Facultad de Ciencias Empresariales

Estudio de la evolución de la música y su relación con los cambios en la sociedad a través de técnicas de análisis de datos

Clave: 201900866
Autor: Marta Ybarra Fernández-Iriondo
Director: Lucía Barcos Redín

MADRID | Junio 2024

Índice

Índice de Figuras

Índice de Tablas

Resumen ejecutivo

La música es un fenómeno universal que, aunque a menudo pueda pasar desapercibido, está mucho más presente en nuestra vida cotidiana de lo que podríamos imaginar. La música es considerada por muchos como la voz o el reflejo de la sociedad. La digitalización, la piratería musical, la llegada de las nuevas tecnologías y el surgimiento de las plataformas de streaming han transformado por completo la industria musical. La forma en que se consume, percibe y utiliza la música ha cambiado.

Este trabajo tiene como objetivo investigar los cambios que ha experimentado la música a lo largo de las últimas siete décadas y determinar si estos cambios están relacionados y reflejan las transformaciones sociales. Para ello, se analiza una base de datos de 1.050 canciones de las siete últimas décadas, teniendo en cuenta tanto sus métricas musicales (como la bailabilidad, la energía ...) como el contenido y temáticas de las letras. Las fuentes para construir esta base de datos han sido las APIs de Spotify y Genius, siendo necesario el uso de la técnica de Web Scraping para poder completar la descarga de las letras de las canciones. El estudio de las métricas musicales se ha realizado mediante un análisis exploratorio, mientras que el contenido de las letras ha sido analizado mediante la técnica de Topic Modeling, más concretamente el algoritmo LDA. Para llevar a cabo la investigación, se ha optado por las listas de éxito "All Out..." de Spotify, las cuales son compilaciones editoriales que abarcan las décadas desde los años 60 hasta la actualidad.

A partir de los análisis realizados y los resultados obtenidos, es razonable afirmar que la música sí que refleja los cambios de la sociedad. Como respuesta directa a la evolución de la sociedad, a los cambios en sus comportamientos, a la globalización y al aumento de la conectividad, la música ha ido cambiando: la duración de las canciones ha disminuido, el contenido inapropiado en las letras ha aumentado y los temas de las canciones han evolucionado. Ciertos temas, como aquellos que reflejan los desafíos urbanos y sociales y que abordan el sentido de la diversión y la felicidad, están cobrando importancia en la actualidad. Además, se ha notado una mayor diversidad tanto en los artistas como en los idiomas de las canciones.

Palabras clave: Topic Modeling, Latent Dirichlet Allocation (LDA), métricas musicales, Spotify, contenido lírico, evolución, industria musical

Executive summary

Music is a universal phenomenon that, although often unnoticed, is far more present in our daily lives than we might imagine. Music is considered by many to be the voice or reflection of society. Digitalization, music piracy, the advent of new technologies, and the emergence of streaming platforms have completely transformed the music industry. The way music is consumed, perceived, and utilized has changed significantly.

This study aims to investigate the changes that music has experienced over the past seven decades and determine if these changes are related to and reflect social transformations. To achieve this, a database of 1,050 songs from the last seven decades was analyzed, considering both their musical metrics (such as danceability, energy, etc.) and the content and themes of their lyrics. The sources for constructing this database were the Spotify and Genius APIs, requiring the use of Web Scraping techniques to complete the download of song lyrics. The study of musical metrics was conducted through exploratory analysis, while the content of the lyrics was analyzed using Topic Modeling techniques, specifically the LDA algorithm. For this research, the "All Out..." hit lists from Spotify were chosen, editorial compilations covering the decades from the 1960s to the present.

Based on the analyses conducted and the results obtained, it is reasonable to affirm that music indeed reflects societal changes. In direct response to societal evolution, changes in behaviour, globalization, and increased connectivity, music has undergone significant transformations: song durations have decreased, inappropriate content in lyrics has increased, and song themes have evolved. Certain themes, such as those reflecting urban and social challenges and those addressing a sense of fun and happiness, are gaining prominence today. Additionally, there is greater diversity in both the artists and the languages of the songs.

Keywords: Topic Modeling, Latent Dirichlet Allocation (LDA), musical metrics, Spotify, lyrical content, evolution, music industry

Capítulo 1. Introducción

1.1 Motivación

La música es un fenómeno universal, que, aunque pueda pasar por desapercibida y simplemente como un par de acordes bien entonados, tiene mucha influencia en la sociedad (Varnun et al., 2021). Está mucho más presente en nuestro día a día de lo que podríamos imaginar. Aproximadamente una cuarta parte del tiempo que pasamos caminando lo dedicamos a escuchar música (Sust et al., 2023). Es más, en 2021, la Federación Internacional de la Industria Fonográfica estimó que el tiempo dedicado a escuchar música a nivel mundial alcanzó las 18,4 horas semanales por persona, lo que equivale a 1,6 horas diarias por persona. Esto significa que en 2021 los aficionados a la música escuchaban, en promedio, 368 canciones por semana o 32 canciones cada día. Estos datos subrayan la relevancia y el papel esencial que tiene la música en nuestras vidas (IFPI, 2022).

Para muchos, la música actúa como "la voz de la sociedad", un espejo que refleja los acontecimientos en todas las culturas del mundo. Muchos la consideran un antídoto frente a las circunstancias actuales, marcadas por la "anormalidad" —una época de incertidumbre, inestabilidad política y económica, pandemias mundiales y conflictos bélicos. El cantautor Andrés Suarez expresó que "lo único que puede salvar a una sociedad cargada de dolor y llantos es la música" (Alonso, 2017). La música posee esa capacidad única de levantar el ánimo de los desalentados y ofrecer consuelo a quienes sufren.

