TRABAJO PRÁCTICO ENTREGABLE 2
Reglas de Asociación

Grupo 4 (martes): Lopez Malizia, Álvaro; Padula, Eliana; Rossi, Fabiana A.

RESUMEN

En el presente trabajo nos propusimos integrar los conocimientos relacionados con el preprocesamiento y gestión de datos desarrollados durante el trabajo práctico 1, con la técnica de reglas de asociación para descubrir la co-ocurrencia de factores interesantes. Para ello utilizamos una Base de Datos NO-SQL (MongoDB) y el IDE R-Studio del lenguaje de programación R. Se realizó un análisis de información relativa a *rankings* semanales de Spotify de los 200 temas más escuchados a nivel mundial. Se complementó esta información con la letra de las canciones interpretadas en idioma inglés e índices de positividad de palabras, y se crearon reglas de asociación en base a la ocurrencia de características y/o términos en dichas canciones.

INTRODUCCIÓN

El volumen de datos que se generan continuamente ha ido creciendo de modo exponencial durante los últimos años, por lo que existe una necesidad urgente de utilizar herramientas computacionales que asistan en la extracción de nueva información, en el proceso de descubrimiento de la información en datasets (KDD). El objeto de KDD es la revelación e interpretación de nuevos patrones en la información, que son imperceptibles en un primer momento debido a su gran volumen. En este trabajo práctico complementamos el análisis realizado sobre los datos explorados durante el primer trabajo práctico, generando reglas de asociación determinados por la co-ocurrencia de características/hechos en los datos.

METODOLOGÍA

Se complementaron los datos obtenidos en el trabajo práctico 1 con la letra de las canciones obtenidas a partir de la colección Genius y el léxico de Sentiwordnet 3.0. La base de datos generada fue importada al procesador de bases de datos MongoDB, y se seleccionaron aquellas canciones en idioma inglés. Dicha base de datos cuenta con información del intérprete, título y letra de cada canción. En primer lugar, se eliminaron los duplicados. Posteriormente, y durante el procesamiento del *corpus* de las letras y creación de la matriz término-documento, se utilizó la lista de palabras vacías (*stopwords*) de la librería *tm*, y se utilizó el parámetro *weightTfIdf*, que realiza una ponderación determinada por la frecuencia de cada término.

Por otro lado, se realizó el procesamiento de la base de datos de atributos de audio de las canciones. Luego de verificar la existencia de duplicados en las colecciones de Spotify (*charts* y *audiofeatures*), se eliminaron dichos registros. Se integraron los datos en una única base de datos, y se promediaron los datos de las variables numéricas correspondientes a las características de audio que presentaban el mismo autor y nombre de tema. Se decidió descartar aquellas variables no pertinentes a nuestro análisis (direcciones de url o imágenes). También se descartaron aquellas variables con poca variabilidad, y algunas fuertemente correlacionadas. Luego de revisar la integridad de los datos, se eliminaron aquellos registros que presentaban un porcentaje de datos faltantes mayor al 50%. En aquellos casos donde existía más de una lista de países de distribución para un mismo tema e intérprete, se eligió aquella de mayor longitud. Se transformó el tipo de dato a aquél más apropiado para cada atributo ('Date', 'float', etc), y se realizaron transformaciones que redujeron el sesgo de algunas variables. Posteriormente se discretizaron las variables utilizando la técnica de k-vecinos más cercanos.

Finalmente, se crearon reglas de asociación integrando información de variables de los audios y de las letras de las canciones en inglés del *chart*, utilizando el paquete *arules*. Asimismo, se crearon reglas derivadas de información acerca de los mercados de distribución de las canciones, los meses en donde fueron escuchadas y los términos presentes en sus letras. En los meses del año sólo fueron consideradas las palabras del léxico de Sentiwordnet 3.0. El mismo fue diseñado explícitamente para respaldar la clasificación de sentimientos y la opinión. Contiene palabras a las que se les asignan puntuaciones numéricas según cuán positivos o negativos resultan según su significado e interpretación. Las reglas propuestas complementan las hipótesis generadas en el trabajo práctico número uno.

Detalle de los atributos utilizados

El *dataframe* ***charts*** presentaba los siguientes atributos: **position** (posición en el *chart*), **track_name** (nombre del tema), **artist_name** (nombre del artista), **streams** (número de reproducciones), **week_start/week_end** (comienzo/fin de la semana para un dado *chart*).

Del *datafarme* ***audiofeatures***, se utilizaron los siguientes atributos: **artist_name** (nombre del artista intérprete), **track_name** (nombre del tema), **danceability** (descripción de cuán apropiada es un tema para bailar: valores entre 0-1), **energy** (representa una percepción de intensidad y actividad: valores entre 0 y 1), **valence** (medida de positividad de un tema: toma valores entre 0 y 1) y **market** (código de país ISO 3166-1 alpha-2). También se utilizaron los atributos **acousticness** (acústica) y **liveness** (detecta la presencia de una audiencia), para los cuales se realizó una transformación de logaritmo natural, lo cual redujo su sesgo.

