TP2-Grupo 11

TRABAJO PRÁCTICO ENTREGABLE N° 2

Reglas de Asociación

Bertolini, Lucas; Di Paola, Matías; Vilca, Stella

Introducción

El presente trabajo práctico tiene como objetivo aplicar las técnicas de reglas de asociación en el conjunto de datos obtenido de la red social Twitter con el fin de encontrar asociaciones que permitan explicar patrones frecuentes en el contexto del COVID-19.

Dos cambios importantes diferencian a este del primer trabajo. Como se menciona arriba, se va a utilizar un algoritmo no supervisado que utiliza solo tipos de datos categóricos, por lo cual todas la variables numérica de interés debieron ser discretizadas. Esta tarea de preprocesamiento requirió de una importante cantidad de tiempo, ya que lo consideramos extremadamente importante para los subsiguientes pasos en la aplicación del algoritmo y obtención de resultados confiables.

El otro cambio no menor es que el volumen de tweets aumenta considerablemente, con lo cual nos encontramos muchas veces con recursos limitados para determinados tratamientos y análisis. Aquí también se invirtió una gran cantidad de tiempo en desarrollar "queries" eficientes que simplificarán lo más posible los cálculos en Rstudio y evitar colapsos por falta de recursos.

La mayoría de la variables estudiadas responden a la "Ley Zipf" con lo cual se aplicaron, dependiendo el caso, distintas transformaciones, principalmente normalizaciones y logaritmo. También en cada consigna se crearon nuevas variables a partir de las originales, para enriquecer la potencia explicativa de las reglas obtenidas.

Consigna 1

Variables seleccionadas

- "Creacion_retweet": Mañana / Tarde / Noche
- "Es_final_semana": Fin de semana / Laborable
- "Verificado": Verificado / Sin Verificar
- "Source"
- "n_char" (Cantidad de caracteres del tweet): Corto / Medio / Largo
- "Popular" (Ratio Followers/Friends[1]): Bajo[2] / Medio / Alto / Muy alto[3]
- "1_respuesta": Instantánea[4] / Rápida / Normal / Hace un rato / Desde el pasado[5]
    - Cantidad de minutos entre el primer retweet que capturamos y el momento en el que fue publicado el tweet original.
- "Adherencia": Baja / Intermedia / Alta / Muy_alta[6]

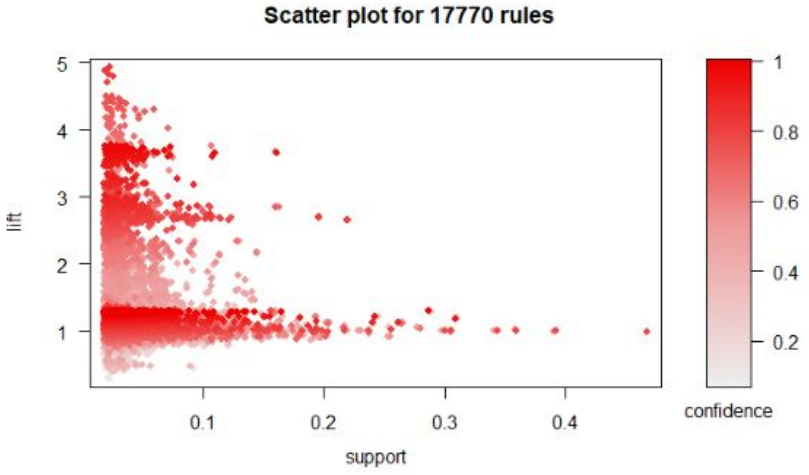
[1] Atributo creado en el trabajo práctico anterior.
[2] Bajo son aquellos que tienen hasta un seguidor por cada cinco personas que siguen.
[3] Muy alto son usuarios que tienen más de dos seguidores por cada persona que ellos siguen.
[4] Aquellas que se dan en los seis minutos siguientes a la creación del tweet.
[5] Con una ventana de tiempo mayor a las diez horas y media.
[6] Bajo son retweets con un máximo de tres favoritos y muy alto son con más de 160 likes.

    - Favoritos para el retweet. Dado que dentro del universo existen varios retweets del mismo contenido, capturas en distintos momentos, tomamos el mayor valor de favoritos.
- "Difusión": Muy_baja / Baja / Intermedia / Alta[7]
    - Retweets que tuvo el retweet.
- "Polaridad": Negativa / Neutro / Positivo
    - Polaridad del texto contenido en el retweet.

El cálculo de la polaridad se realizó con el paquete Syuzhet[8]. También se usó para un cálculo de sentimientos, pero que se descartó posteriormente. Como se comenta en la introducción, con el mayor volumen algunas operaciones se tornaron muy poco performantes. Fue en el momento de intentar incorporar análisis más complejos sobre el texto en que se armó un pequeño pipeline de preprocesamiento. Este punto fue bastante interesante desde lo técnico y ayudó a desacoplar algunas transformaciones y permitir que escalen mejor.

