CURSO: MINERÍA DE DATOS

MAESTRÍA EN EXPLOTACIÓN DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO

TRABAJO PRÁCTICO ENTREGABLE 2

Reglas de Asociación

INTRODUCCIÓN

En este Trabajo Práctico se incorpora la técnica de reglas de asociación para abordar un problema puntual y tratar de entenderlo a partir la observación de co-ocurrencia de factores.

Para la exploración de este tema, se utilizará el IDE R-Studio del lenguaje de programación R con el objetivo de ejercitar los conceptos abordados en las clases teóricas.

El dominio del problema a analizar es el análisis de patrones frecuentes en un conjunto de tweets recolectados en los últimos meses de confinamiento social y preventivo ante la pandemia de COVID19[1]. Los documentos son publicaciones en países hispanohablantes con similares características a los datos provistos para el TP1 pero con mayor volúmen de datos.

OBJETIVO GENERAL

El objetivo general de este trabajo es aplicar la técnica de reglas de asociación sobre 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.

FECHA DE ENTREGA

Lunes 13 de Julio de 2020

DATOS

https://drive.google.com/drive/folders/1_YsBw4fnbfXUM7IhexCk3iu-jwDIG0_G?usp=sharing

PREPROCESAMIENTO

Antes de aplicar reglas de asociación sobre el conjunto de datos se plantean las siguientes pautas de preprocesamiento y transformación de variables. Considere estas pautas como una línea de base e incorpore otras transformaciones en el caso de ser necesario.

[1] Los términos de búsqueda utilizados para recuperar los tweets fueron: "COVID19", "COVID 19", "CORONA VIRUS", "CORONAVIRUS" y "cuarentena"

Tratamiento de Hashtags

Considere generar variables a partir de los hashtags para convertirlos en ítems. Con esta transformación será posible combinar temáticas que aparezcan en las publicaciones tanto con el conjunto de datos de los usuarios como con los tweets.

Discretizaciones

Será necesario generar variables con rangos discretos donde cada etiqueta será un ítem. Un ejemplo de este tipo de transformaciones puede ser ***friends_count***. Tenga en cuenta para las discretizaciones cómo están distribuidos los datos.

Pocos	Medio	Muchos
friends_count < 100	100 ≤ friends_count < 10000	friends_count ≥ 10000

Datos de los usuarios y tweets

Una estrategia de trabajo puede ser considerar los usuarios como “carrito” y como ítems es posible abordar muchas alternativas, a continuación se describen algunas estrategias:

1. Si un usuario es un carrito, los ítems podrían ser los usuarios originales de los retweets (por ejemplo el screen_name). Entonces acá los itemsets frecuentes serían los usuarios más retuiteados. Esta estrategia permitiría analizar cómo se relacionan las cuentas más populares.
2. Considerando de la misma manera a los usuarios como carritos, y como ítems el vocabulario de los términos más frecuentes y/o de los hashtags utilizados en sus tweets. Buscando de esta manera, encontrar asociaciones de temáticas en los tweets.
3. Los ejemplos anteriores podrían ser enriquecidos con otros atributos de los usuarios o de los tweets.
4. Otra estrategia sería usar como carrito a los tweets, y las características de los mismos como ítems.

Tratamiento de textos de los tweets

Desde los textos de los tweets se deberán extraer palabras que permitan caracterizar cada documento y puedan ser utilizadas como ítems. A continuación se proponen una serie de pautas para realizar esta extracción de características.

Extracción de términos desde los tweets

Palabras como columnas:

1. Generar una copia del atributo **text** para realizar este proceso.
2. Convertir las columna a minúsculas[2].
3. Quitar dígitos numéricos[3]
4. Quitar símbolos de puntuación[4].
5. Quitar tildes a las vocales[5].
6. Obtener el listado términos sin repetidos[6].
7. Eliminar palabras vacías en español[7] (preposiciones, artículos, etc).
8. Finalmente luego de aplicar todas estas transformaciones sobre el campo de text, separar en palabras y realizar conteos. Formar un vocabulario de palabras con aquellas palabras que tengan una mínima frecuencia[8].
9. Por cada palabra de vocabulario seleccionada, generar una columna de presencia ausencia. En el caso de ausencia marcar como NA, y en el caso de presencia algún caracter (Ej 'S'). Es recomendable para luego aplicar reglas de asociación, utilizar un prefijo en cada una de estas columnas seguido por la palabra en cuestión[9] (Ej: *termino_casos*, *termino_pandemia*).

Palabras como filas:

Otra alternativa que permitiría abordar más términos que en el caso anterior, es mantener únicamente un atributo que sirva como identificador de transacción, y el atributo texto. Luego se realiza la misma limpieza del texto que en caso anterior (remover acentos, pasar a minúsculas, remover stop words, etc) y se separa en palabras (tokenizar). Hasta acá

[2] Investigue la función *tolower.*
[3] Sugerencia: *tm::removeNumbers.*
[4] Sugerencia: *tm::removePunctuation*.
[5] Sugerencia: *stringi::stri_trans_general(texto,"Latin-ASCII")*.
[6] Sugerencia: *unique() tm::removeWords.*
[7] Sugerencia: *tm::stopwords*.
[8] Sugerencia: *findFreqTerms* o *apriori*.
[9] Sugerencia: *stringr::str_detect*.

tendríamos un dataframe con dos columnas: el identificador de transacción, y el listado de términos en formato de vector.
Finalmente explotamos la columna de términos en filas, es decir por cada palabra se generará una nueva fila, repitiendo el identificador de transacción asociado. Formando un dataframe con pares de transacción-palabra, siendo adecuado para trabajar con el paquete arules.

Otras fuentes de datos y transformaciones

Considere las transformaciones y fuentes adicionales de datos para generar nuevas características que sean de utilidad para la extracción de reglas de asociación.

REGLAS DE ASOCIACIÓN

Una vez que cuenta con el dataset preprocesado, avance sobre la obtención de reglas que le permitan concluir respecto de las siguientes consignas:

Análisis descriptivo: Escoja 6 reglas (para cada consigna) que le permita describir el conjunto de datos.

1. Explique qué aportes se obtienen a partir de las mismas y explique cuál es el peso de cada una de acuerdo a las métricas asociadas.
2. Genere reglas de decisión que expliquen diferentes aspecto de la popularidad de los tweets. Explique cuáles son las reglas más robustas, justifique y explique el conocimiento que estas aportan.
3. Seleccione un grupo de usuarios reducido explique qué factores están asociados a su comportamiento (popularidad, influencia de sus publicaciones, etc) obtiene a partir de reglas de asociación.
4. Formular dos preguntas de KDD y realice una comprobación mediante reglas de asociación. Puede utilizar algunas de las preguntas formuladas en el TP1 y tratar de realizar una comprobación mediante el uso de reglas.

Documente las iteraciones realizadas entre preprocesamiento y la construcción de las reglas y haga referencia a cuáles son las transformaciones con las que obtuvo una mejor configuración.

Las reglas deben ser presentadas en tablas, en formato de regla y con las métricas correspondientes. Todas las tablas con sus respectivas reglas deben estar analizadas e interpretadas, dejando claro el aporte que realizan.

Ejemplo de tabla:

Regla	S	C	Lift	N
{item1, item2..} => {item5}	...	...	...	...