Es un hecho que el mundo que conocemos hoy en día dista notablemente de la realidad que vivieron nuestros abuelos e incluso nuestros propios padres. Nos encontramos en medio de un proceso de transformación profunda y radical, donde los cambios suceden a gran velocidad. La introducción de nuevas tecnologías, como los ordenadores, ha reemplazado por completo a los medios tradicionales como el lápiz y el papel, y la llegada de nuevas plataformas digitales ha revolucionado todas las industrias, ampliando significativamente las posibilidades de disfrute y consumo. Además, la conectividad ha alcanzado un nivel tal que ahora es posible mantener contacto con personas a kilómetros de distancia sin necesidad de salir de casa. En resumen, las reglas del "juego" que regían

el mundo están cambiando, impactando no solo a la sociedad, sino también al entorno en el que operan las empresas y los distintos sectores e industrias.

En el ámbito empresarial, la adaptación a estos nuevos cambios no es una mera opción ni una elección de unos pocos, sino una obligación. Si las empresas no quieren quedar desactualizadas y perder relevancia deben evolucionar en un entorno que está en constante cambio. Y es evidente que la industria musical no es una excepción a esta regla. La digitalización ha transformado radicalmente la industria musical; la forma en la que los consumidores y productores escuchan, perciben y utilizan la música ha cambiado por completo (Stafford, 2010). La llegada de las nuevas plataformas de streaming como Spotify, Apple Music y Amazon Music han transformado la industria musical: los tradicionales CD y radiocasetes se han quedado anticuados, las plataformas musicales y redes sociales se han convertido en los nuevos canales de promoción, y las tiendas de música física han prácticamente desaparecido. En 2021, el tiempo dedicado a la escucha de música a través de plataformas de suscripción como Spotify, Apple Music o iTunes, se incrementó en un 51% con respecto al año anterior (IFPI, 2022).

Por ello, conscientes de los cambios de época y las demandas de las nuevas generaciones, artistas, músicos, discográficas y promotores musicales se han visto obligados a evolucionar. Al fin y al cabo, el paso del tiempo siempre implica cambios: cambios sociales, en las preocupaciones de las personas, en las tendencias, en lo que se busca y se persigue con la música y en el comportamiento de la sociedad debido a acontecimientos históricos. Estas transformaciones impulsan cambios en los formatos, los patrones, las temáticas e incluso la duración y la letra de las canciones para poder adaptarse a estas nuevas tendencias y expectativas. Además, en los últimos años se ha observado un notable incremento en el uso de herramientas de análisis de datos y minería de textos. Estas innovadoras técnicas, como el Topic Modeling, representan una nueva oportunidad para realizar un análisis más profundo de las canciones, específicamente en relación con ciertos cambios en su estructura técnica y contenido.

1.2 Objetivos

Este trabajo tiene como objetivo examinar cómo han evolucionado las canciones a lo largo de las últimas siete décadas (1960-2020) haciendo uso de técnicas de análisis de datos y de minería de textos. Esto implica el estudio, tanto de las métricas musicales, como del contenido de las letras de sus canciones. Además, se busca analizar si los

cambios identificados en las canciones pueden responder a ciertas transformaciones en la sociedad. Así, con este trabajo se pretende dar respuesta a las siguientes preguntas:

- ¿Son las canciones actuales iguales que las canciones de hace varias décadas? Con el paso del tiempo, ¿qué características técnicas de las canciones han evolucionado y cambiado significativamente? ¿Contribuyen todas estas características de igual manera al éxito o popularidad de las canciones?
- ¿Cómo han evolucionado a lo largo del tiempo los temas de los que tratan las canciones?
- ¿Son las canciones reflejo de cambios sociales contemporáneos como la globalización, la mayor conectividad o la reivindicación de la mujer?

1.3 Estructura y Metodología

La estructura del trabajo de investigación se divide en tres secciones: una primera parte en la que se revisa la literatura y se profundiza en las técnicas utilizadas construyendo así el marco teórico del trabajo de investigación, una segunda parte enfocada en la parte analítica del trabajo centrándose tanto en las características musicales como en el contenido de las propias canciones, y finalmente, una última sección en la que se presentan las conclusiones.

La primera sección que busca construir el marco teórico sobre el que fundamentar el trabajo de investigación se presenta a lo largo de los primeros tres capítulos. Se realiza una revisión detallada de estudios previos, seleccionando artículos académicos, informes y trabajos de investigación para analizar la evolución de la industria musical en las últimas décadas. Este análisis incluye el diseño y estructura de las canciones, el desarrollo del contenido musical y la posible relación que ya ha sido investigada entre la música y la sociedad. Para reforzar esta sección y clarificar el análisis posterior, se profundiza en las técnicas de Web Scraping y Topic Modeling (LDA). Esto nos permite establecer una base sólida para iniciar nuestro estudio.

La segunda sección se corresponde con la parte analítica del trabajo que incluye los capítulos 4, 5 y 6. La investigación utiliza las listas de éxitos proporcionadas por la plataforma de streaming Spotify, abarcando las décadas desde los años 60 hasta la actualidad. Para alcanzar el objetivo, se realiza un análisis exhaustivo de las canciones, abordando tanto sus características técnicas—como la energía, la bailabilidad, la

positividad y el volumen, entre otras —como sus aspectos más cualitativos, relacionados con el contenido de las canciones. A través del lenguaje de programación R, de sus modelos y algoritmos, de sus estadísticas descriptivas, sus series temporales y sus visualizaciones se busca estudiar y comparar estas características a lo largo de las últimas décadas para así analizar su evolución y entender los fundamentos que pueden motivar estos cambios. Para el análisis del contenido de las canciones, se emplea el método de Topic Modeling en R, una técnica de minería de textos y más concretamente Latent Dirichlet Allocation (LDA), que es uno de los algoritmos de Topic Modeling más conocidos y aplicados. Este nos permite examinar la evolución de los temas y el contenido musical a lo largo del tiempo.

