Gestión de Datos. Práctica 2

1. Fecha de entrega

El trabajo, que podrá ser hecho en parejas, se deberá entregar a más tardar a las 24hs. del día 17 de diciembre a través de la tarea Moodle creada a tal efecto.

2. Objetivo

El objetivo del ejercicio es poner en prácticas algunas de los conceptos de Procesamiento de Datos vistos en la teoría. Para ello se trabajará con dos ficheros extraídos de un sistema real, aunque con fines de experimentación.

3. Resumen del conjunto de datos

Los datos han sido tomados a partir de la base de datos generada por el sistema de recomendación de películas MovieLens (http://movielens.org). Este conjunto de datos contiene de puntuaciones de películas (utilizando un sistema de 5 estrellas) de MovieLens. Los datos fueron alterados respecto de los ficheros originales por motivos pedagógicos. Contienen 25000095 puntuaciones sobre 62423 películas, creadas por 162541 usuarios entre el 09 de enero de 1995 el 21 de noviembre de 2019.

Los usuarios fueron seleccionados al azar (de entre todos los contenidos en la base de datos), y cada uno de ellos tiene por lo menos 20 películas puntuadas. Cada usuario está representado por un id y no se proveen más datos.

Los datos están almacenados en movies.csv y ratings.csv. Los ficheros originales, así como otros ficheros relacionados, pero no necesarios para este ejercicio están publicados y pueden ser descargados libremente en <http://grouplens.org/datasets/>.

En el texto a continuación los párrafos en cursiva corresponden a partes de los ficheros originales de explicación provistos por los datos.

Usage License

Neither the University of Minnesota nor any of the researchers involved can guarantee the correctness of the data, its suitability for any particular purpose, or the validity of results based on the use of the data set. The data set may be used for any research purposes under the following conditions:

** The user may not state or imply any endorsement from the University of Minnesota or the GroupLens Research Group.*

** The user must acknowledge the use of the data set in publications resulting from the use of the data set (see below for citation information).*
** The user may not redistribute the data without separate permission.*
** The user may not use this information for any commercial or revenue-bearing purposes without first obtaining permission from a faculty member of the GroupLens Research Project at the University of Minnesota.*
** The executable software scripts are provided "as is" without warranty of any kind, either expressed or implied, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The entire risk as to the quality and performance of them is with you. Should the program prove defective, you assume the cost of all necessary servicing, repair or correction.*

In no event shall the University of Minnesota, its affiliates or employees be liable to you for any damages arising out of the use or inability to use these programs (including but not limited to loss of data or data being rendered inaccurate).

If you have any further questions or comments, please email <grouplens-info@umn.edu>

Formato y Codificación

Los ficheros están guardados en formato CSV, con una línea de encabezamiento (nombre de columnas). Las columnas que contienen comas están escapadas usando comillas dobles (`"`). Los ficheros están codificados en UTF-8.

Movie Ids

Permiten relacionar la información de puntuaciones con el fichero de películas.

Estructura del fichero de puntuaciones (ratings.csv)

All ratings are contained in the file `ratings.csv`. Each line of this file after the header row represents one rating of one movie by one user, and has the following format:

userId,movieId,rating,timestamp

The lines within this file are ordered first by userId, then, within user, by movieId.

Ratings are made on a 5-star scale, with half-star increments (0.5 stars - 5.0 stars).

Timestamps represent seconds since midnight Coordinated Universal Time (UTC) of January 1, 1970.

Estructura del fichero de películas (movies.csv)

Movie information is contained in the file `movies.csv`. Each line of this file after the header row represents one movie, and has the following format:

movieId,title,genres

Movie titles are entered manually or imported from <https://www.themoviedb.org/>, and include the year of release in parentheses. Errors and inconsistencies may exist in these titles.

4. Trabajo a hacer

Se deberá realizar un trabajo de procesado de datos que al menos debe contener las fases de carga, tratamiento de valores perdidos, normalización de los datos cuando sea necesario y limpieza de *outliers*. Finalmente, los datos se deben almacenar de forma estructurada que facilite la consulta propuesta. Para ello debe hacer las siguientes suposiciones:

- Los datos son actualizados a las 3:00AM cada día
- La consulta que se pretende implementar a partir de los datos es una que retorne la lista de géneros, ordenados por el promedio de puntuación que han obtenido en la última semana (Nota: suponga que los datos son apropiados para esto; es decir, no se preocupe porque los datos más modernos tengan ya un año de antigüedad).
- La lista de géneros debe ser determinada a partir de los datos

Se pide crear un script utilizando Apache Airflow que todos los días coja los datos actualizados, realice el procesamiento necesario y genere un nuevo fichero (pelis_procesadas.csv) que permita realizar la consulta descrita arriba de forma eficiente. Debe implementar en Python todos los scripts necesarios para este procesamiento. No hace falta que implemente el código de la consulta, pero puede hacerlo para argumentar la conveniencia de la estructura final generada.

Tan importante como implementar los pasos necesarios es justificar cada una de las decisiones tomadas: si se normalizan los valores de los atributos o no, qué se considera valor perdido, qué se hace con ellos, qué formato tiene el fichero generado, etc. El informe debe incluir también estadísticas de valores perdidos y *outliers* encontrados en los datos, así como estadísticas globales antes y después de cada transformación.

Puede entregar su informe directamente a través de un notebook Jupyter.

El trabajo puede ser hecho en pareja o de forma individual.