

Basile Álvarez Andrés José

No. Cuenta: 316617187

Email: andresbasile123@gmail.com

Fecha: 05/10/21

Inteligencia Artificial

Grupo III

Semestre 2022-1

Reporte Práctica 2: Reglas de Asociación

Objetivo: Obtener reglas de asociación a partir de datos obtenidos de un comercio minorista en Francia. Los datos son transacciones de un comercio de un periodo de una semana (7 días).

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 de artículos en un comercio, podemos hacer recomendaciones pertinentes, ofertas sobre ciertos productos o colocar productos que se compren juntos en el mismo lugar.

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, oferta u otros.

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 *store_data.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 *store_data.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 el agua mineral es el elemento más vendido, mientras que los espárragos son el elemento con menor número de ventas.

	Item	Frecuencia	Porcentaje
0	asparagus	1	0.000034
112	water spray	3	0.000102
77	napkins	5	0.000170
34	cream	7	0.000238
11	bramble	14	0.000477
...
25	chocolate	1230	0.041889
43	french fries	1282	0.043660
100	spaghetti	1306	0.044478
37	eggs	1348	0.045908
72	mineral water	1788	0.060893

Figura 3: Impresión de la frecuencia de compra de cada producto. Hubo 1788 transacciones con agua mineral entre sus productos.