Primeramente se aplicaban la mayoría de operaciones en mongo. Agrupar por retweet, seleccionar fechas máximas y mínimas, cantidad de retweet y favoritos máxima. Se etiquetaba el momento del día y el día de la semana. Se aplicaban normalizaciones de tipo logarítmica y redondeo. Estas transformaciones se escribían en una nueva colección que era leída por un script en R, que asignaba la polaridad, sentimiento, y lo volcaba en un archivo csv.

El último tramo era el más abocado a la tarea del trabajo práctico, un notebook en R donde se discretizan las variables y se calculan transacciones y reglas.



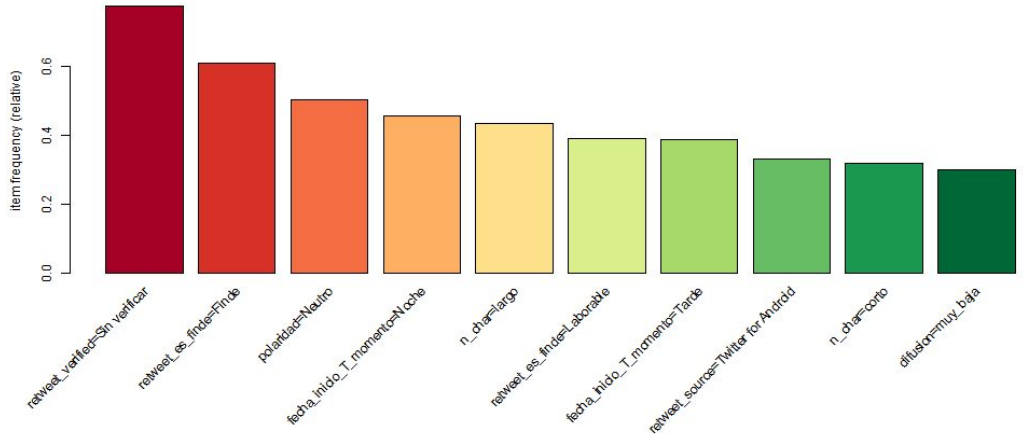
Nuestro universo en este punto son los retweets y constaba de 96.388 itemsets. En este primer inciso, tomamos como base la mayoría de las categorías que teníamos y que creamos, a modo de iniciar la exploración. Nuestros umbrales iniciales fueron una confianza mínima de 0.02 y un soporte de 0.05.

El objetivo de este punto inicial era entender algunas relaciones y el peso de algunas variables. Algunos de los desafíos que nos enfrentamos fueron: discretizar balanceando los bins, para no favorecer alguno, pero creando categorías que mantuvieran un cierto sentido; encontrar el equilibrio en la cantidad de variables, muy pocas variables nos exponían a tener muy pocas reglas, pero al crecer el número de variables aumentaba la superficialidad de la

[7] Una difusión muy baja es cuando son menos de tres retweets, y alta si el número supera los 32.
[8] https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html

información que nos aportan y se repetían mucho las relaciones, aunque con pequeñas variaciones.

Itemsets frecuentes



Reglas

Regla	S	C	Lift	N
{Verificado=Verificado} => {Popular=Muy alto}	0.16	0.72	2.85	15,585

La primera que elegimos es, muy probablemente, la más obvia. Los usuarios verificados tienden a ser populares y representan un séptimo de nuestra muestra, esto es una muy alta representatividad. Con el Lift podemos ver que las variables no son correlacionadas, sino que podemos hallar causalidad. Y podemos ver que la confianza es alta, lo cual nos dice que la probabilidad de que el usuario sea popular, dado que es verificado, es alta.

Regla	S	C	Lift	N
{Source=Twitter for Android, Verificado=Sin verificar} => {Popular=Bajo}	0.14	0.46	1.59	13,588

Esta regla es la contracara de la primera, donde quisimos ver algunos soportes sobre usuarios populares. Aquí lo intuitivo es ver que participan itemsets muy frecuentes. Lo que nos resulta útil aquí es ver que valor tienen las medidas de interés. Vemos que 1.59 es un buena medida para el lift, y que tenemos una buena representatividad de dos características que aparecen en combinación. Lo que sí bajó un poco fue la confianza, pero sigue siendo un buen indicador el 46%.

Regla	S	C	Lift	N
{Source=TweetDeck} => {Popular=Muy Alto}	0.06	0.68	2.67	5,395

TweetDeck fue el primer source que apareció cuando filtramos reglas sobre los más populares. Es una herramienta proporcionada por twitter para gestionar de una forma más profesional la red social. Es parecido a un panel de control.

Este dato no lo sabíamos inicialmente y fue la regla quien nos lo contó. Es un primer insight de información que nos cuenta el dataset, y que no parte de una creencia previa. Es una característica interesante para pensar que es lo que diferencia a los usuarios populares. Siendo dos items independientes y minoritarios en la proporción del dataset, la regla se presenta como una regla fuerte.

