Basile Álvarez Andrés José

No. Cuenta: 316617187

Email: andresbasile123@gmail.com

Fecha: 30/09/21

Inteligencia Artificial

Grupo III

Semestre 2022-1

Reporte Práctica 1: Reglas de Asociación

Objetivo: Obtener reglas de asociación a partir de datos obtenidos de plataformas de películas, donde los clientes pueden comprar o rentar este tipo de contenidos.

Características Generales: El generar reglas de asociación puede ser un factor importante para incrementar las ganancias ya que, al conocer los hábitos y relaciones de compra/renta en las películas de la plataforma, podemos hacer recomendaciones pertinentes que lleven al usuario a consumir más productos. Algoritmos como el presentado a continuación (algoritmo *apriori*) son utilizados por sitios como Netflix, Spotify o Amazon como motor de recomendaciones.

El algoritmo se detiene cuando se cumplen todos los parámetros del soporte, elevación y confianza, obteniendo un conjunto depurado de reglas significativas que posteriormente se pueden utilizar para los sistemas de recomendación.

Desarrollo

Primeramente, tenemos que definir aquellas bibliotecas de Python que nos serán útiles para importar, limpiar y analizar los datos contenidos en el archivo separado por comas *movies.csv.* Estas serán: *pandas* (manipulación y análisis de datos) y *matplotlib* (para la creación de gráficas y visualización de los datos). Además, requerimos la biblioteca *apyori* para poder aplicar el algoritmo *apriori* y obtener las reglas de asociación de los datos.

Más tarde, importamos el archivo *movies.csv* y lo primero que hacemos es guardarlo en un DataFrame, darle formato indicando que no tiene *header* y convertirlo a una lista de dimensión desconocida para poder observar la distribución de la frecuencia de los elementos. Una vez que tenemos los elementos en la lista, creamos un DataFrame y agregamos la columna “Frecuencia”, la cual inicializamos con 1, pero después nos servirá para contar cuántas veces aparece cada elemento. Realizamos la cuenta de los elementos y posteriormente dividimos la frecuencia de cada uno entre el total de datos para obtener el porcentaje de aparición de cada elemento, como se muestra en la figura 1.

```
# Crear matriz (DF) usando lista y agregar columna frecuencia
Lista = pd.DataFrame(Transacciones)
Lista['Frecuencia'] = 1  # valor que después se reemplazará, es nada más para agregar la columna.

#Agrupamos los elementos
#Conteo
# By=[0] empieza en cero.
# as_index para que no aparezca el nombre
Lista = Lista.groupby(by=[0], as_index=False).count().sort_values(by=['Frecuencia'], ascending=True)
Lista['Porcentaje'] = (Lista['Frecuencia']/Lista['Frecuencia'].sum())
Lista = Lista.rename(columns={0:'Item'})

Lista #vemos que el máximo de apariciones en porcentaje es el 6%... esto nos será útil para el soporte
```

Figura 1: Cálculo de la frecuencia de cada elemento y su porcentaje de aparición.

Posteriormente, realizamos una gráfica utilizando *plt.figure* para observar la distribución de la frecuencia de compra/renta de cada una de las películas. En la gráfica, observamos que la película *Ninja Turtles* es la que tiene un mayor número de visualizaciones, mientras que *Vampire in Brooklyn* tiene el menor número de visualizaciones.

	Item	Frecuencia	Porcentaje
106	Vampire in Brooklyn	3	0.000102
63	Lady Bird	5	0.000171
34	Finding Dory	7	0.000239
11	Bad Moms	14	0.000478
118	water spray	29	0.000990
...	...	...	...
25	Coco	1229	0.041944
44	Hotel Transylvania	1280	0.043685
103	Tomb Rider	1305	0.044538
37	Get Out	1346	0.045937
75	Ninja Turtles	1785	0.060919

Figura 3: Impresión de la frecuencia de visualización de cada película.