Además, se crearon los siguientes atributos: **permanencia ininterrumpida** (número de semanas que cada tema permaneció en el *chart* de manera ininterrumpida, para cada una de las instancias que ingresaron al *chart*), **popularidad** (cantidad total de apariciones de un tema en el *chart*) e **índice de éxito del tema** (utilizando datos escalados entre 1 y 100).

$$índice\ de\ éxito\ =\ ln\ (\ \frac{0{,}25\ MPI \times 0{,}5\ PM \times 0{,}25\ ST}{2\ PMA}\ )$$

donde los términos son:

MPI= máxima permanencia ininterrumpida (semanas), **PM**= posición media, **ST**= *streams* totales, **PMA**= posición mínima alcanzada.

Adicionalmente, se utilizaron los **índices de Positividad** y **Negatividad** provistos por el léxico de Sentiwordnet 3.0.

RESULTADOS

Las hipótesis planteadas en este trabajo se formularon sobre la base del trabajo práctico 1, con el objetivo de ampliar e integrar los resultados obtenidos en el mismo.

Hipótesis 1 : Existen atributos y términos que co-ocurren con el éxito alto de un tema

Con el objeto de determinar si existían atributos y/o términos que ocurrieran de manera simultánea con las categorías de la variable **índice de éxito**, en primer lugar se creó una matriz término documento con la información de todas las canciones interpretadas en inglés, y evaluó la frecuencia de cada término en el modelo de bolsa de palabras. Tal como puede observarse en la **Figura 1** (izquierda), algunos de los términos más frecuentes de las canciones eran: "*just*", "*like*", "*yeah*", "*love*", "*never*", "*see*", "*can*", "*let*", "*put*", "*give*", y seguían una distribución de Ley de Zipf (derecha). Dado que la elevada frecuencia de algunos términos podría conducir a la obtención de reglas de asociación superfluas, decidimos remover el 1% de los términos más frecuentes.

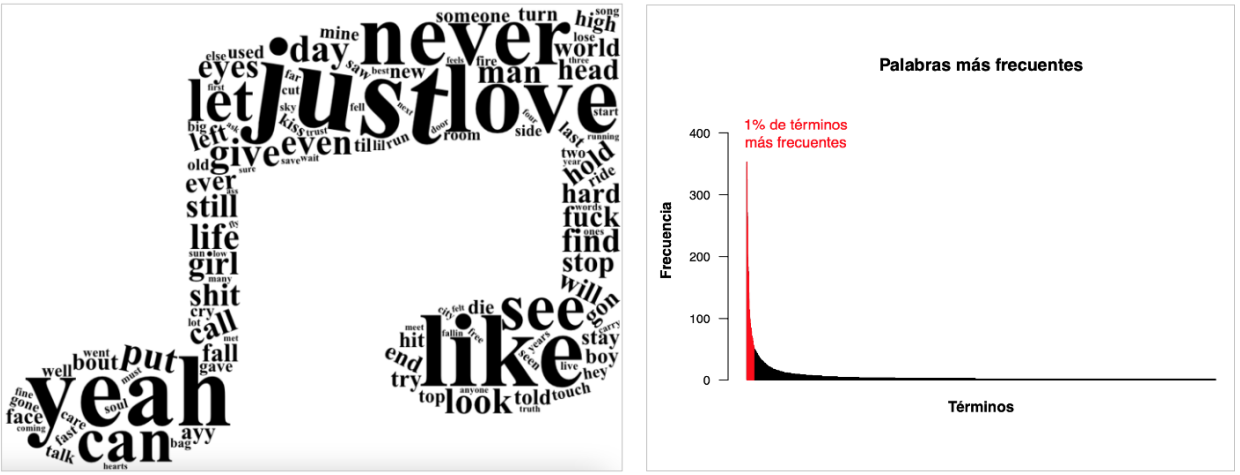

Figura 1- *Representación de los términos asociados a las canciones interpretadas en idioma inglés. A la izquierda se representa una nube de palabras de las letras de las canciones, en donde el tamaño de los términos se correlaciona positivamente con su frecuencia. Derecha: distribución de frecuencias de los distintos términos de la bolsa de palabras. En rojo se indica el 1% de términos más frecuentes.*

A continuación, se generó una matriz término-documento del *corpus* de canciones y atributos de audio discretizados. Luego se crearon reglas de asociación utilizando los parámetros **soporte**= 0,1; **confianza**= 0,2; **longitud mínima**= 2, **longitud máxima**=5 y se obtuvo un total de 1.107 reglas (**Figura 2**).

