



RECOMMENDATION

27 DE SEPTIEMBRE DE 2022

PRÁCTICA 1

Reglas de Asociación - Algoritmo Apriori



Objetivo de la práctica

Por Angel Damian Monroy Mendoza

No. de Cuenta: 316040707

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



Características:

- Por lo general, existe un patrón en lo que ven los clientes. Por ejemplo, superhéroes en la categoría para niños.
- En este sentido, se pueden generar más ganancias, si se puede identificar la relación entre las películas. Esto es, si las películas A y B se rentan juntas, este patrón se puede aprovechar para aumentar las ganancias.
- Las personas que rentan una de estas películas pueden ser empujadas a rentar o comprar la otra, a través de campañas o sugerencias dentro de la plataforma.
- En este sentido, cada vez es común familiarizarse con los motores de recomendación en Netflix, Amazon, por nombrar los más destacados.



Desarrollo

Para el desarrollo de esta práctica dividimos el proceso en cinco secciones:

1. Importación de las bibliotecas necesarias.
2. Importación de los datos.
3. Procesamiento de los datos.
4. Implementación del algoritmo.
5. Análisis de tres reglas de cada configuración.



1. Bibliotecas

En este apartado cabe resaltar que se instaló la biblioteca *apyori* para, posteriormente, importar las bibliotecas necesarias para la manipulación de los datos, creación de vectores y matrices, generación de gráficas y el algoritmo *apriori*.

2. Datos

Posteriormente, se cargaron los datos dados por el profesor y se leyeron como archivo '.csv' con la biblioteca de pandas.

Aquí nos dimos cuenta que la lectura tomaba como encabezado la primera transacción de los datos, lo cual era algo que no se buscaba.

Asimismo, observamos que no había datos (NaN) en algunas columnas de las transacciones (filas), por lo que significaba que esa película no fue rentada o comprada en esa transacción.

Por último, se realizó un tratamiento para que la primera lectura no la tomara como encabezado con el parámetro *header=None*.

	0	1	2
0	The Revenant	13 Hours	Allied
1	Beirut	Martian	Get Out
2	Deadpool	NaN	NaN
3	X-Men	Allied	NaN



Desarrollo

3. Procesamiento

Para tener una mejor idea de nuestros datos, antes de aplicar el algoritmo revisamos la distribución de la frecuencia de los elementos.

Para realizar esto, pasamos todos nuestros datos a una lista y luego los convertimos en un *DataFrame*. Posteriormente, creamos una columna con el nombre de 'Frecuencia' para almacenar las veces que se vieron las películas del conjunto de datos.

Luego, realizamos un conteo a partir del índice 0 de los datos únicos para agruparlos, ordenarlos de menor a mayor y almacenarlos en la columna de 'Frecuencia'.

Después, creamos otra columna de nombre 'Porcentaje', en donde se almacenó el valor de la frecuencia que presentaba la película en cuestión entre la suma de las frecuencias de todas las películas.

En seguida, realizamos una gráfica en la que el eje X representa la frecuencia de la película y el eje Y el nombre de la película, con el objetivo de observar visualmente cuál era la magnitud de las vistas por películas y saber un aproximado del soporte mínimo que debería de tener nuestro algoritmo para que las reglas fueran representativas y relevantes.

Por último, el algoritmo *apriori* necesita que el conjunto de datos tenga la forma de una lista de listas, donde cada transacción es una lista interna dentro de una gran lista, por lo que se realizó el tratamiento necesario para realizar la lista de listas y, con el método '*stack*' se apilaron los datos que son cadena para asegurarnos de librarnos de los 'NaN'.





Watch It Again



Desarrollo

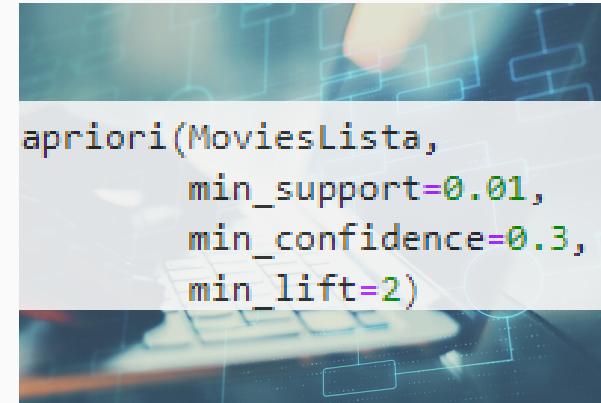
4. Algoritmo

Para obtener los parámetros de nuestro algoritmo, necesitamos definir tres propiedades:

- Soporte: el cual indica la importancia de la regla dentro del total de transacciones y, para este caso, se calcula estableciendo un número de renta de película a la semana (ej. 70 veces a la semana se rentó Moana) y se divide entre el número total de clientes
- Confianza: que indica la fiabilidad de la regla, en donde se propone el valor arbitrariamente, sin embargo, para tener un resultado significativo, la confianza mínima debe de ser del 20%
- Elevación: la cual indica el aumento de posibilidad entre el antecedente y consecuente en la que se busca que sea mayor que 1 para que la relación sea positiva.

Por lo que, una vez establecidos los parámetros que se utilizaron, se realizaron las siguientes tres configuraciones:

- **Configuración 1:**
 - Soporte: 0.01 (1%).
 - Confianza: 0.3 (30%).
 - Elevación: 2.
- **Configuración 2:**
 - Soporte: 0.028 (2.8%).
 - Confianza: 0.3 (30%).
 - Elevación: 1.1.
- **Configuración 3:**
 - Soporte: 0.047 (4.7%).
 - Confianza: 0.25 (25%).
 - Elevación: 1.1.





Desarrollo

```

for item in ResultadosC1:
    #El primer índice de la lista
    Emparejar = item[0]
    items = [x for x in Emparejar]
    print("Regla: " + str(item[0]))

#El segundo índice de la Lista
print("Soporte: " + str(item[1]))
```



```

#El tercer índice de la Lista
print("Confianza: " + str(item[2][0][2]))
print("Lift: " + str(item[2][0][3]))
print("=====
```

4. Algoritmo

Para cada una de las configuraciones se almacenaron las reglas obtenidas en una lista para saber la cantidad de éstas y poder imprimirlas. No obstante, es muy importante recalcar que **no basta** con imprimir las reglas obtenidas, sino también se tienen que interpretar para entender qué es lo que nos están diciendo y poder extraer información de allí.

En este sentido, otra forma de hacer más fácil la interpretación de las reglas, es realizando algún tipo de tratamiento a las listas para que se impriman de manera más ordenada, ya sea con un ciclo *for* como el que fue utilizado, o con algún *DataFrame* de pandas para darle mejor presentación a los datos.

5. Análisis*

Por último, se realizaron los análisis de tres reglas significativas de cada configuración para obtener información relevante con respecto al conjunto de datos. La forma de obtener estas tres reglas es observando la mayor elevación y la mayor confianza, pues esto nos asegura que sea fiable la regla y que sea posible que pase otra vez.

Conclusiones

De esta práctica podemos concluir que entre más alta sea la confianza y la elevación de una regla, es mejor para nuestro sistema de recomendación, por lo que podemos decir que son parámetros clave para decidir si una regla es significativa o no.

Asimismo, observamos que las reglas de asociación son un punto clave para los sistemas de recomendación, pues la información que nos otorgan dichas reglas, ofrecen un patrón de conducta de los clientes que sería muy difícil de identificar sin ayuda de este modelo y del algoritmo apriori.