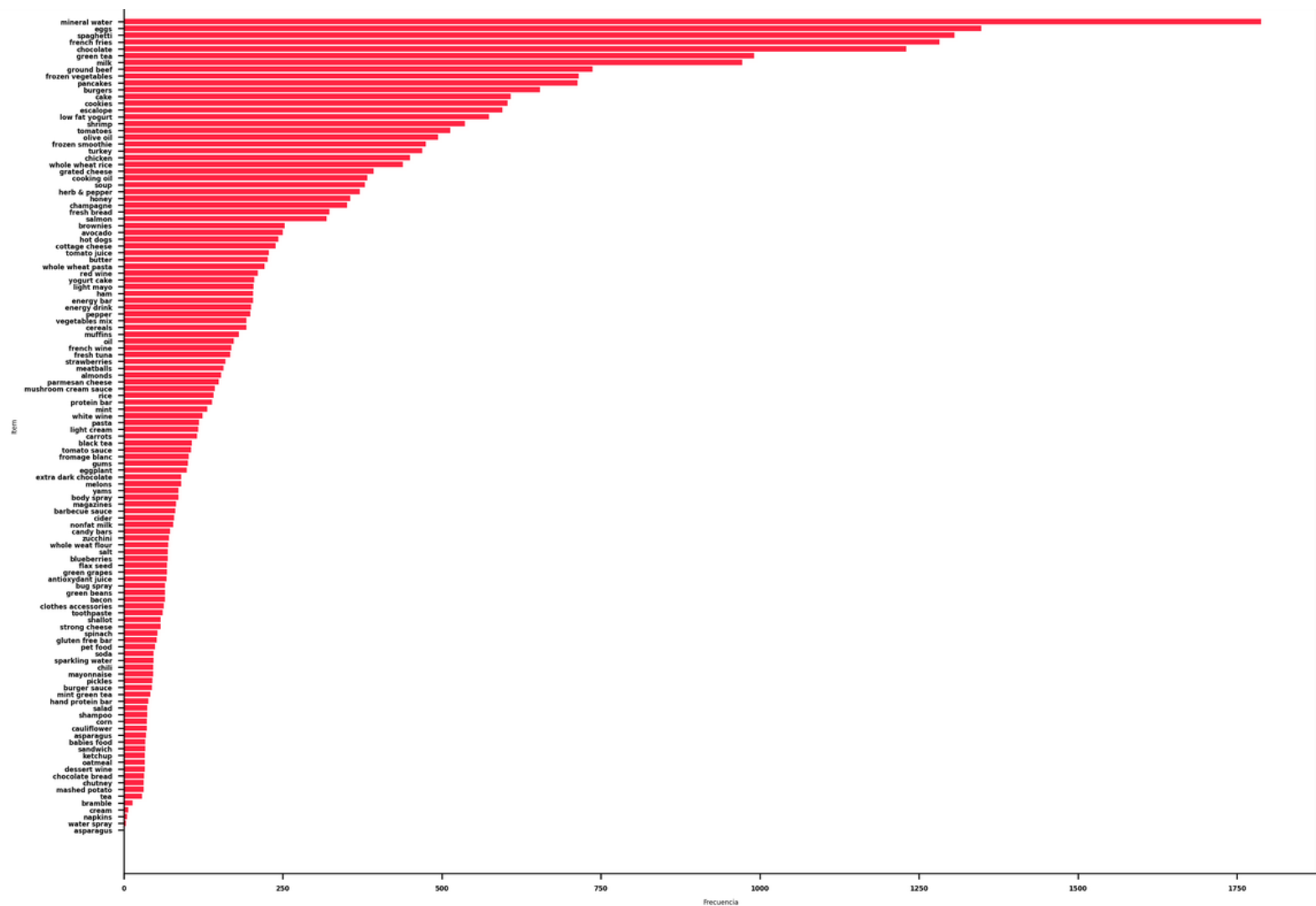


Figura 3: Gráfica de frecuencia de compra de producto.

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
TiendaLista = DatosTiendas.stack().groupby(level=0).apply(list).tolist()
TiendaLista
```

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 aquellos artículos que se compran al menos 5 veces al día (35 veces en una semana). Por lo tanto, el soporte mínimo será de 35/7500 (0.45%). La confianza mínima para las reglas en esta configuración fue definida como el 20% y la elevación mínima de 3. Estos valores fueron elegidos arbitrariamente y, para ellos, obtuvimos 24 reglas de asociación. A continuación, mostramos algunas de ellas:

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

```
Regla: frozenset({'chicken', 'light cream'})
Soporte: 0.004532728969470737
Confianza: 0.29059829059829057
Lift: 4.84395061728395
=====
Regla: frozenset({'mushroom cream sauce', 'escalope'})
Soporte: 0.005732568990801226
Confianza: 0.3006993006993007
Lift: 3.790832696715049
=====
Regla: frozenset({'escalope', 'pasta'})
Soporte: 0.005865884548726837
Confianza: 0.3728813559322034
Lift: 4.700811850163794
=====
Regla: frozenset({'ground beef', 'herb & pepper'})
Soporte: 0.015997866951073192
Confianza: 0.3234501347708895
Lift: 3.2919938411349285
=====
Regla: frozenset({'tomato sauce', 'ground beef'})
Soporte: 0.005332622317024397
Confianza: 0.3773584905660377
Lift: 3.840659481324083
=====
Regla: frozenset({'olive oil', 'whole wheat pasta'})
Soporte: 0.00798933475536596
Confianza: 0.2714932126696933
Lift: 4.122410097642296
=====
Regla: frozenset({'shrimp', 'pasta'})
Soporte: 0.005065991201173177
Confianza: 0.322038983050847
Lift: 4.506672147735896
=====
Regla: frozenset({'frozen vegetables', 'chocolate', 'shrimp'})
Soporte: 0.005332622317024397
Confianza: 0.23255813953488375
Lift: 3.2545123221103784
=====
Regla: frozenset({'spaghetti', 'cooking oil', 'ground beef'})
Soporte: 0.004799360085321957
Confianza: 0.5714285714285714
Lift: 3.2819951870487856
=====
Regla: frozenset({'spaghetti', 'frozen vegetables', 'ground beef'})
Soporte: 0.008665511265164644
Confianza: 0.31100478468899523
Lift: 3.165328208890303
=====
Regla: frozenset({'frozen vegetables', 'olive oil', 'milk'})
Soporte: 0.004799360085321957
Confianza: 0.20338983050847456
Lift: 3.088314005352364
=====
```

Para la primera regla obtenida, vemos que aparecen los elementos *chicken* y *light cream*. Lo anterior tiene sentido debido a que es probable que una persona que compra crema ligera tiene cuidado con lo que come, por lo que es probable que compre pollo en lugar de carne roja. El soporte de esta regla es de 0.45%, con confianza de 29% y elevación de 4.84.

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 *mushroom cream sauce* y *escalope*, presentando un soporte de 0.57%, una confianza de 30% y una elevación de 3.79. Esta relación tiene sentido porque aquel individuo que compra *mushroom cream sauce* probablemente lo haga para servirlo sobre un corte de pollo, cerdo o carne como el *escalope*.

Por otro lado, también obtuvimos la regla de asociación entre *escalope* y *pasta*. Esta regla obtuvo un soporte del 0.586%, con una confianza del 37.2% y una elevación de 4.7. Esta relación tiene sentido porque aquellos que compren carne como el *escalope*, probablemente quieran servirla junto con un poco de pasta.

Otra de las reglas obtenidas involucra los artículos *mineral wáter*, *ground beef* y *herb & pepper*. Esta regla tuvo un soporte del 0.66%, una confianza de 39% y una elevación de 3.9 y tiene sentido ya que los tres artículos pertenecen a un carrito de compra, por ejemplo, para realizar una carne asada.

Segunda configuración

Para la segunda configuración, obtuvimos reglas para aquellos productos que se hayan comprado al menos 210 veces a la semana, por lo que el soporte mínimo será de 210/7500 o 2.8%. La confianza mínima para las reglas es de 25% y la elevación será mayor a 1. Para este conjunto de características, obtuvimos diez reglas de asociación que se muestran a continuación:

