¿Cómo se asocia la música que escuchamos?

Grupo N°5: Flavia Felicioni, Adrian Marino, Claudio Collado

Resumen

Spotify es una aplicación multiplataforma empleada para la reproducción de música vía streaming. Cuenta con 345 millones de usuarios activos y 155 millones de usuarios suscritos, lo cual lleva a la generación de una enorme cantidad de datos con posibilidad de ser explotados para la obtención de conocimiento. En particular spotifycharts provee información diaria y semanal de los artistas que forman parte del top 200 a nivel mundial así como también a nivel país. Tomando esto como base junto con datos obtenidos de la API de Spotify y las letras de las canciones forman el conjunto de datos utilizados en este trabajo práctico de reglas de asociación.

1. Introducción

El presente trabajo práctico tiene como finalidad principal la aplicación de conceptos y metodologías de análisis vistos en la segunda parte de la materia *Data Mining* perteneciente a la *Maestría en explotación de datos y descubrimiento de conocimiento.* En particular se complementa el análisis realizado sobre los datos de Spotify y Charts explorados durante el Trabajo Práctico N°1 generando reglas de asociación que determinen la concurrencia de características/hechos en los datos.

2. Datos disponibles

El conjunto de datos entregados fue un backup de MongoDB en formato JSON. En total fueron cuatro colecciones de MongoDB que almacenan los siguientes conjuntos de datos:

- **Artist:** Incluye información de los artistas que obtuvieron alguna posición en el ranking durante el período analizado.
- **Artist_audio_features:** Guarda metadatos de las canciones que obtuvieron alguna posición en el ranking durante el período analizado.
- **Charts:** Tabla de rankeo de canciones (TOP 200) durante un período determinado
- **lyrics-dm**: Letras de los temas obtenidas de Genius: Para el desarrollo del TP se trabajó con aquellos temas correspondientes a Idioma Inglés.

3. Preprocesamiento y variables utilizadas

Esta etapa del trabajo es la que mayor impacto genera en la calidad de las reglas a obtener y la que mayor tiempo demanda en elaboración, por lo tanto se dedicó un tiempo considerable en su análisis e implementación modular.

3.1 Preprocesamiento de Texto

Al trabajar con base de datos de textos el preprocesamiento resulta fundamental para contar con el conjunto de palabras en condiciones necesarias para la posterior generación de reglas. A continuación se listan en líneas generales las etapas de preprocesamientos realizadas:

- Remoción de aquellos elementos que se generan al momento del scrapping para la obtención de las letras de Genius
- Conversión de todo el texto a minúsculas
- Remoción de palabras vacías, números, puntuaciones, espacios adicionales y todo aquellos que no corresponde a caracteres alfanuméricos.
- Generación de la matrix término-documento del corpus, filtrado de los n términos más frecuentes y binarización de la matriz (presencia del término)

3.2 Preprocesamiento de variables numéricas

El análisis por reglas de asociación corresponde a un algoritmo no supervisado el cual utiliza sólo tipo de datos categóricos y por lo tanto es necesario realizar una transformación de los features numéricos a categorías útiles y representativas.
De la misma forma que en el TP N°1 [3] los features numéricos disponibles son: danceability, acousticness ,energy, duration_ms, liveness, loudness, speechiness, tempo y valence.

Debido a que un mismo tema a lo largo del tiempo aparece en diferentes posiciones del ranking se tienen repeticiones de los features, solamente modificándose aquel correspondiente a la posición. Debido a esto se realizaron las siguientes operaciones:

- Agregación de los features con respecto a tema/artista/álbum
- Agregación por la media de la cantidad de reproducciones de cada tema/artista/álbum
- Agregación por la posición más alta que alcanzó cada tema/artista/álbum

Para una correcta discretización se analizó la distribución de cada feature y se decidió realizar una primera discretización como se observa en la Tabla N°1.