Regla	S	C	Lift	N
{Polaridad=Negativo, Verificado=Verificado} => {Popular=Muy Alto}	0.04	0.74	2.91	3,863
{Polaridad=Negativo} => {n_char=largo}	0.12	0.48	1.11	11,456

El grueso de las polaridades son neutras, y tanto los retweets positivos como negativos fueron muy similares en distribución. La primero fue pensar que cuando no fuese neutra, tal vez habría algo interesante. Elegimos estas dos reglas porque nos ayudaron a despejar esa hipótesis.

Podemos ver por la segunda regla, que muy probablemente, haya dificultad en el método usado para calcular el sentimiento en textos cortos. Está regla sumada a la versión positiva de la misma suman un coverage del 98%.

La primera regla parece más influenciada por verificado y por el consecuente, que por la polaridad. Por lo menos en este contexto, polaridad no aporta ninguna novedad o información extra.

Acá lo útil fue entender que la variable de polaridad sería difícil de explotar y también seguir mejorando los criterios para seleccionar, viendo cómo los distintos elementos se conjugan. Porque en este caso vemos dos reglas con buen soporte, con una confianza aceptable, y dos lift, que aún siendo diferentes, nos hablan de independencia de consecuente y antecedente, pero que analizandolas mejor nos generan dudas sobre la fuerza de la regla.

Regla	S	C	Lift	N
{1_respuesta=Desde el pasado ,Creacion_retweet=Noche, Es_final_semana=Fin de semana} => {Adherencia=Muy_alta}	0.03	0.35	2.08	2,432

Esta regla condensa dos partes que nos resultaban interesantes. Por un lado, en el antecedente, el cuándo, y por el otro, como consecuente que popularidad tuvo el contenido. Esta decisión se explica también a partir de las reglas circundantes. Filtramos reglas que tuvieran como antecedente “Desde el pasado”, retweets cuya primera respuesta capturada era con una diferencia mayor a diez horas.

Las redes sociales tienen una velocidad de creación de contenido muy alta en general. Particularmente en Twitter esto se exacerba, si uno usa la red social, es fácil sentir que uno está navegando un mar de texto. También se percibe cuando uno hace scrapping, el número de tweets escala muy rápido en cuestión de segundo. Por este contexto, debería haber algún valor intrínseco el tweet, o tal vez alguna cuestión temporal que permitiese al usuario navegar más tiempo para hallar un tweet más antiguo..

Las reglas que contenían a “Noche” y a “Finde”, presentaban mayor fortaleza que las reglas con “Tarde”, “Mañana” o “Laborable”. Al mismo tiempo, esta regla era distinta a aquellas que estaban próximas. Reglas que involucraron Verificados/Sin Verificar, que siendo una categoría binaria tomaba mucho protagonismo; o donde el consecuente era “Adherencia=Muy alta” pero dentro del antecedente estaba “Difusión=Alta” (Reglas con muchos retweets y muchos favoritos), lo cual no era muy novedoso.

La explicación que emerge es que los finales de semana, durante la noche, los usuarios tienden a tener más tiempo y a conectarse con contenido más allá de una primera inmediatez. Aunque tiene algo de esperable esta idea, la confianza no es tan alta, un 35%. En esta primera etapa, nos ayuda a calibrar en qué parámetros buscar lo novedoso, e intentar dibujar un poco mejor los límites de lo ‘obvio’.

Consigna 2

Para desarrollar esta consigna nos enfocamos en los tweets originales que fueron retweeteados, aquellos que cumplen con la condición que el atributo “is_retweet” sea igual a “true”.

Luego de una análisis orientado a los aspectos de la popularidad de los tweets, decidimos utilizar el atributo ”retweet_retweet_count” como indicador de la difusión del tweet y el atributo “retweet_favorite_count” como referente de la adhesión al tweet. Para ambos casos, seleccionamos los máximos valores debido a que para un mismo tweet se presentaban más de un valor de estos atributos. Consideramos que a mayor difusión y adhesión el tweet es mas popular y viceversa.

Obtuvimos como resultado un total de 96388 tweets para nuestro análisis

Adicionalmente, generamos las siguientes variables para analizar la popularidad de los tweets:

- primera respuesta al tweet: diferencia entre la fecha del primer retweet y la fecha de creación del tweet.
- actividad del tweet: cociente entre la cantidad de retweet que tuvo el tweet sobre la diferencia entre la última fecha de retweet y la primer fecha de retweet.

Consideramos que una respuesta rápida al tweet, realizando su retweeteo y una actividad alta equivalente a muchos retweeteos en poco tiempo, podrían ser indicadores de popularidad del tweet.

Para las variables de unidad tiempo utilizamos minutos como medida.

Todas las variables fueron transformadas con logaritmo en base 10 debido a que presentan distribuciones muy sesgadas. En particular, para las variables correspondientes a la adhesión al tweet y la diferencia entre la última fecha de retweet y la primer fecha de retweet fue necesario sumar un valor de 0.0001 previamente al cálculo del logaritmo en base 10 porque contienen valores iguales a cero.

