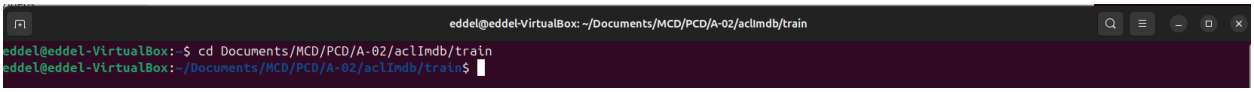
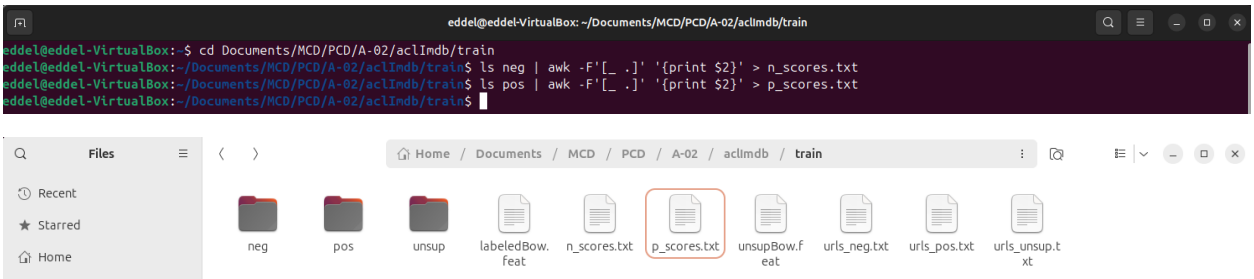
Programación para la Ciencia de Datos
Eddel Elí Ojeda Avilés
17 de agosto, 2024

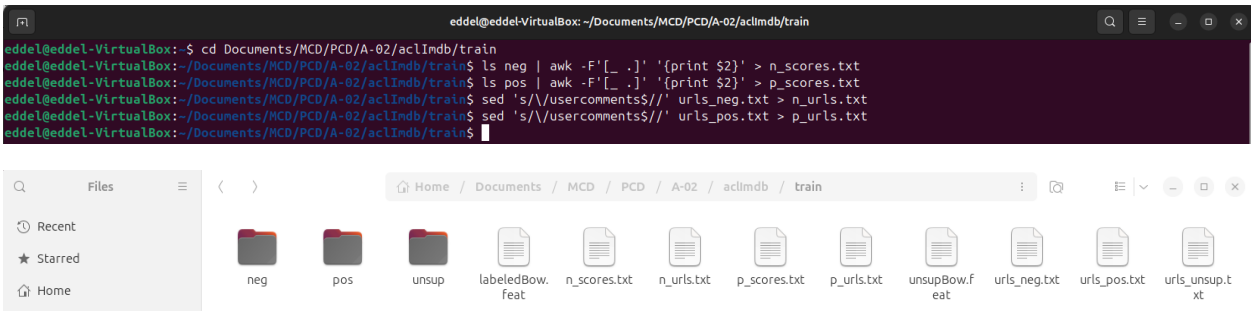
Para encontrar las 10 películas peor y mejor calificadas en promedio del directorio train primero nos posicionamos en dicho directorio.



Posteriormente a partir de la lista archivos en los directorios neg y pos guardamos únicamente la puntuación asignada en las reseñas en los archivos n_scores.txt y p_scores.txt, respectivamente. Esto se hizo usando el comando awk para tomar el segundo elemento en los títulos de los archivos, una vez señalados los separadores "_" y ".".

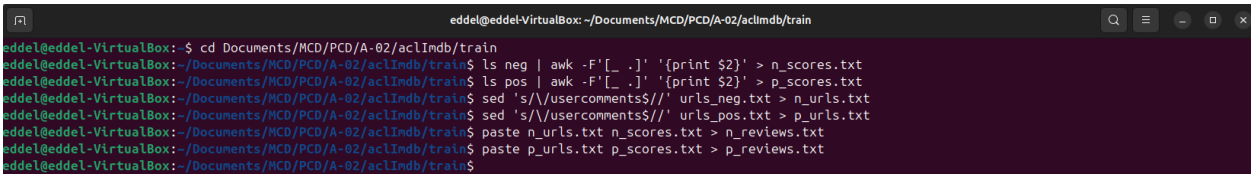


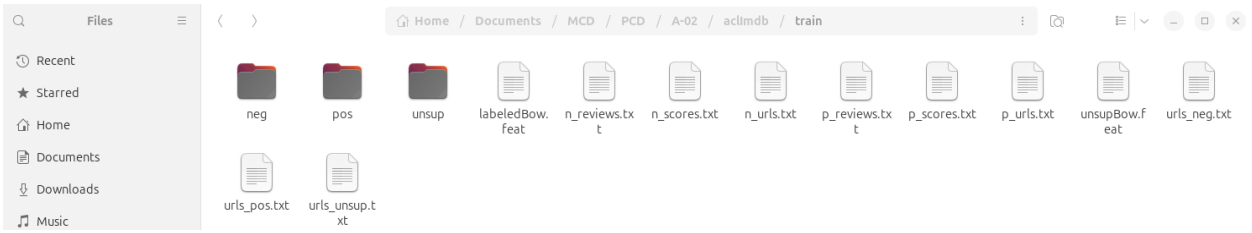
Con el fin de utilizar las urls apropiadas, creamos los archivos n_urls.txt y p_urls.txt a partir de los archivos urls_neg.txt y urls_pos.txt, respectivamente, los cuales contienen las urls resultantes de eliminar "/usercomments" con el comando sed.



La idea a continuación sería crear dos archivos n_names.txt y p_names.txt con los títulos de las películas asociadas a cada una de las urls, esto usando el comando curl, sin embargo, sigo trabajando en eso.

El siguiente paso consiste en utilizar el comando paste para generar los archivos n_reviews.txt y p_reviews.txt mezclando los archivos que contienen las urls, títulos de películas y puntuaciones negativas y positivas, respectivamente.





Finalmente, creamos los archivos n10_ave_revies.txt y p10_ave_revies.txt usando el comando awk donde calculamos la suma de las puntuaciones asociadas a la misma url a la par de que utilizamos un contador para guardar el número de veces que cada url se repite, para posteriormente guardar en un archivo cara una de las url distintas seguidas del cociente de la suma de sus puntuaciones entre cantidad de reseñas, es decir, su puntuación promedio, ordenándolas por su segunda componente de manera descendente o ascendente según el tipo de reseñas a considerar, para finalmente tomar únicamente las primeras o últimas 10 dependiendo de si se buscan los promedios más pequeños o más grandes.