Feature	Categorías	Límites de las Categorías			
		Límite N°1	Límite N°2	Límite N°3	Límite N°4
danceability	low/medium/high	0	0.5	0.75	1
energy	low/medium/high	0	0.52	0.7	1
acousticness	low/medium/high	0	0.5	0.75	1
liveness	low/medium/high	0	0.5	0.75	1
speechiness	low/medium/high	0	0.5	0.75	1
valance	low/medium/high	0	0.52	0.7	1
position	low/medium/high	0.5	1.5	4.5	10.5

Tabla N°1

Es importante destacar que se dejó disponible la posibilidad de modificar estos límites de categorías según las características y particularidades de las preguntas en análisis.
Sumado a lo anterior se incorporaron las siguientes propiedades necesarias para ciertas preguntas:

a) TOP 1: Se definió una nueva variable denominada **top1**, la cual es utilizada para segmentar las canciones entre aquellas que llegaron al TOP 1 de las que no lo hicieron. En nuestro análisis sólo analizamos el top 10. Dado esto tenemos dos grupos:
   - top 1: Contiene las canciones que llegaron al top 1.
   - top 2 a 10: Contiene las canciones que llegaron al top 2 a 10.

   Luego, para discriminar las transacciones agregamos un nuevo término llamado **top1** el cual puede tomar los valores **yes** / **no.**

b) Best Albums: Se definió otra propiedad la cual permite separar o segmentar canciones en dos grupos:
   - Canciones que se encuentran en álbumes que tienen más de dos canciones en el top 10.
   - Canciones que no se encuentran en dichos álbumes.

   Luego para representar ambos grupos utilizamos la propiedad **best_albums** la cual puede tomar los valores **yes** / **no.**

4. Reglas

4.1 Pregunta N°1

Para aquellas canciones que ingresaron al TOP 50 durante el período 2018-2020, interesa analizar la asociación entre los niveles de los atributos musicales y los términos del vocabulario frecuentemente utilizado en sus letras

Considerando que el ingreso al TOP 50 representa un posicionamiento consolidado de los temas resultó interesante analizar cómo se relacionan los atributos musicales de los temas en conjunto con el vocabulario más frecuente utilizado, de forma tal de identificar asociaciones que puedan tal vez ser utilizados por los artistas en la planificacion y elaboracion de sus temas con el objetivo de lograr su posicionamiento en este ranking.

Con respecto al preprocesamiento se utilizó lo indicado en la sección 3. de este trabajo, modificando los valores límites de la categorización de *speechiness* y *position* según se indica en la Tabla N°2 por considerarse estos más adecuados.

Feature	Categorías	Límites de las Categorías				
		Límite N°1	Límite N°2	Límite N°3	Límite N°4	Límite N°5
speechiness	low/medium/high	0	0.07735	0.5	1	----
position	very high/high/medium/low	0	1	10	25	50

Tabla N°2

Se utilizaron como punto de partida (criterios de mínima) un valor de support = 0.08 y un valor objetivo de confidence superior a 0.6. En el análisis se realizaron sucesivos filtrados del conjunto de reglas de forma tal de observar en el consecuente (rhs) los niveles de los atributos musicales considerados de interés.
En la Tabla N°3 se observan las reglas encontradas y consideradas interesantes:

Regla N°	lhs	rhs	support	confidence	lift
1	energy=high,liveness=low,valence=low	danceability=medium	0.106	0.693	1.281
2	aint, bitch, got, like	danceability=high	0.080	0.707	1.928
3	danceability=high,energy=medium	speechiness=medium	0.113	0.650	1.364
4	aint, get, just, like, make	{speechiness=medium}	0.102	0.921	1.934

Tabla N°3

Del análisis de las reglas obtenidas se observa que las mismas tienen como consecuentes (rhs) solo a los atributos *danceability* y *speechiness*. No fue posible encontrar reglas interesantes (o que cumplan con los criterios de mínima establecidos) relacionadas a los atributos musicales: *energy*, *acousticness*, *liveness* , *valance* y *position*.

