
kevin@kevin-VirtualBox: ~/Escritorio/ClaseProg/A-02/aclmdb/train/pos$ cd ..
kevin@kevin-VirtualBox: ~/Escritorio/ClaseProg/A-02/aclmdb/train$ head -n 20 pos_id_cal.txt
0 9
1 7
2 9
3 10
4 8
5 10
6 10
7 7
8 7
9 7
10 9
11 9
12 9
13 7
14 10
15 7
16 7
17 9
18 7
19 10
kevin@kevin-VirtualBox: ~/Escritorio/ClaseProg/A-02/aclmdb/train$
```

2) Creamos la relación url-calificación:

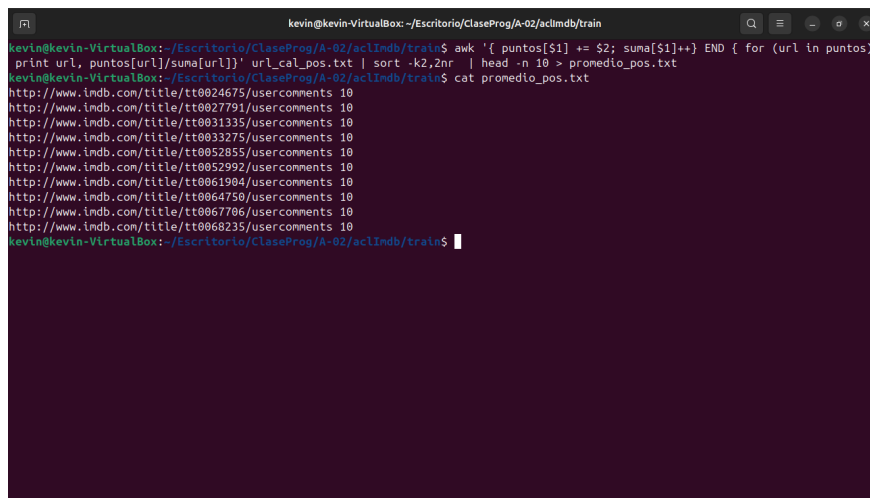
```
paste -d' ' <(tail -n +1 urls_pos.txt) pos_id_cal.txt  
| awk '{print $1, $3}' > url_cal_pos.txt
```



```
kevin@kevin-VirtualBox: ~/Escritorio/ClaseProg/A-02/aclmdb/train  
kevin@kevin-VirtualBox:~/Escritorio/ClaseProg/A-02/aclmdb/train$ paste -d' ' <(tail -n +1 urls_pos.txt) pos_id_cal.txt | awk '{  
print $1, $3}' > url_cal_pos.txt  
kevin@kevin-VirtualBox:~/Escritorio/ClaseProg/A-02/aclmdb/train$ head -n 20 url_cal_pos.txt  
http://www.indb.com/title/tt0453418/usercomments 9  
http://www.indb.com/title/tt0453418/usercomments 7  
http://www.indb.com/title/tt0453418/usercomments 9  
http://www.indb.com/title/tt0064354/usercomments 10  
http://www.indb.com/title/tt0064354/usercomments 8  
http://www.indb.com/title/tt0100680/usercomments 10  
http://www.indb.com/title/tt0100680/usercomments 10  
http://www.indb.com/title/tt0100680/usercomments 7  
http://www.indb.com/title/tt0100680/usercomments 7  
http://www.indb.com/title/tt0100680/usercomments 7  
http://www.indb.com/title/tt0100680/usercomments 9  
http://www.indb.com/title/tt0100680/usercomments 9  
http://www.indb.com/title/tt0100680/usercomments 9  
http://www.indb.com/title/tt0100680/usercomments 7  
http://www.indb.com/title/tt0100680/usercomments 10  
http://www.indb.com/title/tt0100680/usercomments 7  
http://www.indb.com/title/tt0100680/usercomments 7  
http://www.indb.com/title/tt0100680/usercomments 9  
http://www.indb.com/title/tt0100680/usercomments 7  
http://www.indb.com/title/tt0100680/usercomments 10  
kevin@kevin-VirtualBox:~/Escritorio/ClaseProg/A-02/aclmdb/train$
```

3) Sacamos promedio, ordenamos de mayor a menor y guardamos:

```
awk '{ puntos[$1] += $2; suma[$1]++} END { for (url in puntos)  
print url, puntos[url]/suma[url]}' url_cal_pos.txt  
| sort -k2,2nr | head -n 10 > promedio_pos.txt
```



```
kevin@kevin-VirtualBox: ~/Escritorio/ClaseProg/A-02/aclmdb/train  
kevin@kevin-VirtualBox:~/Escritorio/ClaseProg/A-02/aclmdb/train$ awk '{ puntos[$1] += $2; suma[$1]++} END { for (url in puntos)  
print url, puntos[url]/suma[url]}' url_cal_pos.txt | sort -k2,2nr | head -n 10 > promedio_pos.txt  
kevin@kevin-VirtualBox:~/Escritorio/ClaseProg/A-02/aclmdb/train$ cat promedio_pos.txt  
http://www.indb.com/title/tt0024675/usercomments 10  
http://www.indb.com/title/tt0027791/usercomments 10  
http://www.indb.com/title/tt0031335/usercomments 10  
http://www.indb.com/title/tt0033275/usercomments 10  
http://www.indb.com/title/tt0052855/usercomments 10  
http://www.indb.com/title/tt0052992/usercomments 10  
http://www.indb.com/title/tt0061984/usercomments 10  
http://www.indb.com/title/tt0064750/usercomments 10  
http://www.indb.com/title/tt0067786/usercomments 10  
http://www.indb.com/title/tt0068235/usercomments 10  
kevin@kevin-VirtualBox:~/Escritorio/ClaseProg/A-02/aclmdb/train$
```