En las discretizaciones de las variables, los valores extremos de las categorías fueron definidos a nuestro criterio teniendo en cuenta su importancia para la generación de las reglas e intentando mantener el equilibrio entre los bins. También, tuvimos en cuenta que se conserven las características de las distribuciones originales. A continuación, presentamos las categorías resultantes:

Atributos de los tweets:

- “Difusion” : { **muy baja**: < 6; **baja** : 7 - 50 ; **media**: 51 - 501; **alta**: > 502}
- “Adhesion” : { **muy baja**: < 10; **baja** : 11 - 99 ; **media**: 100 - 999; **alta**: 1000- 9997; **muy alta**: > 10000}
- “Primera respuesta”:{ **"rapidisima"** < 10 **,"muy rapida":** 10 - 100**, "rapida":** 101- 999**, "menos rapida" :** 1000-10000, **"lenta"** > 10000}
- “Actividad”:{ **baja** : < 0.01 ; **media** < 1; **alta**: > 1}

Para enriquecer el trabajo decidimos incorporar en el análisis el año de creación de la cuenta del usuario que retweetea el tweet. Identificamos que dichos años se encontraban entre 2006 y 2020.

Realizamos tareas de preprocesamiento para obtener un dataset que contenga una fila por cada tweet y una columna por cada año, desde 2006 a 2020. La columna de año se completa con “si” cuándo corresponde con el año de la creación de la cuenta del usuario que retweeteo el tweet, caso contrario se completa con “na ”. Adjuntamos un ejemplo de un tweet que fue retweeteado por usuarios cuyas cuentas se crearon en los años 2007 y 2020.

retweet_status_id	*año_2006*	*año_2007*	*año_xx*	*año_xx*	*año_xx*	*año_xx*	*año_2020*
1232734362833244161	na	si	na	na	na	na	si

Las variables de años se transformaron a tipo factor para ser utilizadas como categorías para la generación de reglas.

Previamente a la generación de la reglas se generó un dataset con las categorías de adhesión al tweet, difusión del tweet, primera respuesta, actividad del tweet y las correspondientes a los años.

Reglas

Para las reglas de asociación, creamos las transacciones con las categorías de la adhesión al tweet, difusión del tweet, primera respuesta, actividad del tweet y las correspondientes a los años de las creaciones de las cuentas de los usuarios que retweetearon los tweets, una por cada año.

Luego de varias pruebas, análisis y ajustes concluimos en emplear un soporte igual a 0.002 y una confianza igual a 0.6 para la generación de las reglas.

Como el objetivo era encontrar reglas que asocian los atributos de la popularidad del tweet, en las búsquedas aplicamos filtros para que el consecuente contenga a categorías de difusión y adhesión altas y el antecedente incluya alguna de las categorías correspondientes a la primera respuesta (rápida), actividad del tweet (alta) y los años de creación de la cuenta del usuario. Además, en todas nuestras búsquedas incluimos la condición que el lift sea mayor a 1.

A continuación, se detallan las reglas seleccionadas, ordenadas por su robustez:

Reglas	S	C	Lift	N
{cat_rpta=rapida,cat_adhesion_RT=alta,cat_2009=SI}=>{cat_difusion_RT=alta}	0.0039	0.765	18.36	382
{cat_rpta=rapida,cat_adhesion_RT=alta,cat_2009=SI,cat_2016=SI} => {cat_difusion_RT=alta}	0.0024	0.8	19.19	240
{cat_rpta=rapida, cat_adhesion_RT=alta, cat_2009=SI, cat_2015=SI} => {cat_difusion_RT=alta}	0.0025	0.8	19.42	242
{cat_actividad=alta,cat_adhesion_RT=alta,cat_2010=SI}=> {cat_difusion_RT=alta}	0.0020	0.74	17.87	269
{cat_actividad=alta,cat_adhesion_RT=alta,cat_2011=SI}=> {cat_difusion_RT=alta}	0.0020	0.72	17.47	276
{cat_actividad=alta,cat_difusion_RT=alta,cat_2019=SI}=> {cat_adhesion_RT=alta}	0.0020	0.71	15.75	229

Para definir este orden, primero seleccionamos las reglas que tuvieran un lift mayor a uno, que da cuenta que el consecuente y antecedente serían dependientes y están relacionados. Sobre las seleccionadas, evaluamos en conjunto los valores de la confianza y el soporte buscando un equilibrio que refleje la robustez de la regla.

Las valores de las confianzas de las reglas son superiores a 0.7. Consideramos que es un muy buen valor para representar la frecuencia de que el antecedente se presente en las transacciones en las que también aparece el consecuente.

Consideramos que los valores de soporte son muy bajos, la fracción de transacciones que contienen al antecedente y consecuente está aproximadamente entre el 0.2 % y el 0.3 %.

En cuanto al conocimiento que aportan las reglas, observamos:

- La adhesión y la difusión, variables que elegimos para la popularidad del tweet, se presentan en las reglas con categorías semejantes. Sería un posible indicador que los tweets con popularidad alta tienen difusión alta y adhesión alta.

- Se presenta una difusión alta cuando el antecedente es una adhesión alta, la respuesta al tweet es rápida (entre 10 y 100 minutos) y el año de creación de la cuenta de los usuarios que retweetean es 2009.