En la Tabla N°4 se observa el vocabulario asociado a los atributos musicales que se observan en el consecuente (rhs) de la Tabla N°3, correspondiente a un conjunto mayor de reglas de las observadas en la Tabla N°3:

Feature	Nivel	Vocabulario
danceability	high	get,just,like,yeah,aint,bitch,got
danceability	medium	know,like,just,see,yeah,cause,always,never
speechiness	medium	aint,make,fuck,like,man,yeah,see

Tabla N°4

En general el vocabulario que se observa se repite en algunos casos para los features analizados. En particular los términos “like”, “know” y “yeah” es posible considerarlos para su inclusión en la lista de *stopwords* extras debido a la poca información que pueden aportar. En la misma línea también es posible descartar palabras como “get” o “got” y “aint”.

4.2 Pregunta N°2

Para las canciones que ingresaron al TOP 10 durante el período 2018-2020 ¿Cuáles son los niveles de los atributos y los términos del vocabulario frecuente de las letras que más inciden

en la permanencia de las canciones en el puesto número 1 respecto de aquellas que nunca lo alcanzan?

En este análisis utilizamos la propiedad top1 para inicialmente separar o segmentar las transacciones/canciones entre las que top 1 y las que solo llegaron a tops entre el 2 al 10.

4.2.1 Generación de reglas

Para generar las reglas en base a las transacciones se seleccionó el support en 0.08 y confidence a 0.1, para tener un rango de reglas extenso y no filtrar solamente reglas generales. Luego, se realizaron consultas para analizar subsets de reglas y así analizar features y su vocabulario asociado en cada caso. La Figura N°1 muestra **lift** vs. **support** y el nivel de **confidence** en cada punto para el set de reglas generado:

Figura N°1

De la Figura N°1 interesa encontrar reglas con un lift entre 1 y 2 y una alta confianza, pero no tan alta, ya que en ese caso las reglas sólo proporcionan información trivial.

4.2.2 Features encontrados en el top 1 vs top 2 a 10

Al filtrar por el *top1=yes* se encontraron las features que se observan en la Tabla N°5:

Regla N°	lhs	rhs	support	confidence	lift
1	top1=yes	liveness=low	0.152	1	1.0058
2	top1=yes	speechiness=low	0.152	1	1

Tabla N° 5

Como se puede apreciar para el caso de reglas para canciones que llegaron al top 1 sólo se encontraron dos features que se analizan a continuación:

Speechiness

- Mide el grado de presencia de voces habladas en el audio de una canción.
- En este caso se puede decir que hay baja presencia de voz hablada en los temas.
- Si repasamos los hallazgos encontrados en el TP1 esto también coincide. En la Figura N°2 se aprecia que el grado de voz hablada al llegar al top 1 decae drásticamente.

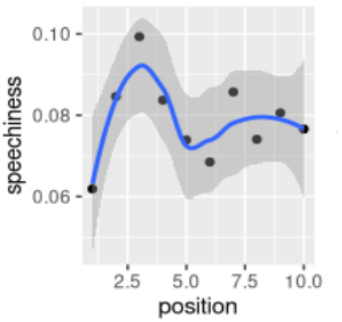

Figura N°2

Liveness

- Mide el grado de presencia de audiencia en la grabación de un tema.
- Se aprecia que ningún tema en el top 1 tiene audiencia, solo se tienen valores bajos.
- Es interesante notar que se llegaron a los mismos resultados para el top1 en el TP1: Según se observa en la Figura N°3 los temas correspondientes al top 1 no tienen presencia de audiencia.
- En definitiva, sólo se tienen temas grabados en estudio.