Esta aproximación integral nos proporciona una visión más clara de la evolución de la música, permitiéndonos alcanzar conclusiones significativas y bien fundamentadas. Esta sección se organiza siguiendo el orden de las fases del análisis: primero se construye la base de datos (capítulo 4); posteriormente se realiza el análisis de las métricas musicales (capítulo 5); seguido el estudio del contenido de las canciones (capítulo 6).

Y finalmente, la tercera sección (capítulo 7) que incluye la discusión de los resultados para alcanzar las conclusiones derivadas de nuestro análisis, las limitaciones encontradas durante la investigación y las posibles líneas de investigación para estudios futuros.

Capítulo 2. Revisión de la literatura sobre aspectos relacionados con la estructura actual de las canciones

La música, como componente fundamental de nuestra sociedad y que lleva existiendo muchas décadas, ha sido objeto de numerosos estudios e investigaciones. En esta sección, examinamos en detalle algunas contribuciones académicas relacionadas con nuestro tema de estudio.

Con la creciente presencia de la era digital en nuestras vidas, el panorama de la industria musical ha cambiado por completo provocando un aluvión de cambios, elevando esta industria a nuevos niveles y alterando la relación entre las discográficas, los artistas y los consumidores. Las discográficas han perdido todo su poder frente a los artistas y propios consumidores. Los sellos discográficos ya no son imprescindibles para los artistas pues ahora estos tienen la capacidad suficiente para gestionar sus propios proyectos de manera independiente, promocionarse a sí mismos y ser dueños de su propio trabajo (Stafford, 2010). Además, las discográficas no solo han perdido importancia, sino que también han experimentado dificultades financieras. A la desaparición de los CDS y discos, se le añade la aparición de un nuevo desafío: la piratería musical. Esta descarga ilegal de canciones ha dificultado la generación de ingresos para las entidades musicales, ya que ahora es posible lanzar y distribuir canciones sin que las discográficas o los artistas reciban beneficios económicos, así como descargar, enviar y recibir música sin coste alguno. La industria musical se encuentra en plena disputa legal para terminar con este desafío y poder recuperar su valor e importancia. Este es uno de los grandes cambios a los que se está enfrentando la industria musical actualmente. Por ello, se están reinventando para conseguir recuperar la posición que tenían hace décadas (Stafford, 2010).

La existencia de la "piratería musical" no sólo ha perjudicado las cuentas de pérdidas y ganancias de las discográficas, sino también a la supervivencia de las canciones en las listas de reproducción. Las nuevas formas digitales de compartir música han revolucionado el rol de los artistas, haciendo más sencillo y accesible el proceso de ganar reconocimiento y alcanzar a audiencias globales, una posibilidad que antes de la revolución tecnológica resultaba inimaginable. Esta nueva era digital brinda la oportunidad a los músicos emergentes de grabar sus canciones, subirlas a las plataformas y darlas a conocer sin la necesidad de una gran inversión ni del apoyo de las grandes

discográficas (Bhattacharjee et al., 2007). Esto ha alterado la trayectoria tradicional de los artistas. Anteriormente, los artistas seguían caminos más predefinidos y estructurados, dependiendo completamente de intermediarios de la industria. En la actualidad, cuentan con mayor independencia, ya que pueden darse a conocer sin la necesidad de terceros gracias a las redes sociales como Instagram y TikTok, y sin seguir un recorrido predeterminado desde el inicio.

A esto se le suma que la llegada de estas nuevas plataformas de streaming, que ofrecen descargas e intercambios gratuitos y rápidos de canciones, han dificultado la capacidad de las canciones y álbumes para mantenerse en las listas de éxitos. Las canciones y los álbumes que se encuentran en la parte inferior de las listas de éxitos están sujetos a constantes cambios, a una menor permanencia y a una mayor incertidumbre sobre su desempeño tras su lanzamiento. Debido a la disminución de barreras tanto regulatorias como económicas a la hora de lanzar una canción, ahora cualquier persona puede crear su propia música, dejando de ser un privilegio exclusivo de las grandes discográficas. Como consecuencia, el panorama competitivo se ha vuelto más complicado y la supervivencia en listas de éxito es todo un desafío (Bhattacharjee et al., 2007).

La llegada de las nuevas tecnologías, la piratería musical y plataformas de streaming, representan sólo algunos de los múltiples factores que han impulsado los cambios en la industria musical en las últimas décadas. A lo largo de este período, tanto el diseño estructural como el contenido lírico de las canciones han evolucionado significativamente, convirtiéndose en objeto de numerosas investigaciones académicas.

Por un lado, en referencia al estudio de la estructura de las canciones, numerosos investigadores han analizado cómo las composiciones pueden llegar a influir en el éxito de las canciones en el mercado. Algunos, como Interiano et al. (2018) en su trabajo de investigación, evaluaron las características musicales de las canciones y su correlación con el éxito, analizando hasta qué punto estos atributos contribuían al éxito de estas. El objetivo de su estudio era descifrar las dinámicas del éxito musical y ver si era posible prever el éxito futuro de las canciones. Tomando como base los datos extraídos de "MusicBainz" y "AcousticBainz", una "enciclopedia de información musical de código abierto mantenida por la comunidad" (MusicBrainz, s.f.), demostraron que no existen pautas específicas que, al seguirse de manera sistemática, garanticen el éxito de una canción. En otras palabras, generalizar el éxito era una tarea más compleja de lo que en un principio se podía pensar. La única conclusión obtenida fue que, para que una canción

lograse el éxito, debía tener esa capacidad distintiva, estar compuesta por unos patrones únicos y difíciles de replicar y distanciarse de las tendencias más generales (Interiano et al., 2018).