- Se observa una difusión alta cuando el antecedente es una adhesión alta, la respuesta al tweet es rápida (entre 10 y 100 minutos) y los años de creación de las cuentas de los usuarios que retwetean son 2009 y 2016. Esta regla es semejante a la anterior, tiene valores próximos de confianza, lift y soporte, sin embargo, decidimos seleccionarla porque agrega la información sobre las cuentas creadas en el año 2016.

- En las últimas tres reglas, se observa una difusión alta cuando el antecedente es una adhesión alta, la actividad del tweet es alta (más de un tweet por minuto) y los años de creación de las cuentas de los usuarios que retwetean son 2010, 2011 y 2019. Las reglas tienen valores semejantes para la confianza, lift y soporte y cada una de ellas nos aportan diferente información sobre los años de creación de las cuentas de los usuarios.

- La información que nos brindaron las reglas nos resultó muy interesante, teniendo en cuenta que las cuentas de los usuarios que retweetearon fueron creadas entre el 2006 y 2020.

Consigna 3

Para la presente consigna decidimos enfocarnos en la búsqueda de patrones de co-ocurrencia de diferentes atributos de los usuarios que permitan caracterizar y diferenciar el comportamiento “retweetear/retweeteado”.

Para los “retweeteros”, agrupamos al dataset por usuarios y contamos la cantidad de veces que un mismo usuario había retweeteado. Más del 75% de los usuarios que retweetearon lo hicieron una única vez. Para intentar detectar tendencias más marcadas se puso un umbral de corte de 5, resultando un total de 5355 usuarios que retweetearon 5 o mas veces. Lo mismo se hizo pero ahora con los más retweeteados, eligiendo un umbral

también de 5 veces (o mas) retweeteados, obteniendo un total 5191 usuarios. Para evitar ruido en el análisis se eliminaron 309 usuarios que pertenecían a ambos grupos.

Se trabajó en la selección y construcción de variables que describieran diferentes aspectos de los usuarios. Luego de sucesivas iteraciones se seleccionaron “followers_count” (cantidad de seguidores), “friends_count” (cantidad de amigos o personas seguidas), “favourites_count” (cantidad de “likes” o “me gusta” totales recibos) y también se construyeron dos nuevas: “actividad” como la cantidad de posteos por año (statuses_count / 2021 - account_created_at) y “afinidad” como la cantidad de ‘me gusta’ recibidos por posteo (favorites_count / statuses_count).

Todos las variables fueron transformadas con logaritmo en base 10 ya que todas responden a la ley Zipf y presentan distribuciones muy sesgadas. En el caso de la variable afinidad se tuvo que realizar previamente un re-escalado de tipo min-max (proyectando los valores entre 0-1) y luego aplicar el logaritmo ya que se observó que solo aplicando este último los valores seguían con un sesgo importante hacia la izquierda impidiendo realizar adecuadamente la discretización. Esto se debía a que había muchos valores menores a 1 y algunos mucho mayores.

Las discretizaciones se realizaron en todas las variables mencionadas arriba, pero sobre los usuarios totales presentes en el dataset (283.129). Se hizo de manera “manual”, como se mencionó previamente, intentando un equilibrio entre un tamaño de categoría (‘bin’) que tengan un sentido práctico/explicativo de análisis pero a la vez que las frecuencias de las observaciones ni sean muy altas que saturen la generación de reglas ni muy bajas de manera que tenga alguna probabilidad de co-ocurrencia con otras categorías de otras variables.

Atributos de los usuarios:

- “Seguidores” : { **baja** : < 50 ; **intermedio**: 50 -250; **alta**: 250- 1000, **muy_alta** : > 1000 }
- “Amigos” : { **baja** : < 100 ; **intermedio**: 100 - 500; **alta**: 500- 2000, **muy_alta** : > 2000 }
- “Favourites”: { **baja** : < 500 ; **intermedio**: 500-5000; **alta**: 5000-20000, **muy_alta** : > 20000 }
- “Actividad” (tw/dia) : { **baja** : < 1; **intermedio**: 1 - 3.3; **alta**:3.3- 13.3, **muy_alta** : > 13.3 }
- “Afinidad” : { **baja** : < 0.25 ; **intermedio**: 0.25 - 1; **alta**: 1- 2.5, **muy_alta** : > 2.5}

Reglas

Para que no queden dudas las reglas se generaron a partir de los dos grupos (con las mismas variables) por separado, y luego se eligieron 3 reglas de cada uno.

Los parámetros de búsqueda de reglas intentaron ser lo más restrictivos posibles, buscando la combinación más alta de mínimo soporte y lift de manera de encontrar reglas que mejor representaran a ambos grupos de estudio. A partir de los gráficos de “soporte vs lift” se fijó un minsupp entre 0.1-0.15 y un lift entre 2-4 para los “retweeteros” y para los “retweeteados” un minsupp mayor 0.05 y un lift mayor a 2.