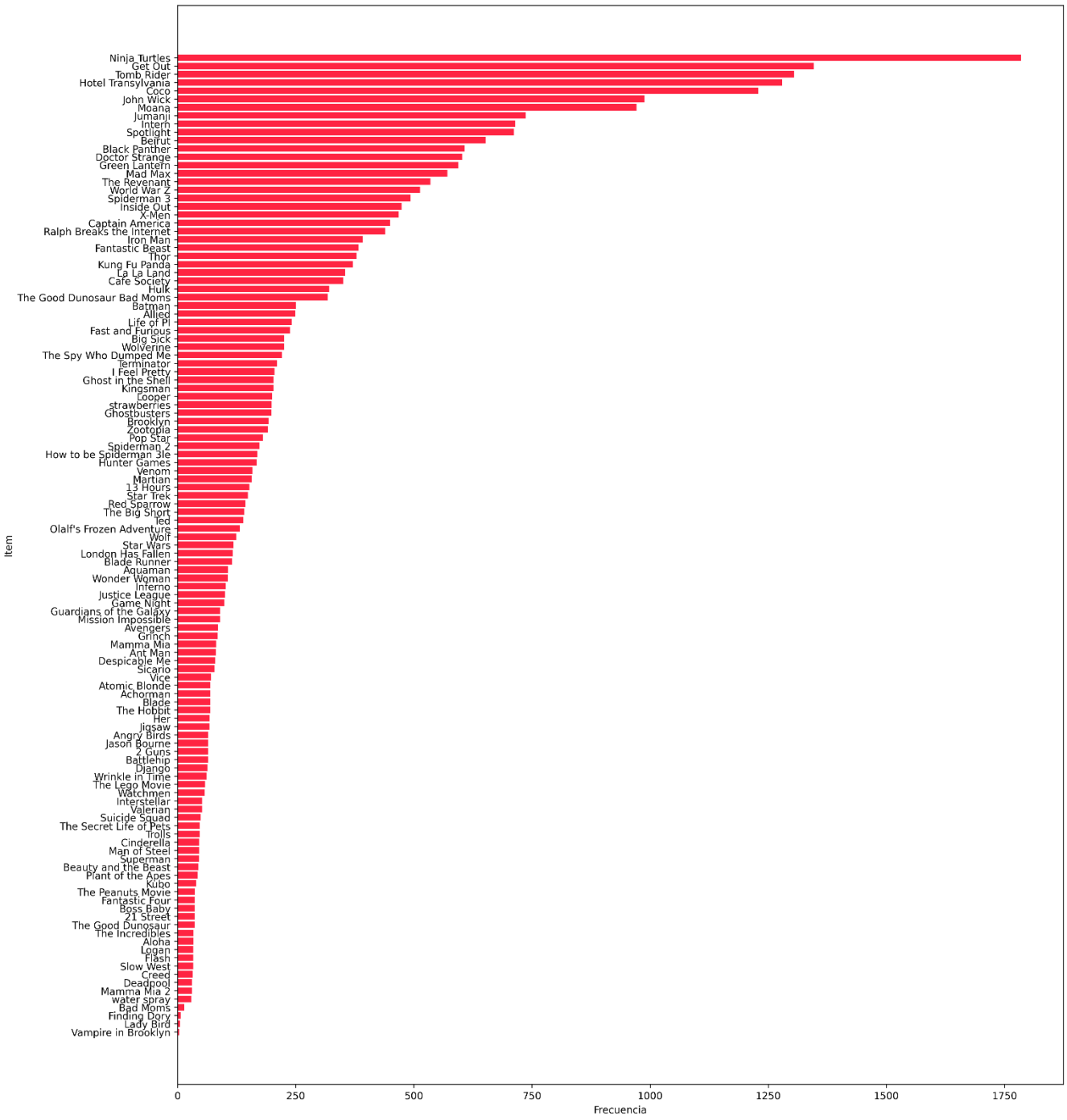

Figura 3: Gráfica de frecuencia de visualización de cada película.