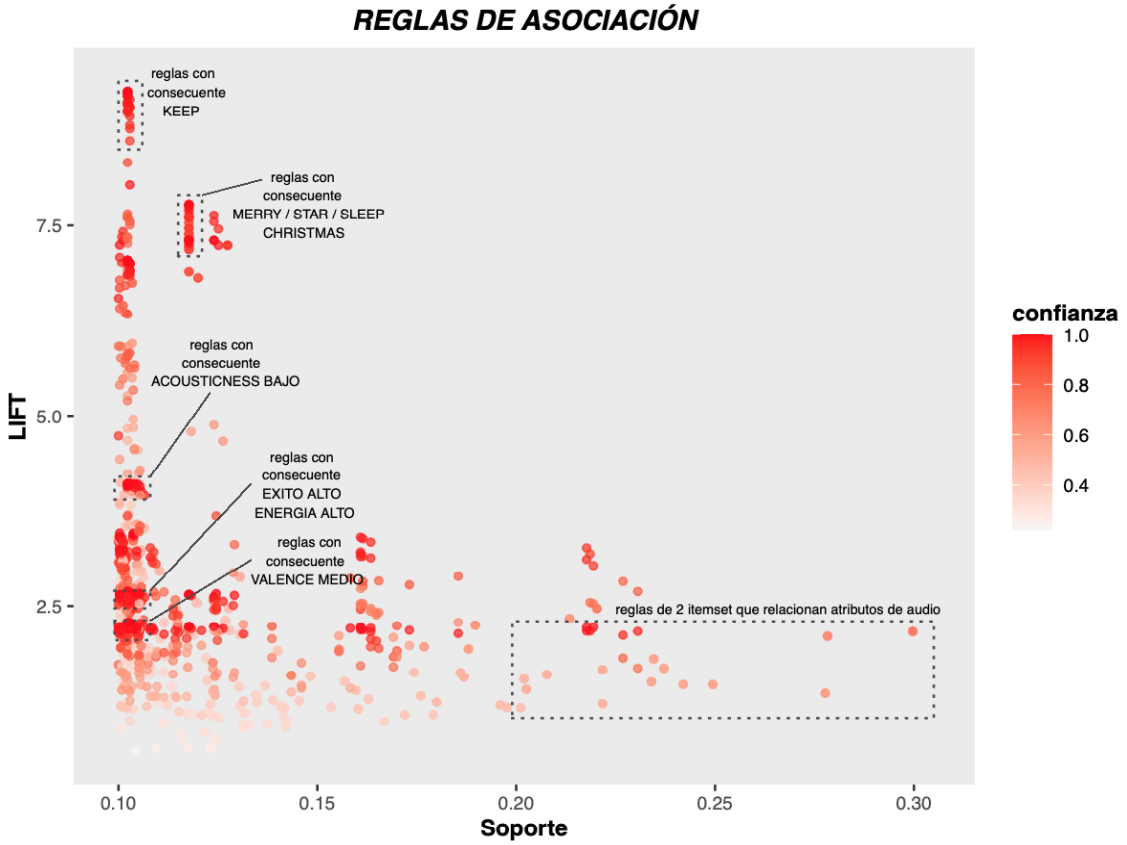

Figura 2- *Representación de reglas de asociación de la matriz término documento de las letras de las canciones y atributos de las mismas. Cada punto representa una regla de asociación, con su soporte y lift asociados. Los mismos están coloreados con una escala de colores blanco-rojo, que representan los valores menores-mayores de confianza (respectivamente).*

A partir de los resultados de la **Figura 2**, se exploraron las reglas formuladas. En términos generales, reglas de alto *lift* mostraron reglas previsibles y de bajo interés, como ser reglas que tenían al término "*keep*" como consecuente. Asimismo, se observaron reglas de alto *lift* en las cuales "*merry*", "*christmas*", "*star*" eran parte del consecuente, y cuya constitución indicaba reglas que posiblemente estén relacionadas con canciones festivas. Por otro lado, se observó que las reglas de menor confianza en general presentan menor *lift*, y que aquellas de mayor soporte estaban conformadas por 2-*itemsets* que relacionaban atributos de audio.

A fin de enriquecer el espacio de reglas en aquellas relacionadas con una categoría **medio** y **alto** para la variable **índice de éxito,** se tomó un *subset* de las reglas generadas y se definió **índice de éxito** como parte del consecuente.

Tal como puede observarse en la **Figura 3**, se obtuvieron 152 reglas de asociación a partir de este *subset*. Las mismas poseían una **confianza** mínima de 0,27; un **soporte** mínimo de 0,1 y un **lift** mínimo de 0,75. Entre las mismas, se encontraban reglas con la categoría bajo para el índice de éxito (~3%), medio (~47%) y alto (~50%).

A continuación, se realizó un análisis de aquellas reglas del *subset* con valores de *lift* mayores a 1,25, y valores de soporte mayores a 0,11.