Por otro lado, con respecto al análisis de las letras de las canciones, cada vez más investigadores se dedican al estudio de las letras de las canciones debido a su creciente relevancia y su consideración como el "alma de la música" (Fischer y Greitemeyer, 2006). Las letras de las canciones tienen un gran poder y pueden impactar profundamente en la sociedad, en sus costumbres, cultura y valores.

En estos últimos años, muchos investigadores han estudiado acerca de la simplicidad y homogeneización de las canciones. Varnum et al. (2021), a través de un estudio sobre la comprensibilidad lírica de 14.661 canciones en el Billboard Hot 100 de 1950 a 2016, confirmaba un cambio en la complejidad de las letras. Utilizando el índice de compresibilidad, que evalúa la repetitividad y la densidad informativa de las letras, y medidas de correlación no paramétricas, demostró que las letras de las canciones más populares estaban volviéndose progresivamente más sencillas. Para describir este fenómeno, Varnum et al. (2021) se refirieron al "efecto de mera exposición", una teoría que sugiere que las canciones más simples y repetitivas tienden a captar más la atención de la audiencia musical. Las personas se suelen sentir más atraídas por aquellos elementos que se presentan de manera constante, debido a que, con la capacidad cognitiva limitada del ser humano, sólo procesamos la información esencial y buscamos atajos al tomar decisiones. En otras palabras, las personas prefieren aquello que es memorable, fácil de recordar y de transmitir. En las últimas décadas, el aumento en el número de canciones ha llevado a una preferencia por opciones más sencillas, lo que ha llevado a una reducción gradual de la complejidad lírica.

En cambio, otros investigadores, como por ejemplo Meindertsma (2019), optaron por investigar sobre la creciente homogeneidad de las canciones de los últimos tiempos. Utilizando un enfoque de macro analítico y examinando los cambios en las varianzas del nivel de agrado y del ratio de "tokens", demostró que las canciones del US Billboard Hot 100 son cada vez más similares entre sí. Estas nuevas incorporaciones presentan diseños más simples con menos singularidades. Mientras que muchos atribuyen este cambio a una creciente concentración en la propiedad de los medios de comunicación, con apenas tres grandes conglomerados globales de sellos discográficos dominando el mercado, otros lo relacionan con la globalización y la estandarización de los géneros musicales.

Además de estos estudios, algunos investigadores se han dedicado a estudiar la influencia que el contenido lírico de las canciones puede tener en el bienestar psicológico de las personas. En su trabajo de investigación, Fischer y Greitemeyer (2006) realizaron tres experimentos para analizar el impacto de las canciones con contenido sexualmente agresivo en los comportamientos y pensamientos hacía personas del mismo o del sexo opuesto. Los resultados mostraron que los hombres expuestos a letras de canciones agresivas eran más propensos a desarrollar actitudes de venganza hacia las mujeres, ya que recordaban predominantemente aspectos negativos sobre ellas, en comparación con aquellos que escucharon letras neutrales. Con esto demostraron que el contenido de las canciones sí que puede influir significativamente, tanto de manera positiva como negativa, en el comportamiento de la sociedad. Esto reafirma una vez más la relevancia y repercusión de la música en nuestra sociedad (Fischer y Greitemeyer, 2006).

Y finalmente, otros investigadores han profundizado aún más, examinando cómo el estudio de las canciones puede ser útil para industrias ajenas al ámbito musical. En su estudio, Dodds y Danforth (2010) analizaron la evolución del grado de felicidad transmitida en textos extensos, como títulos y letras de canciones, entradas de blogs y discursos políticos en los Estados Unidos. El objetivo era investigar la posible relación entre esta felicidad transmitida y ciertos acontecimientos históricos, como guerras, crisis o elecciones presidenciales que ocurrían en ese momento. Se observó que en Estados Unidos durante la elección de Kennedy como presidente se experimentó un notorio aumento en la expresión de felicidad, en contraste con lo ocurrido durante la Segunda Guerra Mundial. Mientras en momentos de crisis e incertidumbre, los textos de tono tristes y melancólicos se hacen más frecuentes, en épocas de estabilidad y tranquilidad política, la alegría y la positividad suele predominar. Este estudio destacó que ciertos textos como canciones o blogs sirven como indicadores del estado emocional de la sociedad en ese momento y su respuesta a eventos concretos. En consecuencia, Dodds y Danforth defienden que el análisis de grandes textos - comprender cómo, cuándo y por qué las personas experimentan diversas emociones – puede ser muy práctico para ayudar a los políticos, por ejemplo, a mejorar sus políticas públicas, o a los científicos a entender los fenómenos económicos y sociales.

En sus análisis sobre los cambios en la industria musical, la mayoría de los investigadores toman como referencia las listas de éxitos de la US Billboard Hot 100. Esta lista, que clasifica las canciones más populares del mercado musical estadounidense,

es publicada de manera exclusiva y semanal por la revista Billboard. Su clasificación se determina en función de las ventas de música, tanto físicas como digitales, del streaming en línea y de la radiodifusión a nivel nacional. A pesar de tratarse de una fuente fiable, tiene una limitación: sólo refleja las tendencias musicales estadounidense, por tanto, no considera las canciones que alcanzan popularidad fuera de los Estados Unidos. En cambio, algunos estudios han optado por la fuente de información de "MusicBrainz" para construir sus bases de datos. Se trata de una base de datos pública construida por las aportaciones de los propios voluntarios lo que puede llevar a un sesgo debido a las preferencias de las personas que cargan las canciones (Interiano et al., 2018). Incluso muchos investigadores, para facilitar el tratamiento de datos, utilizan subconjuntos de datos (por ejemplo, sólo teniendo en cuenta las diez primeras canciones de cada lista), limitando así el alcance y el detalle de la investigación.