Reglas				
+ Retweet(eros)	**S**	**C**	**Lift**	**N**

{seguidores=intermedia} => {amigos=intermedia}	0.136	0.61	2.13	729
{actividad=muy_alta, amigos=muy_alta, favourites=muy_alta} => {seguidores=muy_alta}	0.148	0.92	2.54	793
{afinidad=intermedia, seguidores=muy_alta} => {amigos=muy_alta}	0.104	0.63	2.51	560
+ Retweet(eados)	**S**	**C**	**Lift**	**N**
{actividad=muy_alta, afinidad=alta, seguidores=muy_alta} => {favourites=muy_alta}	0.069	0.92	2.75	161
{actividad=muy_alta, afinidad=intermedia,amigos=muy_alta} => {favourites=muy_alta}	0.115	0.83	2.48	267
{afinidad=alta,amigos=muy_alta, seguidores=muy_alta} => {favourites=muy_alta}	0.067	0.72	2.15	157

Lo primero que se puede advertir es que las reglas de los “+retweeteros” tienen un mayor soporte que los “+retweeteados” (excepto la segunda con 1.115) lo que marca una mayor robustez de las reglas de los primeros.

El lift para todas es mayor a 2, lo que marca asociación entre consecuente y antecedente, y no de manera azarosa. La confianza varía entre 0.6 y 0.9, un rango más que aceptable de confiabilidad de las reglas generadas.

+ Retweet(eros)

Se destaca de la primer regla que las variables seguidores y amigos toman el mismo valor. Esto nos sugiere que alrededor del 10% de los retweeteadores tienen seguidores y amigos en la misma proporción, y con valores no muy altos.

En la **segunda regla**, otra vez aparecen las variables amigo y seguidores con el mismo valor de categoría (proporción semejante), pero esta vez seguidores en el consecuente lo que marca la interdependencia entre una y otra, como si hubiese una lógica subyacente de “si vos me seguis yo te sigo”. Los otros dos términos del antecedente son muy alta actividad y muy alta cantidad de “me gusta” acumulados. Si bien esto podría estar generando mayor cantidad de seguidores, no implica que la afinidad sea alta, de hecho los muy altos valores de la variables explican, quizás, por que no aparece afinidad, ya que la mayor actividad implica más cantidad de posteos y eso “diluye” el efecto de los “me gusta”, y simplemente acumulan por la antigüedad de la cuenta.

En la **tercera regla**, vuelve a repetirse el par seguidores muy alta y amigos muy alta, pero esta vez aparece, ahora sí, afinidad con valor intermedio.

En conclusión la relación seguidores/amigos parecería estar cercana a 1, siguiendo la la logica “yo te sigo si vos me seguis”. Ambas variables pueden tomar valores de intermedios a muy altos. La actividad parecería ser muy alta, lo que explicaría una afinidad de intermedia para abajo, ya que aunque puede tener muchos “me gusta”, al tener alta actividad (mas posteos) los “diluye”.

+ Retweet(eados)

Se puede apreciar que en las 3 reglas en el antecedente se encuentra afinidad con valores de intermedio- alta, y en el consecuente la variable “favourites” con valores de la categoría más alta. Esto no debería sorprender ya que afinidad se construye a partir de favourites y están relacionadas de manera directa.

Sin embargo, altos valores de me gusta, tienen en el antecedente en las primeras 2 reglas una actividad muy alta, esto nos indica que los usuarias postean mucho, reciben muchos likes, pero los likes que reciben son proporcionalmente mayores a los posteos realizados, y por eso obtenemos valores de intermedios a altos en las dos primeras reglas.

También podemos observar la aparición de valores muy altos de amigos y de seguidores, pero lo hacen por separado (no en la misma regla), ni uno está implicando al otro como en el grupo de los “”+retweeteros”.

En conclusión, este grupo de usuarios tiene un alto número de me gusta, pero en mayor cantidad que los posteos que generan (que son muchos, “actividad muy alta”), haciendo que su afinidad sea de intermedia a alta. También se observan alto números de amigos y de seguidores, pero no parecería que están en la misma proporción como en el primer caso.

Consigna 4.1

Retomando lo trabajado en las consignas 2 y 3, la pregunta KDD que se intentó explorar fue que influencia tenían las variables de los usuarios definidas en la consigna 3 sobre la popularidad de los retweets. Para intentar ganar claridad en el análisis se redefinió popularidad como una combinación de las variables “retweet_retweet_count” y retweet_favourites_count”

Nos enfocamos en todos aquellos usuarios que fueron retweeteados pero que además cumplían con la condición de pertenecer al dataset original. Por lo tanto, el total de transacciones analizadas fueron 19.074.

Para definir popularidad del tweet se tomaron los retweets de estos usuarios previamente seleccionados, y se tomaron los valores más grandes (si es que había más de un tweet por usuario) de “retweet_retweet_count” y de “retweet_favourite_count”. De alguna manera estas dos variables reflejan cuán difundido y cuanta adherencia genera el tweet publicado y que juntas consideramos que podían definir la “popularidad” de un tweet. En este caso como consideramos que cuanto más adherencia y más retweetiado un tweet es, más popularidad tiene, entonces normalizamos ambas variables y las promediamos para generar la nueva variable. Previamente normalizamos ambas con min-max, ya que si lo hubieramos hecho por zscore obtendriamos valores negativos con lo cual no se podría aplicar logaritmo posteriormente.