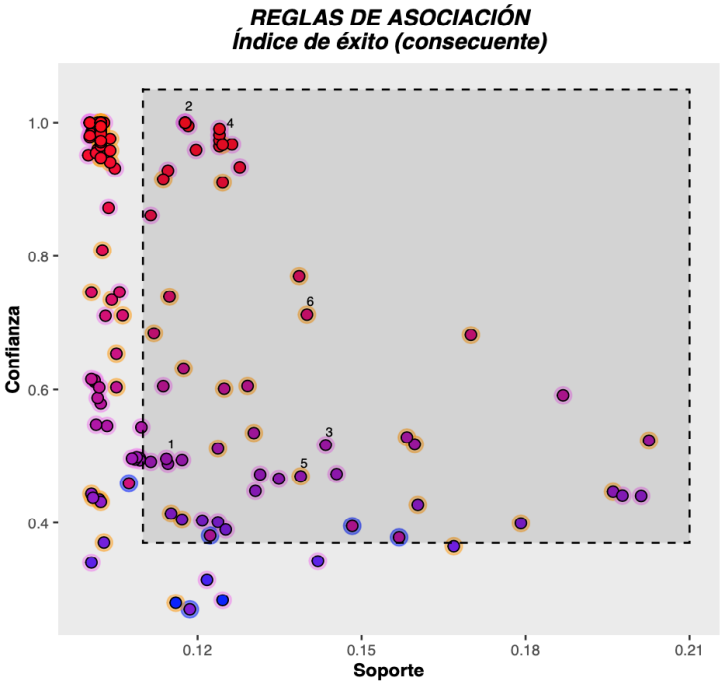

Figura 3- *Representación de reglas de asociación con consecuente equivalente a* ***índice de éxito****. Cada punto representa una regla de asociación, con su soporte y confianza asociados. Los mismos están coloreados con una escala de colores azul-rojo, que representan los valores menores-mayores de lift (respectivamente). El borde de los puntos indica la categoría del consecuente de cada regla (bajo, medio o alto). La zona sombreada indica las reglas con lift mayor a 1,25 y soporte mayor a 0,11. Los números indican la referencia de las reglas en la* ***tabla 1****.*

Entre las distintas reglas de asociación obtenidas se seleccionaron siete, las cuales se formulan en la **Tabla 1.**

Nº	ANTECEDENTE	CONSECUENTE	SOPORTE	CONFIANZA	LIFT
1	acousticness_alto,danceability_bajo	índice_medio	0,114	0,488	1,294
2	término_*christmas*,término_*merry*,término_*star*	índice_medio	0,117	1,000	2,649
3	acousticness_alto,valence_bajo	índice_medio	0,143	0,516	1,368
4	acousticness_bajo,energy_alto,liveness_medio	índice_alto	0,124	0,966	2,598
5	danceability_alto	índice_alto	0,138	0,469	1,262
6	término_*dancing*	índice_alto	0,140	0,712	1,914

Tabla 1- *Detalle de un subset de reglas de asociación con consecuente equivalente a* ***índice de éxito.***

Las reglas esbozadas en la **Tabla 1** son robustas con un soporte mínimo de 0,11; una confianza mínima de 0,48 y un *lift* mínimo de 1,42. Las reglas 1 a 3 tienen como consecuente un índice de éxito en su categoría medio, y las reglas 4 a 6 un índice de éxito con nivel alto.

La regla 1 muestra la co-ocurrencia de temas con éxito medio y que además tienen un valor de *acousticness* alto y un valor de *danceability* bajo. La misma informa acerca de los temas que posiblemente son más "tranquilos" y "poco bailables", que no son tan exitosos. Por otro lado, la regla 2 asocia términos relacionados con navidad, tales como *"christmas"*, *"merry"*, *"star"*, e índices con nivel medio. La conclusión que deriva de esta regla es que los temas navideños son medianamente exitosos. Por otro lado, la regla 3 indica que bajos niveles de *acousticness* se asocian con temas de baja valencia, los cuales además tienen un éxito intermedio.

Por otro lado, las reglas con consecuente índice de éxito igual a "alto" muestran la ocurrencia simultánea de temas con *acousticness* bajo, *energy* alto, *liveness* medio y un alto éxito. También aparecen reglas que incluyen al término *"dancing"* y un nivel de *danceability* alto. Estos resultados parecen indicar que los temas más exitosos son además "bailables" y tienen un nivel de energía alto, *liveness* moderado y *acousticness* bajo.

Hipótesis 2 : Existen términos que co-ocurren con los mercados de distribución de los temas.

Con el objetivo de estudiar los términos de las canciones presentes en los distintos países, se realizó una matriz término documento que incluyó información sobre los **mercados de distribución** de los temas. Debido a la gran cantidad de países disponibles se decidió agruparlos por continentes. La distribución resultante en cada una de las categorías quedó determinado por su pertenencia (o ausencia) y fue balanceada: América del Norte (715), América del Sur (711), América Central (712), Asia (713), Oceanía (720) y África (718).