Para este trabajo de investigación, se utiliza la plataforma de música digital Spotify como principal fuente de información. Spotify no solo destaca como una de las plataformas líderes en la actualidad, debido a su extenso catálogo de canciones y su gran base de usuarios, sino que también ofrece una API accesible. Esta API proporciona el acceso a una base de datos amplia y diversificada, que incluye canciones de múltiples géneros y décadas, además de información y detalles actualizados (Spotify, s.f.). A diferencia del US Billboard Hot 100, Spotify es utilizado y escuchado en todo el mundo, recopilando canciones de diversos países y no sólo de los Estados Unidos. En Spotify existen tres tipos de listas: las generadas por usuarios, las algorítmicas y las editoriales (Leighton, 2023). Para nuestro trabajo, estas últimas han sido las elegidas para recopilar las 150 canciones más exitosas de cada una de las últimas siete décadas, sumando un total de 1.050 canciones. Consideradas "la crème de la crème" para artistas emergentes, estas listas son gestionadas por el equipo interno de Spotify, que tiene un profundo conocimiento de las tendencias musicales y de las mejores canciones tanto a nivel local como global. Utilizando estos datos, se busca estudiar la evolución de la música, en concreto de las características musicales y del contenido lírico a lo largo de las últimas décadas para evaluar su posible relación con los cambios en la sociedad y contrastar algunas de las conclusiones obtenidas en las diferentes investigaciones.

Capítulo 3. Técnicas utilizadas para la construcción de la base de datos y para la identificación de los tópicos de las canciones

Para este trabajo de investigación, se utiliza exclusivamente el entorno de R para el procesamiento y análisis de datos. Este lenguaje es ampliamente utilizado en Ciencia de Datos y ofrece una variedad de paquetes que facilitan diversas técnicas de análisis (Lander, 2014). Para acceder a los datos y construir la base de datos sobre la que se apoya nuestro trabajo, se utilizan diversas APIs y la técnica de Web Scraping. Para el análisis de las variables musicales y las letras de las canciones, nos apoyamos en las visualizaciones en R y en la técnica de Topic Modeling (LDA). En este capítulo se profundiza sobre el Web Scraping y el Topic Modeling desde una perspectiva teórica, sin detallar su aplicación en nuestro trabajo.

3.1 Web Scraping

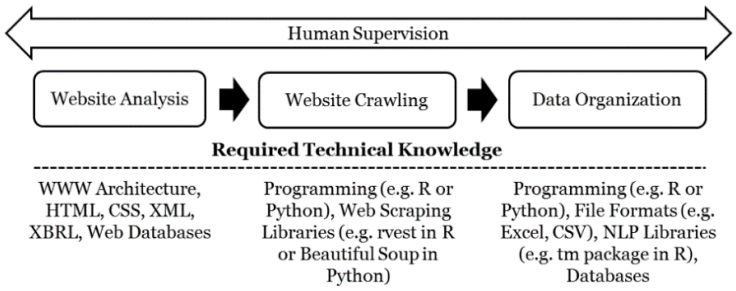
Gracias a la llegada de las nuevas tecnologías y al Big Data, las fuentes y plataformas donde poder encontrar datos ha crecido exponencialmente, al igual que la variedad en los tipos de datos disponibles. Actualmente, la fuente web es el tipo de fuente más utilizado ya que se trata de una red de alcance global que permite el acceso universal y que proporciona un gran volumen de datos y de información. Sin embargo, está asociado a unos problemas comunes como son la variedad, la velocidad y la veracidad de los datos (Sirisuriya, 2015). Los datos que se recogen varían en formato y calidad y su fiabilidad es incierta. Al final cualquier persona puede publicar información en internet sin necesidad de autorización previa y los datos al estar en línea están cambiando a tiempo real (Krotov y Silva, 2018). Aunque se trate de una fuente que requiere un largo procesamiento de los datos, trabajar con ella es muy beneficiosa porque permite la extracción de información valiosa y la obtención de resultados completos y consistentes.

Para recopilar información de fuentes web, se utiliza el Web Scraping, una técnica de extracción y transformación de datos web. Esta técnica tiene como propósito extraer la información de distintas páginas web, procesarla y adaptarla a formatos más adecuados para un análisis. Dicho en otras palabras, busca convertir la información no estructurada o semiestructurada en información estructurada, organizada y operable, para así almacenarla en hojas de cálculo o bases de datos estructuradas (Sirisuriya, 2015). Estos

formatos estructurados facilitan el trabajo con los datos y permiten la aplicación de un mayor número de algoritmos y modelos, obteniendo así un mayor provecho de los datos.

De acuerdo con Krotov y Silva (2018), el desarrollo y la implementación de esta técnica se puede dividir en tres fases *(Figura 1)*. En primer lugar, la fase de "Análisis del sitio web", una etapa en la que se busca analizar la estructura del sitio web de la que se quiere extraer información para entender su almacenamiento de datos a nivel técnico. Para ello, se requiere un conocimiento básico de los distintos lenguajes de marcado (por ejemplo, HTML, XML, etc.) y de las posibles bases de datos web (por ejemplo, MySQL). En segundo lugar, la fase de "Rastreo de web", una fase en la que mediante la ejecución del script (se suele utilizar R o Python), se procede a la navegación y exploración de manera automática por la página web con el fin de recoger los datos correspondientes para el propósito de la investigación. Y finalmente, una vez seleccionados los datos necesarios, se procede a la última fase de "Estructuración y organización de datos". Para aprovechar estos datos en análisis posteriores, es crucial limpiarlos, ordenarlos y estructurarlos adecuadamente en columnas y filas, eliminando redundancias e impurezas. A pesar de tratarse de un proceso mayoritariamente automático, sigue requiriendo de la supervisión humana para corregir errores, como la recopilación de datos incorrectos de otras páginas webs (Krotov y Silva, 2018). Esta técnica se utilizará para recoger las letras de las canciones de una página web llamada "Genius", la cual se explicará con más detalle en el próximo capítulo.