Atributo del retweet

- “Popularidad” : { **baja**; **intermedia**; **alta** }

Reglas

Como el objetivo era descubrir reglas que asocian los atributos del usuario con la popularidad del tweet, nos interesaban aquellas que en el consecuente tuviesen la variable popularidad y encontrar patrones de co-ocurrencia con las categorías de las otras variables de los usuarios en el antecedente.

Esto fue particularmente intrincado, ya que o bien no se encontraban reglas o bien una de las categorías monopolizaba todas las reglas. Se tuvo que ir "tuneando" los diferentes parámetros, intentando conservar el soporte y la confianza más grande posible. También, en el mismo proceso iterativo, se probaron otras variables, se modificaron las categorías en cantidad y ancho, se probaron diferentes largos de regla hasta obtener algún resultado razonable. Nos quedamos con un *minsupp* de 0.004 y una *confianza* de 0.6.

Reglas	S	C	Lift	N
{afinidad=muy_alta, amigos=intermedia, seguidores=muy_alta} **=>** {popularidad=alta}	0.0049	0.678	1.85	95
{actividad=baja, afinidad=muy_alta, seguidores=muy_alta} => {popularidad=alta}	0.0040	0.663	1.81	77
{actividad=muy_alta, afinidad=muy_alta, seguidores=muy_alta} => {popularidad=alta}	0.0047	0.602	1.64	91
{afinidad=baja, seguidores=baja} => {popularidad=baja}	0.0058	0.63	2.32	111
{actividad=intermedia, seguidores=baja} => {popularidad=baja}	0.0045	0.63	2.32	87
{afinidad=intermedia,seguidores=baja} => {popularidad=baja}	0.009	0.62	2.30	172

Popularidad alta

Como se puede apreciar, teniendo en cuenta el tamaño del dataset y el bajo conteo de casos (N), las 3 reglas encontradas para alta popularidad son bastante "débiles", es decir, que la co-ocurrencia entre antecedente y consecuente es bastante poco frecuente. Esto podría deberse múltiples razones, desde algún error el preprocesamiento de las variables, como mala elección de variables e incorrecta discretización, pocos datos con mucho ruido, o bien que las variables no se encuentren asociadas, es decir, que los atributos elegidos del usuario no influyan en la popularidad de un tweet. Sin embargo la tendencia existe y los resultados encontrados "tienen sentido".

En las 3 la cantidad de seguidores es "muy alta", es decir, más de 1000 seguidores, lo cual tiene sentido que para que un tweet sea altamente retweeteado y tener muchos "likes" el usuario tiene que tener mucha gente que mire sus posteos.

Quizás lo más novedoso sea que en todas las reglas los usuarios tengan una afinidad alta, es decir, que se estaría sugiriendo que un usuario que acumula "likes" en sus posteos, es probable que los futuros también reciba "likes".

En cambio, actividad, solo aparece en 2 de las 3 reglas, y lo hace en una con la categoría más baja y en la otra con la más alta. Esto parecería indicar que el comportamiento del

usuario en cuanto a la frecuencia de posteo no estaría relacionado con la popularidad del tweet.

El lift es bajo, levemente superior a 1, con lo cual consecuente y antecedente podrían ser independientes y no estar relacionados, sin embargo, también hay que tener en cuenta que la variables popularidad está dividida en 3 categorías con un soporte cercano a 0.3 cada uno de ellas, con lo cual se esperan lift bajos si el antecedente es popularidad.

Popularidad baja

Al considerar las 3 reglas se nota un aumento tanto del soporte como del lift respecto a las de alta popularidad, lo que las hace un poco más robustas que las anteriores.

A la inversa que en el caso anterior ahora los usuarios de tweets poco populares tiene muy poco seguidores (< 50) lo cual tiene sentido. Se nota en cambio que la variable afinidad no se encuentra en todas las reglas y tampoco lo hace en la categoría más baja. Parecería que los usuarios tienen condiciones “menos restrictivas”, es decir, por más que tengan una afinidad intermedia ya con poseer muy pocos usuarios el tweet es muy difícil que sea popular.

Consigna 4.2

En esta segunda parte del punto cuatro, intentamos entender cómo es el contenido más difundido. Una de las reglas tiene que ver con el momento en que se publica, y las dos siguiente con el tipo de dispositivo desde el que hizo el tweet principalmente. El universo de fueron los retweets, al igual que en el punto uno y dos.

Para este último punto, adicionamos dos métricas extras, Imbalance Ratio (IR) y Kulczynski (Kulc). Consultando a la bibliografía, el libro “Data mining. Concepts and techniques”, cabecera de la materia, la intención era sumarlas para orientarnos en lo valioso de una regla. IR es *null invariant*, esto significa que no es sensible a las transacciones que no contienen los items que estamos investigando.