Como primera aproximación, se construyó el conjunto de reglas de asociación para el total de los términos presentes. Los parámetros de asociación utilizados fueron: **soporte**= 0,09; **confianza**= 0,5; **longitud mínima**= 2, obteniéndose un total de 75.161 reglas (**Figura 4**).

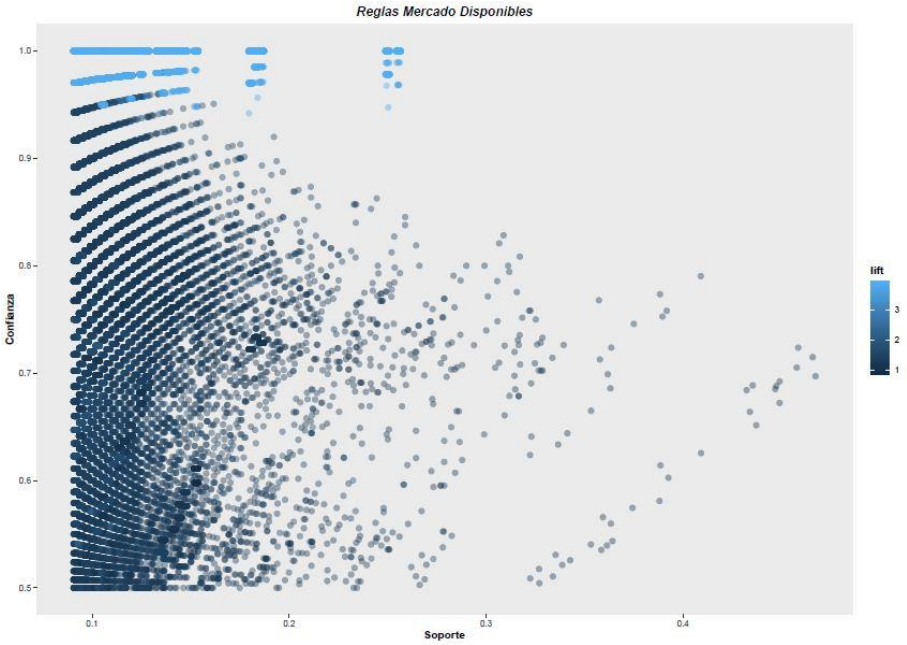

Figura 4- *Representación de reglas de asociación en relación a los* ***mercados disponibles****. Cada punto representa una regla de asociación, con su soporte y confianza asociados.*

Las reglas de asociación de los sectores geográficos se caracterizaron por un alto lift. Los puntos coloreados en azul claro, en la parte media superior, pertenecen a las asociaciones obvias entre los continentes (**Figura 4**).

A continuación se buscaron reglas de orden 2 (2 *itemset*) que permitieran conectar regiones en particular, ya sea mediante la condición de consecuente o de antecedente. Ordenando de forma decreciente de acuerdo al valor de lift, se encontraron reglas de asociación directas para la condición de consecuente. Sorprendentemente, las regiones de Asia y América del Norte no evidenciaron reglas que se ajustaran a dicho criterio. Dichas reglas se representaron en un gráfico de cuerdas (**Figura 5**), en donde la conexión entre regiones geográficas representa los ítem dentro de las reglas encontradas.

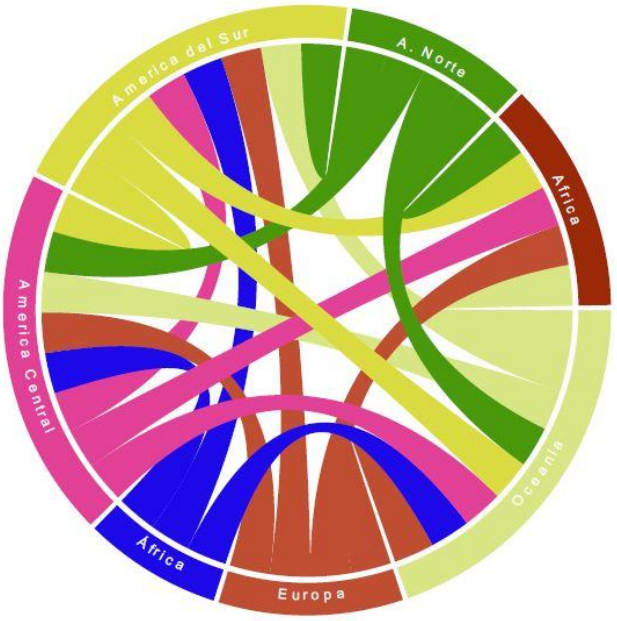

Figura 5- *Representación de reglas de asociación entre los distintos continentes. Cada color representa una región. Las reglas de asociación diferenciales de longitud 2 sólo se encontraron cuando el continente se encontraba como consecuente.*

En la **Figura 5** se puede observar que África es consecuente para los temas que se escuchan en América del sur, América Central y Oceanía. América del Norte, por el contrario, no es consecuente de ninguna otra región.