Para poder aplicar el algoritmo *apriori*, requerimos hacer una preparación previa de los datos, agrupándolos en una lista de listas donde removemos los datos que aparecen como *NaN*, utilizando el método *stack()*.

```
#Removiendo "NaN" y creando lista de listas a partir del dataframe.
#level=0 especifica desde el primer índice.
#stack, los que tienen una cadena... quitar los NaN
MoviesLista = DatosPeliculas.stack().groupby(level=0).apply(list).tolist()
MoviesLista
```

Figura 4: Preparación de los datos para la aplicación del algoritmo apriori.

Primera configuración

Una vez que los datos se encuentran preparados, se definen las características de la primera configuración sobre la que ejecutaremos el algoritmo *apriori*. En esta primera configuración, se obtendrán reglas para aquellas películas que se visualizan al menos 10 veces al día (70 veces a la semana). Por lo tanto, el soporte mínimo será de 70/7460 (1%). La confianza mínima para las reglas en esta configuración fue definida como el 30% y la elevación mínima de 2. Estos valores fueron elegidos arbitrariamente y, para ellos, obtuvimos 9 reglas de asociación mostradas a continuación:

Figura 5: Reglas de asociación generadas con la primera configuración.

```
Regla: frozenset({'Jumanji', 'Kung Fu Panda'})
Soporte: 0.016087947446038343
Confianza: 0.32345013477088946
Lift: 3.278008906597914
====================================
Regla: frozenset({'Tomb Rider', 'Jumanji'})
Soporte: 0.03941547124279394
Confianza: 0.39945652173913043
Lift: 2.2831771614192906
====================================
Regla: frozenset({'Moana', 'Thor'})
Soporte: 0.015283550073736427
Confianza: 0.3007915567282322
Lift: 2.310611968729026
====================================
Regla: frozenset({'Tomb Rider', 'Terminator'})
Soporte: 0.010323099611207937
Confianza: 0.36492890995260663
Lift: 2.0858273864647456
====================================
Regla: frozenset({'Ninja Turtles', 'Jumanji', 'Get Out'})
Soporte: 0.01018903338249095
Confianza: 0.5066666666666667
Lift: 2.117213818860878
====================================
Regla: frozenset({'Ninja Turtles', 'Intern', 'Moana'})
Soporte: 0.011127496983509854
Confianza: 0.30970149253731344
Lift: 2.3790560585332865
====================================
Regla: frozenset({'Ninja Turtles', 'Jumanji', 'Moana'})
Soporte: 0.011127496983509854
Confianza: 0.503030303030303
Lift: 2.1020185043714457
====================================
Regla: frozenset({'Tomb Rider', 'Ninja Turtles', 'Jumanji'})
Soporte: 0.017160477275774234
Confianza: 0.4169381107491857
Lift: 2.38309683377638
====================================
Regla: frozenset({'Tomb Rider', 'Ninja Turtles', 'Spiderman 3'})
Soporte: 0.010323099611207937
Confianza: 0.3737864077669903
Lift: 2.136454264776997
====================================
```

Para la primera regla obtenida, vemos que aparecen los elementos *Kung Fu Panda* y *Jumanji.* Lo anterior tiene sentido debido a que es probable que una persona que ve películas familiares de corte infantil veo otras películas del mismo tipo. El soporte de esta regla es de 1.6%, con confianza de 32% y elevación de 3.27.

Es importante mencionar que el soporte hace referencia a cuán importante es una regla dentro del total de transacciones, la confianza indica la fiabilidad de una regla y la elevación el nivel de relación entre el antecedente y consecuente de la regla (nivel de posibilidad o aumento de posibilidad de que ocurra lo que menciona la regla nuevamente). Describamos entonces la segunda regla de asociación. Para este caso, obtuvimos la relación entre *Jumanji* y *Tomb Raider,* presentando un soporte de 3.9%, una confianza de 39.9% y una elevación de 2.28. Esta relación tiene sentido porque aquel individuo que vea la película de Jumanji, la cual tiene que ver con aventuras y acción, posiblemente verá la película de *Tomb Raider,* la cual también trata de acción y aventuras.