Figura 1. Fases para la implementación de la técnica del Web Scraping



Fuente: Krotov y Silva (2018)

3.2 Topic Modeling

Como se ha comentado en el apartado anterior, al encontrarnos en plena era digital, la mayor parte de la información se encuentra digitalizada y almacenada en plataformas en línea como blogs, páginas web y redes sociales (Tong y Zhang, 2016). Actualmente, la cantidad de datos disponibles excede ampliamente nuestras capacidades de procesamiento, lo cual nos presenta el desafío de gestionar y aprovechar estos datos de manera efectiva. Se estima que las empresas sólo logran utilizar aproximadamente el 1% de los datos no estructurados que almacenan (DalleMule y Davenport, 2017). Esto demuestra que no solo es importante el proceso de extracción y almacenamiento de datos, sino también el análisis y la exploración detallada del contenido recopilado.

Para poder darle un sentido a los datos extraídos, que, en nuestro caso, al ser letras de canciones se tratan de datos de textos sin estructura definida, es necesario acudir a las técnicas de Text Mining. Estas desempeñan un papel fundamental en la consecución de nuestro objetivo de estudiar la evolución de las letras de las canciones a lo largo de las últimas décadas ya que permiten una mayor exploración, organización y análisis de datos de textos. Según Tong y Zhang (2016, p.212), la minería de textos permite "obtener información de alta calidad del texto y generalmente involucra el proceso de estructurar el texto de entrada, encontrar patrones dentro de los datos estructurados, y finalmente la evaluación e interpretación del producto".

Parte del análisis realizado en este trabajo busca identificar de forma automática los tópicos de los que se hablan en las canciones y estudiar cómo evoluciona la prevalencia de estos tópicos a lo largo de las décadas. Y para alcanzar este objetivo, se utiliza la técnica de Topic Modeling, una herramienta de Text Mining, que permite el estudio de las letras de las canciones recogidas a través del Web Scraping. Se trata de un modelo generativo que busca comprender y resumir de forma automática grandes archivos de textos e identificar aquellos temas predominantes mediante el uso de un marco probabilístico (Blei et al., 2010; Tong y Zhang, 2016). Para entender bien el funcionamiento de este modelo, es importante tener una comprensión general de la composición de los documentos y corpus. La palabra es la unidad básica de los datos, por lo que una secuencia de N palabras conforma un documento y a su vez, un conjunto de M documentos crean un corpus.

En este contexto, la aproximación elegida para el desarrollo de este modelo es el "Latent Dirichlet Allocation (LDA)", que Tong y Zhang (2016, p.213) describen como "la técnica probabilística de modelado de texto más popular en el aprendizaje automático". Se trata de un método que utiliza un modelo probabilístico para extraer la estructura latente del texto, generando así tópicos en cada documento y determinando las probabilidades de temas en cada documento y de palabras en cada tema (Laoh et al., 2018).

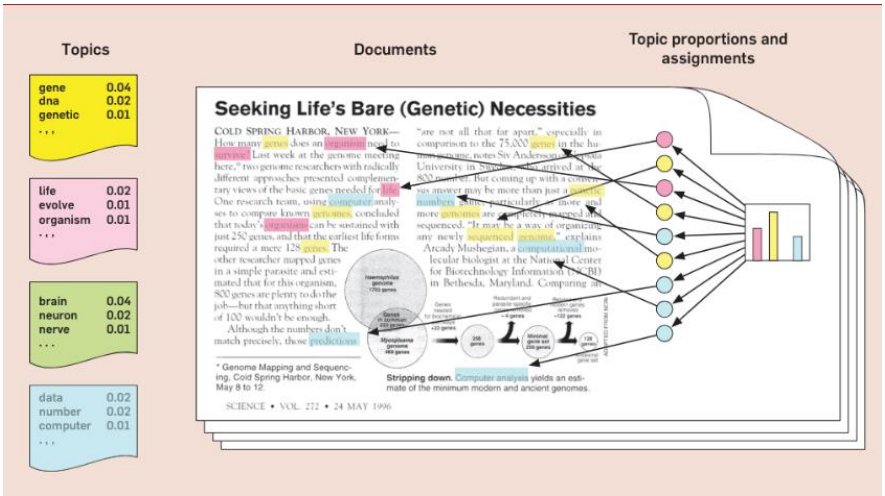
Esta técnica entiende que un documento está compuesto por varios temas y que un conjunto de palabras que tienden a aparecer juntas da lugar a un tópico. Luego, un documento, al estar formado por varios temas, tiene una distribución de probabilidad sobre temas y estos temas tienen una distribución sobre un vocabulario fijo (Tong y Zhang, 2016).

La siguiente representación gráfica, publicada por Blei (2012), permite entender su funcionamiento de una forma más visual, clara y sencilla *(Figura 2)*. El funcionamiento del modelo implica la participación de diversos elementos.