La presencia de estas reglas sugiere que, en menor frecuencia, los temas que se escuchan en América del norte y Asia no provienen de otras regiones continentales. Una hipótesis es que debido a su peso demográfico y penetración de Spotify, poseen temas reproducidos únicamente

en sus límites que ingresan al *chart* global. Contrariamente, no se encontraron diferencias entre las regiones para las reglas filtradas mediante la condición del antecedente.

Finalmente, se realizó un nuevo preprocesamiento del corpus para optimizar la generación de reglas en el contexto de la información asociada a continentes. Como se mencionó anteriormente, los términos presentaron una distribución de acuerdo a Ley de Zipf, por lo que se decidió eliminar el 1% de los términos más frecuentes.

Se generaron nuevas reglas con un valor de **soporte** de **0,001** (para valores superiores no fue posible generar reglas de orden 2 entre continentes y términos). A continuación se realizó un *subset* de las mismas, filtrando los continentes como consecuentes, y se obtuvieron un promedio de 200 reglas por grupo geográfico. Se encontraron las siguientes reglas para América del Norte y Europa (**Tabla 2**).

	ANTECEDENTE	CONSECUENTE	SOPORTE	CONFIANZA	LIFT
América del Norte	término *anywhere*	América del Norte	0,011	0,500	1,894
	término *barely*	América del Norte	0,011	0,500	1,894
	término *losin*	América del Norte	0,013	0,500	1,894
	término *stand*	América del Norte	0,019	0,538	2,040
	término *blow*	América del Norte	0,027	0,500	1,894
Europa	término *easy*	Europa	0,027	0,500	1,935
	término *anyone*	Europa	0,016	0,545	2,111
	término *full*	Europa	0,022	0,533	2,064
	término *half*	Europa	0,025	0,529	2,049

Tabla 2- *Detalle del conjunto de subreglas únicas de asociación Europa y América del Norte.*

De acuerdo a las reglas observadas, existe un conjunto de términos que se asocia con las regiones de América del Norte y Europa. Dichos términos no se encuentran en asociación directa con el resto de los continentes.

Hipótesis 3: Existe una asociación entre las palabras que utilizan las canciones según el mes del año, y las mismas presentan un grado de positividad/negatividad diferencial según el mes.

Con el objetivo de determinar la estacionalidad en las palabras de las canciones del *ranking* mundial de Spotify, se realizó un análisis de términos agrupando las canciones por mes. Se limitó el estudio a los términos que estuviesen contenidos en el léxico Sentiwordnet 3.0, el cuál permitió la identificación de la emocionalidad negativa o positiva de las palabras. De aquellos términos encontrados frecuentemente en las canciones, se extrajeron 3 cuya frecuencia era relativamente alta comparada con los demás (**Figura 6**).

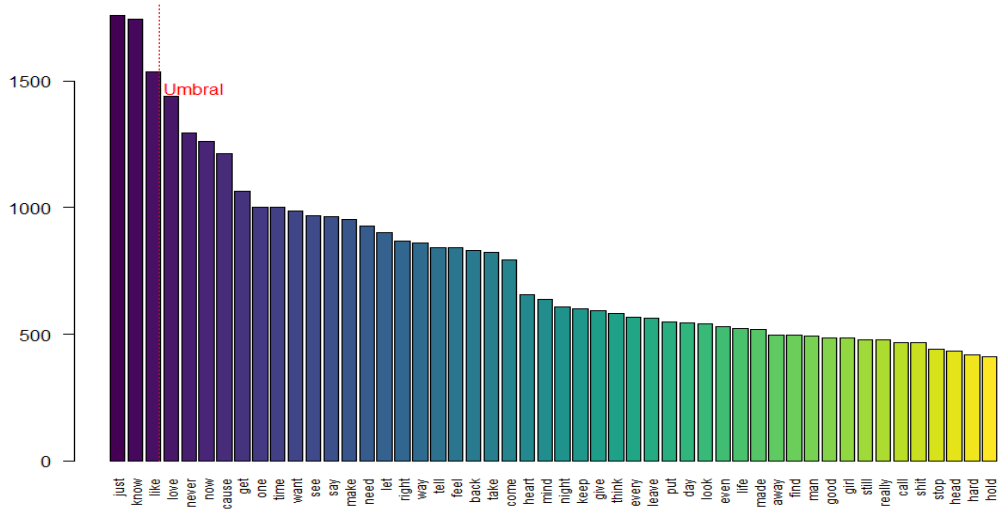

Figura 6- *50 palabras más frecuentes utilizadas en el ranking de Spotify que se pueden hallar en Sentiwordnet 3.0. Se observa el valor umbral utilizado. Los términos "just, know y like" quedan fuera de dicho análisis.*