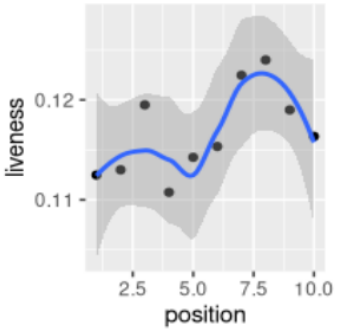

Figura N°3

Luego al filtrar por el *top1=no* se encontraron las features que se observan en la Tabla N°6:

Regla N°	lhs	rhs	support	confidence	lift
1	top1=no, can, know, right	danceability=medium	0.087	0.78	1.5
2	top1=no, position=medium, just, now	valence=low	0.11	0.86	1.491
3	top1=no	danceability=high	0.087	0.6	1.486
4	top1=no, aint, get, see	energy=medium	0.087	0.6	1.486

Tabla N° 6

Danceability

- Mide cuán bailable es una canción.
- Se aprecia que fuera del top 1 también se tienen temas de bailabilidad alta como se encontró en el TP1 y además se tienen niveles medios.
- Como se vio anteriormente en el TP1 y se observa en la Figura N°4 a medida que nos alejamos del top 1 la bailabilidad baja, coincidiendo esto con los resultados obtenidos.

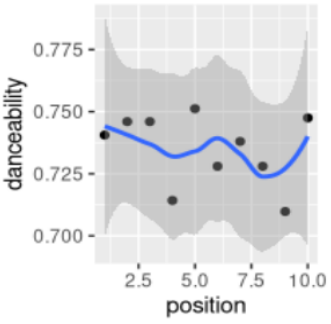

Figura N°4

Energy

- Se puede apreciar que para el top 2 al 10 los niveles de energía son medios indicando que son temas musicales con una intensidad media.
- Nuevamente vemos coincidencia con el TP1 y según se observa en la Figura N°5 donde en las misma posiciones tenemos presencia de niveles medios y bajos de energía.

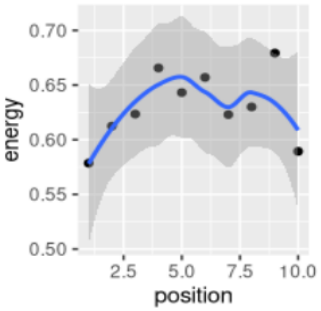

Figura N°5

Valencia o Positividad

- Es una medida de la positividad que transmite el tema, felicidad, alegría, euforia.
- En este caso los niveles de positividad encontrados son bajos, discrepando con los resultados obtenidos en el TP1 y según se observa en la Figura N°6. Entendemos que esto tiene que ver con la decisión de los rangos utilizados al discretizar los valores de esta variable.

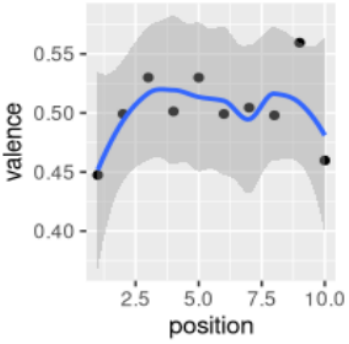

Figura N°6

4.2.3 Vocabulario por top y feature

En la Tabla N°7 se observan los vocabularios encontrados segmentados por top y características encontradas en cada uno:

Top	Feature	Nivel	Vocabulario
1	valence	Bajo	let, back, know, just, love, now, yeah y get.
1	speechiness	Bajo	let, back, know, just, love, now, yeah y get.
1	valence	Bajo	way, back, want, aint, never, right, make y aint.
1	danceability	Medio	can, one, say, see, around, make, really, leave y wanna.
1	danceability	Alto	need, put, niggas, fuck, tell, put, never, aint, gotta, day, man, want y way.
1	energy	Medio	can, aint, let, leave, need, man, can, shit, can, see, put y cause.

Tabla N° 7

4.2.4 Observaciones

- Podemos apreciar que no se encontraron diferencias al segmentar por las categorías encontradas (valence y speechiness) en el top 1. En ambos casos el vocabulario no difiere. También se aprecia que el vocabulario condicen con los niveles de los features donde vemos palabras como “love”, “back” que condicen con positividad. No encontramos una relación con el feature temas hablados pero es entendible que temas que hablan de amor en general no son hablados.
- Para el top 2-10 se observa un vocabulario distinto a los demás para bailabilidad alta, donde predominan términos fuertes como “fuck”, “niggas” y otros que suelen acompañarlos. Para energía alta se observa que también predominan palabras fuertes como “shit”. La positividad tiene términos similares y otros diferentes al top 1 como: “back”, “never”, “want”. En general, los términos expresan sentimientos parecidos pero no son los mismos.