Por otro lado, también obtuvimos la regla de asociación entre *Tomb Raider* y *Terminator*. Esta regla obtuvo un soporte del 1.03%, con una confianza del 36% y una elevación de 2.08. Esta relación tiene sentido porque aquellos que vean la película de acción y aventuras *Tomb Raider*, probablemente también querrán ver la película de acción *Terminator.*

Otra de las reglas obtenidas involucra las películas *Tomb Raider, Spiderman 3* y *Ninja Turtles.* Esta regla tuvo un soporte del 1.03%, una confianza de 37% y una elevación de 2.13 y tiene sentido ya que las tres películas pertenecen al género de acción y existe una fuerte relación entre las últimas dos ya que ambas tienen un corte infantil.

Segunda configuración

Para la segunda configuración, obtuvimos reglas para aquellas películas que se hayan visto al menos 210 veces a la semana, por lo que el soporte mínimo será de 210/7460 o 2.8%. La confianza mínima para las reglas es de 30% y la elevación será mayor a 1. Para este conjunto de características, obtuvimos ocho reglas de asociación que se muestran a continuación:

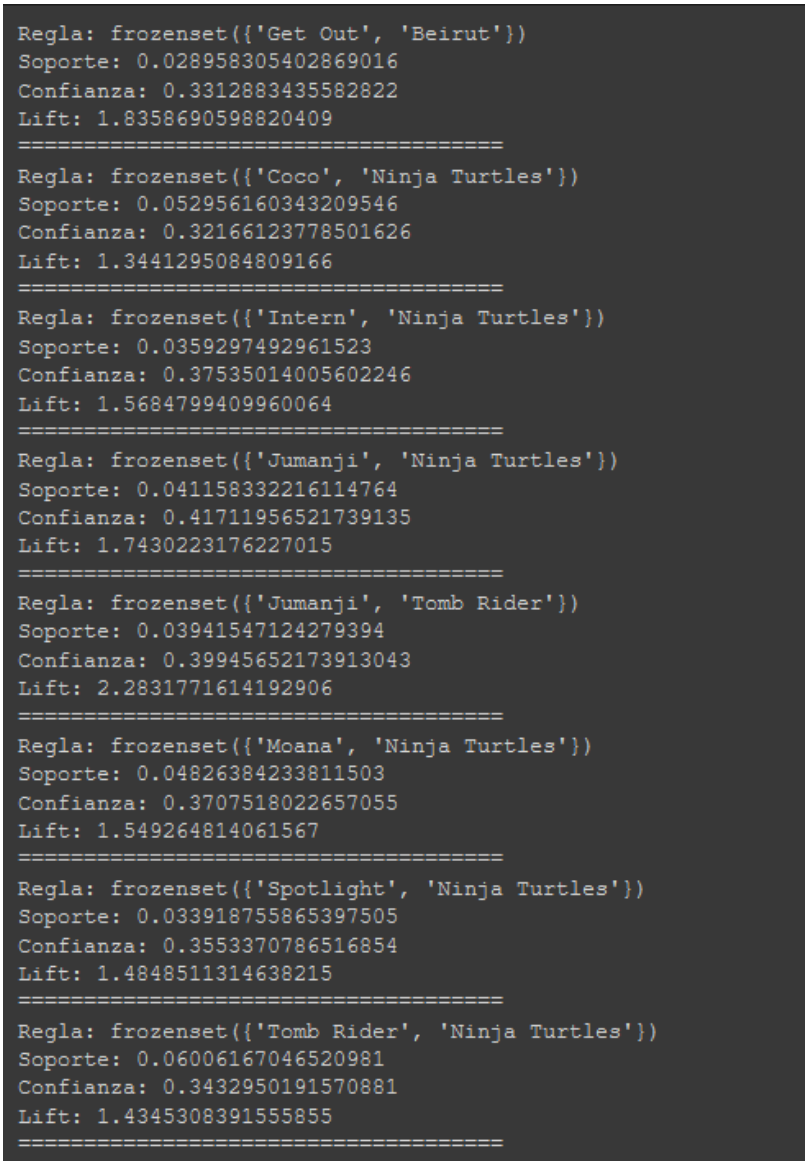

Figura 6: Reglas de asociación para la segunda configuración utilizada.

En este caso, podemos interpretar que, para la primera regla, por ejemplo, es lógico pensar que aquellas personas que vean la película *Beirut* de espionaje pueden tener gustos afines con películas de terror, como *Get Out*, con un soporte de 2.8%, confianza del 33% y elevación de 1.83.

La segunda regla de asociación obtenida involucra a las películas *Coco* y *Ninja Turtles.* En este caso, se obtuvo un soporte del 5.2%, una confianza del 32.1% y una elevación de 1.34. Lo anterior tiene sentido ya que ambas películas son de corte familiar/infantil y, probablemente, aquellos que vean la primera querrán ver una película de corte similar.

Nuevamente obtuvimos la regla *Jumanji – Tomb Raider*, ahora con un soporte del 3.9%, una confianza del 39.9% y una elevación de 2.28. Esta regla es lógica debido a lo que mencionamos en la primera configuración, pero es interesante ver que se repita para ambas configuraciones.