Resumen de las medidas de calidad			
	Soporte	**Confianza**	**LIFT**
Mínimo	0,05047	0,3070	1
Primer cuartil	0,05748	0,3505	1,043
Mediana	0,06355	0,3750	1,098
Media	0,06745	0,3845	1,118
Tercer cuartil	0,07290	0,4139	1,157
Máximo	0,34019	0,5273	1,408

Una vez realizado el preprocesamiento de datos, se crearon reglas de palabras de canciones cuyo consecuente implicase un mes en particular. Se tomó un nivel de soporte mínimo de 0,05 para poder incrementar el número de candidatos y obtener información de todos los meses. Fue posible encontrar reglas con *lift* mayor 1, es decir que tengan alguna relación y con una confianza promedio de 0,38 (**Tabla 3**).

Tabla 3- *Resumen de las medidas de calidad de las reglas de asociación.*

Las reglas generadas encontraron sólo 1 palabra asociada a cada mes y se le asignó el índice de positividad y negatividad promedio que tienen según Sentiwordnet 3.0 (**Figura 7**).

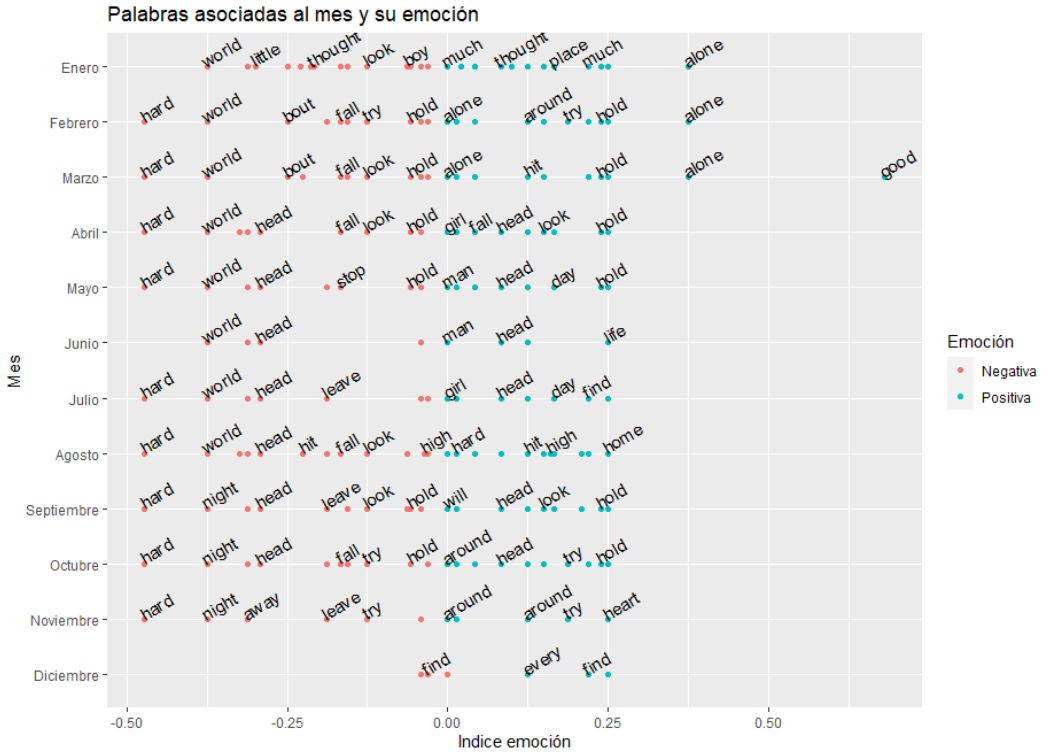

Figura 7- *Índice de positividad y negatividad según el léxico de Sentiwordnet para cada palabra hallada en las reglas asociadas al mes de reproducción.*

En términos generales, se puede observar que se hallaron las mismas palabras frecuentes en cada mes. No parecerían existir diferencias en cuanto a la frecuencia de nivel de positividad y negatividad de las palabras, excepto para el mes de diciembre. Particularmente, en este mes no se encontraron reglas de asociación con palabras de alto impacto negativo de emoción. Por otra parte, nuestros resultados indicaron que el mes de marzo contiene la palabra frecuente "*good*", catalogada con la máxima positividad.

Finalmente, con el objeto de comparar la similitud entre las reglas encontradas en los distintos meses, se calculó el coeficiente de similitud de Jaccard (**Figura 8**)**.** Se observó un valor de similitud cercano a 0,5 entre los meses analizados, a excepción de diciembre. Para dicho mes, la similitud de Jaccard alcanzó un valor máximo de apenas 0,23 al compararlo con julio.