4.3 Pregunta N°3

En las letras de las canciones de los artistas con discos más exitosos editados entre 2018 y 2020 ¿El vocabulario utilizado en la composición es más limitado respecto del utilizado frecuentemente en otras canciones del TOP 10? ¿Los términos del vocabulario utilizado en estas letras permiten caracterizarlas como mayormente positivas y vincularlas con alguna de sus propias características musicales?

Para responder a esta pregunta se utiliza la variable **best_albums** definida en la sección 3.2 b. En cuanto al vocabulario utilizado en estas canciones se encontró que existen algunas palabras que aparecen en el 70% de las canciones analizadas, como por ejemplo “like”, “know” y “yeah”. Dichos términos se agregan a la lista de *stopwords* extras debido a la poca

información que pueden aportar al ser ítems tan frecuentes. También, se descartan de esta búsqueda algunas palabras como “get” o “got” y “aint” ya que son conjugaciones de verbos que no aportan valor. Tras haber realizado este filtrado las 10 palabras más frecuentemente encontradas fueron: now (52%), cause (47%), love (47%), one (41%), baby (41%), say (41), see (40%), make (40%), never (37%), way (37%).
Se obtuvieron 13 reglas que tienen como consecuente la variable “best_album” (habiéndose también excluído los términos frecuentes presentes en ambos tipos de álbumes). Las más interesantes se listan en la Tabla N° 8:

Regla N°	lhs	rhs	support	confidence	lift
1	bad, make	best_album=yes	0.105	0.98	1.97
2	even , time	best_album=yes	0.105	0.66	1.46
3	cause, never	best_album=yes	0.122	0.65	1.43
4	back, keep	best_album=no	0.105	0.75	1.38
5	back, tell	best_album=no	0.111	0.70	1.3

Tabla N°8

4.3.1 Observaciones del vocabulario por tipo de álbum

- Del análisis de reglas se desprende que hay términos que aparecen frecuentemente en el vocabulario de las letras de los artistas de los álbumes exitosos (“bad”, “cause”, “even”, “time”, “let”, “make”, “never”, “one”, “take”, “time”, “want”) y en cambio pocos que aparecen en el resto de álbumes (“back”, “keep” y “tell”).
- Si bien no se puede concluir decisivamente en cuanto a la diversidad de vocabulario utilizado en las letras, se puede apreciar una mayor variedad de los términos frecuentes del vocabulario en las canciones de artistas con álbumes exitosos.

Para poder abordar el análisis si las letras utilizan mayormente términos positivos o negativos se utilizó el paquete **tidytext** de R [2] y en particular el diccionario de propósito general **affin** [1] (sólo para términos en inglés). Dicho diccionario asigna un score en un rango entre -5 y 5 para caracterizar si el sentimiento de la palabra es negativo o positivo (donde -5 indica un término muy negativo, un 5 un término muy positivo y el 0 un término neutral).
Claramente, no podemos concluir si una canción expresa un sentimiento positivo o negativo a partir del análisis de palabras aisladas y más aún teniendo en cuenta que se han eliminado términos que podrían negar determinada expresión, por ejemplo ‘not’. Por esto, sólo se pretende investigar si la tendencia a la utilización de los términos frecuentes en las letras se orienta más hacia la positividad, la neutralidad o la negatividad.
Para ello se calcula la variable numérica *positiveness* como la suma de los scores de los términos, de acuerdo a **affin**) respecto de la cantidad de palabras identificadas en cada canción. La versión discretizada de *positiveness* permite generar la variable categórica *positive* que asume alguno de los niveles {*low*, *medium*, *high*}, en el cual *medium* se aplica a canciones con valores entre -0.5 y 0.5 mientras *low* y *high* resultan los casos por debajo de -0.5 y por encima de 0.5, respectivamente.
Las proporciones de esta categorización son 50% *low*, 31% *medium* y 19% *high*, con lo cual tienen un soporte razonable para la generación de reglas. Se puede observar la nube de las palabras de las canciones más frecuentes identificadas dentro del diccionario **affin** (que tengan una presencia de al menos un 15% en las canciones).