- Por un lado, se encuentra el corpus, un conjunto de documentos que a su vez están formados por un grupo de palabras y que son la base sobre la que se aplica el modelo LDA.
- Seguido, tenemos los tópicos, que son los temas identificados por el modelo LDA y que están compuestos por un conjunto de palabras y sus probabilidades correspondientes. Estas distribuciones, representan la importancia de esa palabra en ese tópico. Para la identificación de los tópicos, el modelo se basa en la distribución de las palabras dentro de cada tópico (Φ) y la distribución de los tópicos dentro de cada documento (θ) (Blei at al., 2010; Laoh et al., 2018). Los tópicos siguen una distribución Dirichlet con el hiperparámetro β, mientras que los documentos, con el hiperparámetro α. A diferencia de otros modelos, el número de tópicos a extraer no es determinado por el propio modelo, sino que es un parámetro de entrada que debe ser decidido previamente (k) (este paso se detalla más adelante).
- Y finalmente, está la asignación y distribución de palabras a temas. En la figura 2, la gráfica de barras muestra la proporción de cada tópico dentro

del documento mientras que las líneas representan la distribución de probabilidades asociadas con cada tópico presente en el texto. El modelo LDA identifica patrones en la distribución de palabras para descubrir temas, asignando a cada documento una distribución de estos temas que refleja su relevancia. Se podría decir que el modelo proporciona dos resultados: la lista de palabras clave por tópico y la asignación de cada documento a los tópicos (Bail, 2020).

Figura 2. Representación visual del modelo LDA



Fuente: Blei (2012)

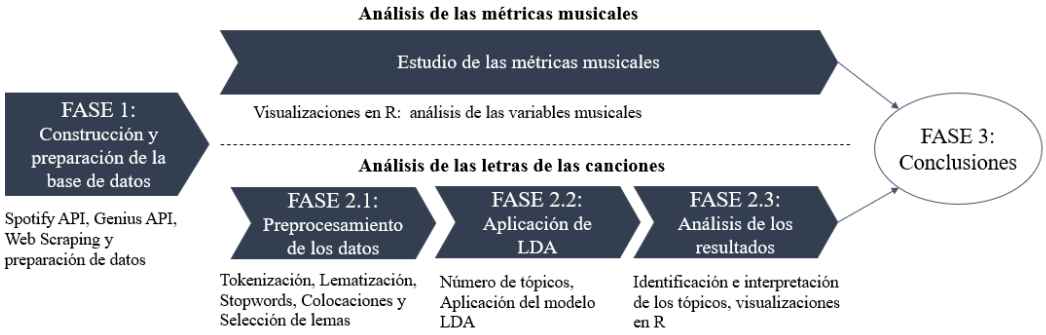
Se trata de un proceso iterativo que comienza con una asignación aleatoria de palabras a los temas. Se asume que todas las palabras tienen la misma probabilidad de pertenecer a cualquier tópico. Seguida de esta asignación, se asignan los tópicos a los documentos y se calculan las distribuciones de probabilidad Φ (palabras dentro de cada tópico) y θ (tópicos dentro de cada documento). Este proceso se repite ajustándose continuamente conforme a las distribuciones de probabilidad hasta identificar los tópicos relevantes del corpus. Es importante tener en cuenta que, para ejecutar el algoritmo y calcular estas distribuciones, son necesarios tres parámetros: (1) el hiperparámetro α, que define la distribución Dirichlet de los documentos; (2) el hiperparámetro β, relacionado con distribución Dirichlet de los tópicos; y (3) el número de tópicos a extraer (Blei et al., 2010). De forma resumida, el modelo sigue los siguientes pasos:

1. Se recopilan los documentos y se procede al preprocesamiento del corpus (Tareas de tokenización, eliminación de stopwords, lematización del texto, entre otras).
2. Se define el número de tópicos a extraer (más detalle en el apartado 6.1)
3. En una primera iteración, se asignan de forma aleatoria cada palabra a uno de los temas. Cada documento está compuesto por un conjunto de temas con sus proporciones respectivas, y estos temas a su vez por un conjunto de palabras con una cierta proporción. Es común que, en esta primera iteración, debido a su aleatoriedad, las asignaciones sean imprecisas y con posibilidad de mejora.
4. El modelo LDA revisa las asignaciones realizadas anteriormente, realiza los ajustes necesarios y repite este proceso hasta que las asignaciones de palabras a temas comiencen a estabilizarse.
5. Después de un número suficiente de iteraciones, las asignaciones de palabras a temas y de temas a documentos convergen en una distribución que da lugar a la estructura temática final.
6. Una vez verificada la relevancia de los tópicos, se procede al análisis e interpretación de los resultados. El modelo LDA no da como resultados el nombre de los tópicos, sino que son los investigadores los que basándose en las palabras que componen el tema, determinan y asignan un nombre adecuado (Bail, 2020).

En resumen, el modelo LDA, a partir de las distribuciones de probabilidades y según la relevancia de las palabras dentro de los tópicos y de los tópicos dentro de los documentos, identifica los tópicos más destacados del corpus (Tong y Zhang, 2016). Un aspecto peculiar y destacable del modelo es que las palabras no son exclusivas de un solo tema; pueden aparecer en varios temas, y su importancia varía según el contexto en el que se utilicen (Blei, 2012).

En este trabajo de investigación, se sigue la siguiente metodología (*Figura 3*). Se detalla cada paso en los siguientes capítulos.

