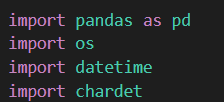
Universidad del Valle de Guatemala  
Data Preparation  
Alexis Fernando Hengstenberg Chocooj

**Proyecto Final**

**Algoritmo**

Se utilizo el lenguaje de programación de Python, el cual es muy utilizado para realizar este tipo de tareas, a continuación, una explicación del algoritmo.

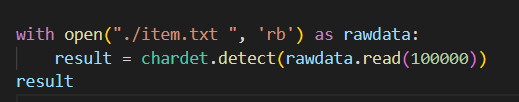
Se utilizaron las siguientes librerías:



Las cuales son fundamentales para este trabajo.

* Pandas: librería utilizada para la manipulación y transformación de datos.
* Os: librería utilizada para leer los archivos en un directorio.
* Datetime: librería para trabajar con fechas.
* Chardet: se utilizo para detectar la codificación de los archivos de texto y poder abrirlos.

En el siguiente fragmento utilizamos la librería de chardet para leer que tipo de codificación tienen los archivos de texto y poder convertirlo y utilizarlo para nuestra investigación.



A continuación, leemos el archivo item.txt y le asignamos encabezados, lo mismo sucede con el archivo de uses. Eliminamos los datos duplicados y eliminamos la columna que no utilizaremos en este caso no\_used

A screenshot of a computer

Description automatically generated with medium confidence

A continuación, leemos todos los archivos CSV que contienen las calificaciones y los unimos para tener la información disponible. Le eliminamos los duplicados y las calificaciones que sean mayores a 6.

Text

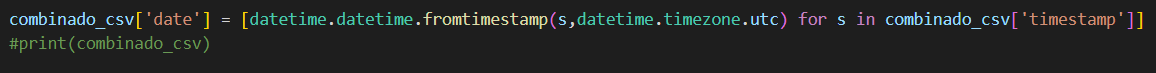
Description automatically generated

Luego asociamos los datos de las películas, los usuarios y las calificaciones para tenerlos completos.

Text

Description automatically generated

Luego convertimos la fecha a una estructura mas accesible y legible para todos los usuarios.

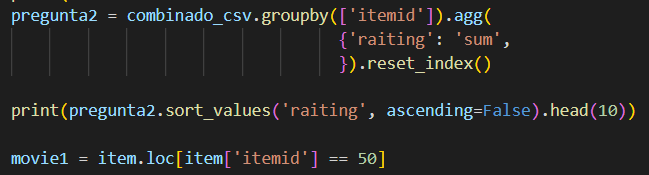


Para finalizar utilizamos distintas funciones que tiene la librería Python para ir respondiendo cómodamente las pregunta ya con la información limpia y preparada.

Para la pregunta 1 se agrupan por la película y luego se cuentan todas las calificaciones para saber cual fue la mas popular, luego se ordena de mayor a menor para tener las mas populares en la lista.



Para la pregunta 2 se crea una variable nueva que es agrupada por la película y a su ves se va sumando el rating para encontrar las películas con las calificaciones mas altas.



Luego se intento realizar el mismo procedimiento, pero en vez de suma, quisimos obtener el promedio para tener una mejor idea, pero podemos observar que al existir películas con 1 sola calificación puede afectar en la investigación y se decidió no incluir estos datos.

Timeline

Description automatically generated

Para la pregunta 3 al ser una variable de unos y ceros, se puede contar fácilmente la columna de cada una de las películas que se calificaron en cada género.

A screenshot of a computer

Description automatically generated with low confidence

En la pregunta número 4 se realizó el mismo procedimiento anterior, pero en vez de contar o sumar la columna por cada genero se agruparon por género y se sumó el rating para saber que genero tenía mejores calificaciones.A screenshot of a computer

Description automatically generated with medium confidence

En la pregunta 5 solo se calculo el promedio de las edades de las personas que realizaron una calificación.



Para lograr responder la pregunta 6 se realizaron distintas gráficas, para ver el comportamiento de las personas y si tenia algo que ver la edad con la calificación que daban, para esto se utilizó la librería de matplot para generar graficas.

Graphical user interface, text

Description automatically generated

Continuamos con la pregunta numero 7 y para aquí se eliminaron las calificaciones de personas que eran menores de 50 años para luego agrupar por película y ver que película tenía más calificaciones.