La discretización de las variables categóricas se realiza de acuerdo a lo establecido en la Tabla N°1 y lo mencionado en Tabla N°2.
Se obtuvieron una gran cantidad de reglas, de las cuales tenemos mayor interés en aquellas que involucren la variable *positive*. Las más interesantes se listan en la Tabla N°9:

Regla N°	lhs	rhs	support	conf.	lift
1	call, cause, like, love	positive=low	0.102	0.88	1.77
2	liveness= low, positive=medium, valence=low	speechiness=low	0.102	0.72	1.62
3	danceability=medium, positive=low, speechiness=medium, valence=low	liveness= low	0.102	1	1.81
4	energy=medium, positive=low, valence=low	liveness= low	0.128	0.90	1.65
5	danceability=high, liveness= low, positive=low	speechiness=medium	0.105	0.88	1.61
6	positive=high	danceability=medium	0.123	0.67	1.3

Tabla N°9

4.3.2 Observaciones del análisis de emoción del vocabulario

- Se aprecia que si bien las palabras que aparecen con mayor frecuencia son “love” y “like” con score de 2 y 3 en el diccionario **affin**, las canciones son mayormente calificadas con valor bajo de positividad. Esta observación se puede derivar de la nube de palabras y se respalda con la regla N°1 de Tabla N°8.
- Para los valores *high* de positive no aparecen gran asociación con otras variables categóricas, se presenta sólo la regla 6 que indica que está asociada con un nivel medio de *danceability*. En cuanto a la opción *medium* de *positive* se puede observar que se relaciona con bajo nivel de *speechiness* (*low*), mientras que para la opción *low* de positive las reglas más interesantes informadas 3, 4 y 5 coinciden con valores *bastante bajos* de *speechiness* (*medium*), pero hay versiones tanto con niveles de *danceability* medios y bajos.

5. Conclusiones

- La etapa de preprocesamiento de los datos disponibles demandó la mayor parte del tiempo invertido en la elaboración del TP. La correcta realización de este preprocesamiento repercute directamente en la calidad de las posibles reglas a obtener, el cual a su vez toma diferentes caminos (generación de variables y análisis particulares) en función del tema/pregunta que se este analizando y pretenda resolver
- La generación y evaluación de aquellas reglas consideradas útiles para el problema en análisis fue un desafío en sí mismo ya esto se corresponde a un proceso iterativo que incluye pruebas, combinación de variables de interés, cumplimiento de los valores de confianza/soporte de mínima y luego su interpretación en forma de reglas. Por momentos fue frustrante y llevó, en algunos casos, a replantearnos las preguntas utilizadas como disparadoras del TP por la imposibilidad de encontrar reglas acordes.
- Si bien los resultados de análisis de vocabulario no son decisivos se pudieron apreciar algunas asociaciones distintivas en el vocabulario utilizado en las canciones de los álbumes más exitosos en cuanto a la diversidad y la emoción que expresan.

6. Anexo

Github: [magistery-tps/dm-tp2]

[1] F. Nielsen. "AFINN", (2011) "Informatics and Mathematical Modelling, Technical University of Denmark", http://www2.imm.dtu.dk/pubdb/pubs/6010-full.html.

[2] J. Silge, D. Robinson. (2017). Text Mining with R: A Tidy Approach 1st Edition. O'Reilly Media https://www.tidytextmining.com/

[3] Grupo 5. Trabajo práctico N° 1. Datamining 2021. Maestría en Explotación de Datos y Descubrimiento de Conocimiento. Universidad de Buenos Aires.