```

Regla: frozenset({'eggs', 'burgers'})
Soporte: 0.02879616051193174
Confianza: 0.33027522935779813
Lift: 1.8378297443715457
=====
Regla: frozenset({'mineral water', 'chocolate'})
Soporte: 0.05265964538061592
Confianza: 0.3213995117982099
Lift: 1.3483320682317521
=====
Regla: frozenset({'mineral water', 'eggs'})
Soporte: 0.05092654312758299
Confianza: 0.28338278931750743
Lift: 1.188844688294532
=====
Regla: frozenset({'mineral water', 'frozen vegetables'})
Soporte: 0.03572856952406346
Confianza: 0.37482517482517486
Lift: 1.57246288387228
=====
Regla: frozenset({'mineral water', 'ground beef'})
Soporte: 0.040927876283162246
Confianza: 0.41655359565807326
Lift: 1.7475215442008991
=====
Regla: frozenset({'spaghetti', 'ground beef'})
Soporte: 0.03919477403012932
Confianza: 0.3989145183175034
Lift: 2.291162176033379
=====
Regla: frozenset({'mineral water', 'milk'})
Soporte: 0.04799360085321957
Confianza: 0.3703703703703704
Lift: 1.5537741320739085
=====
Regla: frozenset({'spaghetti', 'milk'})
Soporte: 0.03546193840821224
Confianza: 0.27366255144032925
Lift: 1.5717785592296398
=====
Regla: frozenset({'mineral water', 'pancakes'})
Soporte: 0.03372883615517931
Confianza: 0.3548387096774194
Lift: 1.4886158620191963
=====
Regla: frozenset({'mineral water', 'spaghetti'})
Soporte: 0.05972536995067324
Confianza: 0.2505592841163311
Lift: 1.4390851379453289
=====

```

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 compran huevos y hamburguesas lo hagan por la facilidad de preparar una comida con estos ingredientes. Esto con un soporte de 2.8%, confianza del 33% y elevación de 1.83.

La segunda regla de asociación obtenida involucra a los artículos *mineral water* y *chocolate*. En este caso, se obtuvo un soporte del 5.2%, una confianza del 32.1% y una elevación de 1.34. Vemos que esta regla tiene un porcentaje relativamente alto de soporte. Sin embargo, no alcanzo a notar una relación clara para realizar la compra de estos dos productos.

También obtuvimos la regla *eggs – mineral water*, con un soporte del 5.09%, una confianza del 28.3% y una elevación de 1.34. Esta regla es lógica porque aquellas personas que compran huevos, probablemente busquen un estilo de vida saludable que complementen con la compra de agua mineral en lugar de refrescos de cola.

También encontramos la regla *spaghetti – ground beef*. Esta regla obtuvo un soporte del 3.9%, con una confianza del 39.89% y una elevación de 2.29. Esta regla tiene sentido debido a que estos productos suelen servirse juntos en una comida. En realidad, vimos una regla muy parecida a esta (*escalope y pasta*) en la primera configuración, lo que refuerza la probabilidad de que un usuario compre estos artículos juntos.

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 compradores.

Al utilizar el algoritmo *apriori*, logramos obtener reglas de asociación a partir de datos obtenidos de un comercio minorista francés, donde los clientes pueden comprar distintos tipos de productos. Las reglas de asociación obtenidas muestran cierta tendencia de usuarios hacia la compra de artículos con los que se pueda cocinar (como ingredientes de un mismo platillo).

No obstante, creo que también obtuvimos reglas de asociación que, en un principio, no parecen tener demasiada relación, como la que relaciona *mineral water, frozen vegetables, milk y spaghetti* con un soporte de 0.45% y una confianza y elevación de 28.8% y 3.022, 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 compras de productos 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 a comprar productos que puedan ser de su interés o agregar promociones para que lleven estos productos (que ya compraban de manera regular) junto con algún otro producto similar.

Sobre dónde se encuentra el aprendizaje en este algoritmo, podríamos decir que el aprendizaje se encuentra en la forma en cómo 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.