A screenshot of a computer

Description automatically generated with medium confidence

En la penúltima pregunta se utilizaron los datos combinados de usuarios, películas y calificaciones para ver cuantas calificaciones tenían las películas, pero estas utilizadas agrupadas por género.

Text

Description automatically generated

Por último, para la pregunta numero 9 se utilizo la columna de la fecha ya convertida con anterioridad para facilitarnos encontrar los fines de semana, con la librería de date time podemos identificar día de la semana es una fecha en específico, en este caso eliminamos los que días que no fueran sábados ni domingos, luego eliminamos las horas que estén fuera del horario de 8 de la mañana a 10 de la noche y agrupamos las calificaciones por película y sumamos todas para ver cuáles eran las más populares.

Text

Description automatically generated

**Preguntas**

* **¿Cuáles son las 3 películas más populares? Es decir, las que más calificaciones han recibido, sin importar la calificación otorgada.**

Table

Description automatically generated with medium confidence

Las películas mas populares se encuentran con el código 50, 258, 100

Text

Description automatically generated

Las películas mas populares por nombre son:

* + Star Wars (1977)
  + Contact (1997)
  + Fargo (1996)
  + Return of the Jedi (1983)

En ese orden.

* **¿Cuáles son las 10 películas mejor calificadas por los usuarios?**
  + Top 10 de pelicumas con mejores calificaciones sumadas

Table

Description automatically generated with low confidence

* + **Son las siguientes :**
  + Star Wars (1977)
  + Fargo (1996)
  + Return of the Jedi (1983)
  + Contact (1997)
  + Raiders of the Lost Ark (1981)
  + Godfather, The (1972)
  + English Patient, The (1996)
  + Toy Story (1995)
  + Silence of the Lambs, The (1991)
  + Scream (1996)

Text

Description automatically generated

Para realizar un promedio de las calificaciones se obtiene una mayor distorsión ya que las películas con menos calificaciones y con mejor calificación pueden llegar a salir primero como en el siguiente ejemplo

Table

Description automatically generated

Text

Description automatically generated

Con lo cual no puede ser fiable confiarnos del promedio. A pesar de que mientras más calificaciones tenga una película puede beneficiarse con la suma de la calificación es mucho mas confiable el primer análisis.

* **¿Cuántas películas hay de cada género? Si una película pertenece a más de un género, puede contarse más de una vez.**

En total de películas de cada genero son las siguientes:

Text

Description automatically generated

* **¿Qué género de películas tiene las calificaciones más altas?**

El género que tiene calificación más alta es drama.

Text

Description automatically generated

* **¿Cuál es la edad media de todos los usuarios que calificaron las películas?**
  + La edad media de todas las personas que generaron una calificación es de 34 aproximadamente

Graphical user interface, text

Description automatically generated

* **¿Existe alguna relación entre la edad del usuario y la calificación que otorgó a la película?**
  + No existe ninguna relación entre la edad y la calificación esto se puede ver en la siguiente grafica donde nos damos cuenta de que la dispersión es igual para todas las edades.

Table

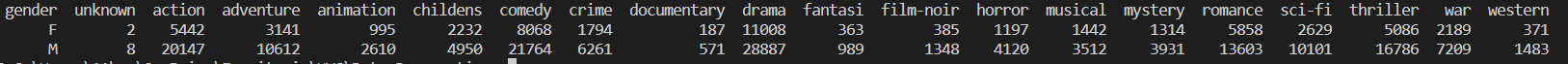
Description automatically generated

* **¿Cuáles son las películas más populares entre los usuarios mayores de 50 años?**
  + Las películas mas populares para los usuarios mayores de 50 años son:

Text

Description automatically generated

* + English Patient.
  + Fargo
  + Air Force One
  + Star Wars
* **¿Hay algún género de película que sea más popular entre los usuarios mujeres que entre los usuarios hombres?**



Podemos ver tenemos mayores calificaciones de hombres que de mujeres, y que en la mayoría de las películas los hombres han dado mas calificaciones que las mujeres, lo que nos dice que los hombres ven más películas. Y que en géneros como acción aventura si existe una diferencia visible, así como en algunos otros.

* **¿Cuál es la película mejor calificada durante fines de semana, en horario de 8 a 10 PM?**

**La película mejor calificada durante los fines de semana es**

A picture containing text

Description automatically generated

La película mejor calificada en horarios del fin de semana entre un horario de 8 de la mañana a 10 de la noche son:

* + Star Wars.
  + Fargo.
  + Twelve Monkeys.