Es un buen complemento al Lift que tomamos como referencia, y lo apoya en un punto flojo. Lift está calculado sobre las probabilidades de ocurrencia, y nos ayuda a entender independencia de las partes. IR se calcula sobre soportes, similar a la confianza, y esto evita que la medida tome en cuenta las transacciones que no contienen los elementos que nos importan. Lo importante es hallar reglas con un ratio lo más cercano a 1 posible.

Kulczynski es una medida similar a la confianza, pero en ambas direcciones. Permite analizar P(A|B) y P(B|A), y evitar elegir reglas donde B es significativo para A, pero que al invertir la relación, vemos que A no es relevante dado B. Vamos a desarrollar este punto con la primera regla de ejemplo.

Regla	S	C	Lift	N	IR	Kulc

{Creacion_retweet=Noche, Es_final_semana=Fin de semana} => {Difusion=alta}	0.08	0.31	1.17	7,463	0.02	0.30
{Source=Twitter for iPhone, Verificado=Verificado} => {difusion=alta}	0.02	0.60	2.30	1,975	0.83	0.34
{Es_final_semana=Fin de semana,Source=Twitter for iPhone, Popular=muy muchos} => {Difusion=alta}	0.02	0.71	2.72	1,773	0.87	0.39

Las tres reglas seleccionadas hablan de aspectos que hacen a la difusión de un tweet. La primera sigue una línea que se había planteado en el primer punto, como hay mayor probabilidad de que un contenido tenga un mayor alcance los finales de semana y durante la noche. El soporte es alto en comparación a muchas reglas, aunque la confianza no. Pero lo interesante surge cuando miramos las nuevas variables. El IR tan cercano a cero habla de una igualdad entre P(A|B) Y P(B|A). Esto es interesante para entender mejor la relación de las métricas y habla de una mayor fuerza en la relación. Se pueda ver lo mismo con la métrica Kulczynski, que es un promedio de ambas probabilidades.

En la segunda y la tercera regla, Kulczynski es menor a la confianza. Viendo esto, y el imbalance ratio (IR). Dando este antecedente es muy probable que haya una alta difusión, pero no el caso inverso.

Esto es interesante, porque encuentra un conjunto de características que dan como consecuencia lo que estamos estudiando. Si una persona con iPhone y verificada, es retweeteada hay altas chances de que se difunda mucho el tweet. La proporción de reglas en otro tipo de dispositivos es menor. En Android que quedó dentro de los soportes seleccionados, fue probablemente por una mayor cantidad de dispositivos de esta clase.

El soporte de la regla “Source=iPhone => difusion=Alta” era bastante alto, alrededor del 7%. Estas dos reglas más específicas, continúan teniendo un soporte, una confianza y un lift alto. Lo cual habla de una buena frecuencia y una causalidad.

Conclusión

Durante la elaboración del presente trabajo tuvimos la oportunidad de repasar y utilizar varias técnicas del preprocesamiento de datos empleadas en el primer trabajo práctico. Principalmente debido a que el algoritmo utilizado (*arules*) requiere convertir todas las variables numéricas de interés a categóricas. Esta etapa nos requirió una inversión de tiempo considerable, entendimos desde un principio la importancia de la misma, donde una mala elección de las variables y de sus categorías (ancho y frecuencia de las bandas) repercute dramáticamente en las reglas obtenidas. Existe un delicado equilibrio entre este proceso y obtener reglas robustas y confiables.

A su vez la generación y evaluación de reglas fue un desafío en sí mismo, un proceso iterativo en el que tuvimos que realizar numerosas pruebas, combinando diferentes variables de interés con valores de confianza y soporte que produjeran resultados interesantes y robustos a la vez. Por momentos fue frustrante, resultandonos muchas veces dificultoso la interpretación de las reglas, sin embargo, pudimos aplicar las técnicas aprendidas para

generar reglas de asociación, modificando los parámetros de las reglas (soporte, confianza lift, IR y Kulc) y evaluando su robustez para los distintos casos analizados.

Nos queda la sensación de haber avanzado mucho. De haber podido complejizar aún más el preprocesamiento, encontrando optimizaciones e iterando de una forma más rápida sobre esa parte que ya conocíamos y también extrapolar esas habilidades a lo nuevo. Nos queda la sensación y el entusiasmo por seguir una metodología exploratoria, que tal vez con más tiempo y más datos podríamos mejorar esas tendencias, complejizarse aún más y complementarlas con otras técnicas.

Cerramos intentando contar esa sensación que atravesó ambos trabajos, que a cada paso el camino se bifurca infinitas veces. Cada puerta que abrimos nos muestra un mundo por conocer, y hace sentir más pequeño aún ese paso que terminamos de dar. En ese sentido nos inquieta no haber podido avanzar tanto como hubiésemos querido, y no haber podido tocar más temas y puntos, o hallar conclusiones más disruptivas. Pero eso también nos conecta porque estamos acá, en una de las áreas más desafiantes de la actualidad y nos deja tranquilos haber podido dar algunos pasos firmes en ese camino.