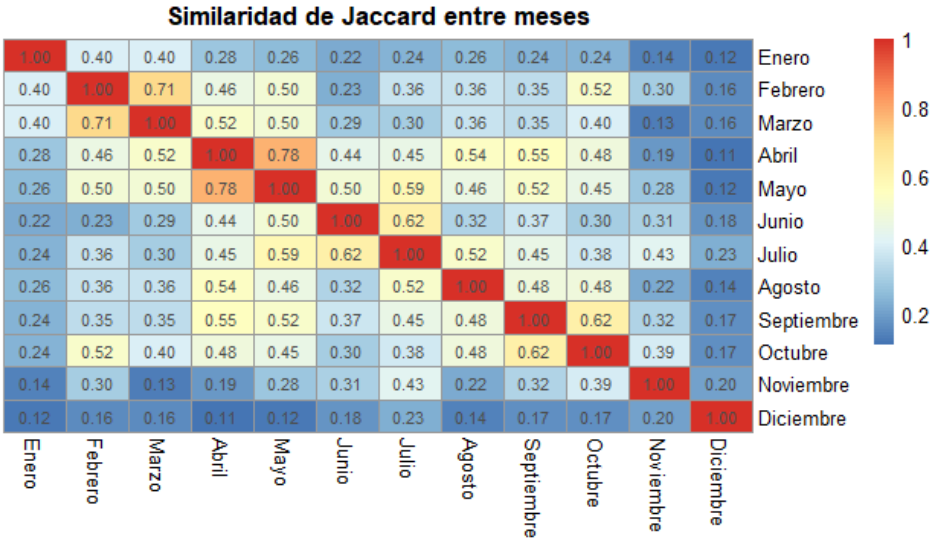

Figura 8- *Similitud de Jaccard calculada como: sim(A, B) = supp(A & B) / (supp(A) + supp(B) - supp(A & B)) donde A y B representan el conjunto de palabras encontradas como regla para el mes A y B, respectivamente.*

DISCUSIÓN

En el presente trabajo se realizó un análisis exploratorio de asociación de palabras relativa a las letras en inglés de *rankings* semanales de Spotify de los 200 temas más escuchados a nivel mundial. Las hipótesis planteadas en este trabajo se formularon sobre la base del trabajo práctico 1, con el objetivo de ampliar e integrar los resultados obtenidos en el mismo.

La primera hipótesis de este trabajo estipulaba que existen **atributos o términos que co-ocurren con el éxito alto de un tema.** Para evaluar esta hipótesis, se generaron reglas de asociación a partir de una matriz término-documento que integraba información relativa a las características de audio de las canciones y los términos en sus letras (descontando el 1% más frecuente). Dicho análisis permitió arribar a las siguientes conclusiones: los temas de éxito medio tienen un valor de *acousticness* alto y un valor de *danceability* bajo. Dichos temas son posiblemente más "tranquilos" y "poco bailables", y no tan exitosos. Existen términos asociados a canciones festivas, las cuales son medianamente exitosas. Los temas de baja valencia y *acousticness* medio también tienen éxito intermedio. Por el contrario, canciones de alto éxito se asocian con niveles de *acousticness* bajo, *energy* alto y *liveness* medio, así también como con el término *"dancing"* y un nivel de *danceability* alto. Estos resultados parecen indicar que los temas más exitosos son además "bailables" y tienen un nivel de energía alto, *liveness* moderado y *acousticness* bajo.

A continuación, evaluamos la hipótesis de que **existen términos que co-ocurren con los mercados de distribución** de los temas. La incorporación de los atributos de regionalidad al análisis permitió estudiar, de forma general, la tendencia sobre la distribución de los temas en los continentes. Si bien el objetivo inicial de esta hipótesis fue testear la existencia de palabras que se asociaran con regiones específicas, observamos que debido a la homogeneidad de la distribución de los temas, las reglas que se impusieron fueron las relativas a las relaciones entre los continentes. Por otro lado, a pesar de la distribución global de los temas de Spotify pudimos identificar, aunque de manera débil, algunos términos asociados con Europa y con América del norte.

Por último, se exploró la posibilidad de que existiese una **asociación entre las palabras contenidas en las canciones según el mes del año**. Se encontraron reglas de asociación entre determinados términos y distintos meses del año; en particular se encontró que, para el mes de diciembre, las reglas difieren de todos los meses y también la emocionalidad promedio de las mismas.

En conclusión, en este trabajo se exploraron las palabras contenidas en las canciones de habla inglesa asociadas a otras variables, tales como los atributos de las canciones, su éxito, el mercado donde se encuentra disponible y el período del año donde fueron exitosas. La técnica de reglas de asociación nos permitió obtener nueva información acerca de la relación entre variables.