También encontramos la regla *Moana – Ninja Turtles.* Esta regla obtuvo un soporte del 4.8%, con una confianza del 37.07% y una elevación de 1.54. Esta regla tiene sentido debido a que ambas películas son de corte infantil o familiar y, bajo el supuesto de que un individuo quiera ver la primera película, se puede concluir que también querrá ver la segunda por ser del mismo género.

Conclusiones

A partir de lo realizado en la práctica, podemos concluir que los sistemas de recomendación permiten filtrar contenido y darle una cierta valoración que representa información que puede ser de mayor interés para ciertos usuarios.

Al utilizar el algoritmo *apriori*, logramos obtener reglas de asociación a partir de datos obtenidos de plataformas de películas, donde los clientes pueden comprar o rentar este tipo de contenidos. Las reglas de asociación obtenidas muestran cierta tendencia de usuarios hacia la visualización de contenidos relacionados, generalmente por género (por ejemplo, un usuario que ve *una* película de acción probablemente vea *otra* película de acción).

No obstante, creo que también obtuvimos reglas de asociación que, en un principio, no parecen tener demasiada relación, como la que relaciona la película de *Ninja Turtles* con la película *Intern* con un soporte de 3.5% y una confianza y elevación de 37% y 1.568, respectivamente; resultado que no consideraría muy obvio en la realidad.

En general, las reglas de asociación obtenidas representan fielmente lo que podrían ser las tendencias de visualización de películas por parte de los usuarios y, siempre y cuando se manejen los datos de manera adecuada, aplicando el algoritmo correctamente, podremos obtener información de gran utilidad. Utilizar un algoritmo como el anterior en un sistema real podría llevar a generar más ganancias para "empujar" o convencer a ciertos usuarios de rentar o comprar otras películas que pueden ser de su interés (o mantener una suscripción a un servicio de *streaming* como lo es *Netflix*).

Sobre dónde se encuentra el aprendizaje en este algoritmo, podríamos decir que el aprendizaje se encuentra en la forma en como se quitan los elementos menos significativos, quedándonos únicamente con unas cuantas reglas o patrones de verdadera importancia.

En conclusión, puedo decir que se cumplieron con los objetivos de la práctica y que esta fue útil para comprender de mejor manera el algoritmo *apriori* para obtener reglas de asociación en un conjunto de datos.