Figura 3. Procedimiento del Trabajo de Investigación

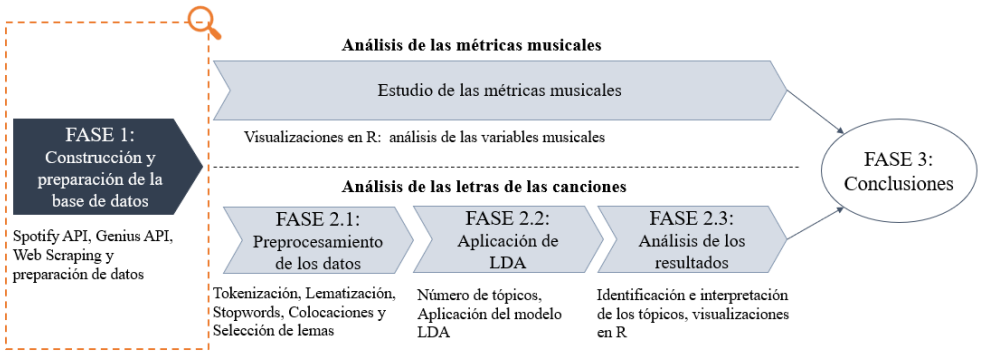


Fuente: Elaboración propia

Capítulo 4. Construcción y preparación de la base de datos

Siguiendo la estructura previamente indicada para la resolución de este trabajo de investigación, este capítulo se centra en la primera fase: la construcción y preparación de los datos (*Figura 4*).

Figura 4. Procedimiento del Trabajo de Investigación: Fase 1



Fuente: Elaboración propia

4.1 Recopilación de datos

En este estudio se investiga la evolución de las canciones, analizando tanto aspectos técnicos como cualitativos, como las letras. Para construir la base de datos, se han utilizado dos fuentes distintas: las listas de éxitos de Spotify para seleccionar las 150 canciones más populares de cada década (1960-2020) y Genius para obtener las letras de estas canciones identificadas.

Spotify ha sido elegida la fuente para la extracción de las características musicales ya que es la plataforma líder en el mercado de transmisión de música, capturando más de un 31% de la participación, y teniendo aproximadamente 550 millones de usuarios activos mensuales en todo el mundo y 220 millones de suscriptores (Mena Roa, 2023). En concreto, se han utilizado sus listas de reproducción “All out …” de las últimas siete décadas. Éstas, que son universales y accesibles a todos los usuarios de Spotify, recogen las 150 canciones más populares y exitosas de cada década. Además del título de la canción, estas listas proporcionan información adicional como el año de publicación, nombre del artista o algunas de sus características musicales específicas. Las listas "All Out..." son listas editoriales que se distinguen del resto por su gran número de seguidores

y la facilidad de acceso. De hecho, dos de las listas utilizadas en este trabajo, "All Out 80s" y "All Out 2000s", se encuentran entre las 10 listas editoriales con más seguidores (Leighton, 2023). La segunda fuente utilizada es Genius, un sitio web que funciona como una enciclopedia musical que contiene las letras de la mayoría de las canciones.

Dado que ambas son fuentes web y los datos se almacenan en las plataformas digitales de Spotify y Genius, para acceder a ellos de manera legal y legítima, se hace uso de sus propios Web Apis – Spotify API y Genius API. Estas APIs son interfaces de programas de aplicación que son utilizadas como un mecanismo de interconectividad que permite a dos componentes de software comunicarse entre sí a través de Internet (Sohan et al., 2015). Por lo tanto, a través de las respectivas APIs de Spotify y de Genius, las aplicaciones pueden comunicarse con los servidores correspondientes y proporcionarnos la información deseada. Sin embargo, para poder utilizar estas APIs, primero es necesario adquirir las credenciales, denominadas como “tokens”, que actúan como un mecanismo de autentificación. Para ello, hay que acudir a los servicios de cuenta de Spotify y de Genius, hacerse una cuenta y obtener así esa credencial. En ambos casos, la extracción y descarga de los metadatos de Spotify y las letras de las canciones de Genius, se realiza a través de R, en concreto del paquete “Spotifyr” y “Geniusr”.

4.2 Integración y preparación de los datos para conseguir la base de datos final

La base de datos final se construye fusionando dos conjuntos de datos: (1) la base de datos de métricas musicales obtenida a través de la API de Spotify (*Tabla 1)* y (2) la base de datos de letras de canciones obtenida mediante la API de Genius y técnicas de Web Scraping (*Tabla 2*).

Por un lado, el API de Spotify, y en concreto su función “get_playlist_audio_features”, proporciona 61 variables para cada una de las canciones que conforma la lista (Antal, 2022). Para acceder a estas playlists, simplemente se ha tenido que añadir su URL correspondiente. El estudio incluye siete listas de reproducción, una por cada década desde 1960 hasta 2020, con 150 canciones seleccionadas por década, dando un total de 1.050 canciones. De las 61 variables iniciales, muchas, como 'track.external_urls.spotify', 'video_thumbnail.url' y 'track.disc_number', han sido eliminadas para el análisis de este trabajo. Mientras que otras 22 variables, dada su importancia, han sido elegidas. Por tanto, el tamaño de la base de datos se ha reducido de

1050x61 a 1050x22. Cada canción está asociada a un identificador (ID) y a otras 21 variables, que incluyen el título de la canción, el nombre del artista y del álbum, la fecha de lanzamiento, la década correspondiente además de otras variables técnicas detalladas y explicadas en el Apéndice 1 (*Tabla 1*). Entre ellas, se encuentran las siguientes variables: “popularity”, “duration”, “loudness”, “danceability”, “energy”, “speechiness”, “acousticness”, “instrumentalidad”, “valence”, “mode” y “explicit”. Para agilizar el tratamiento de la información y para que las visualizaciones sean más concluyentes, se han revisado y modificado, en caso de necesidad, algunas de estas variables. La fecha de lanzamiento se ha ajustado para conservar solo el año de publicación, la duración de las canciones se ha convertido de segundos a minutos, la variable “explicit”, antes codificada como "True" o "False", ahora se representa como 1 si contiene letras explícitas e inapropiadas, y como 0 si no, la variable “mode”, antes codificada como “major” o “minor”, ahora se representa como 1 y 0 respectivamente. Y, por último, la variable de Sonoridad (“Loudness”) ha sido normalizada para asegurar que todos los valores estén en una misma escala entre 0 y 1, facilitando así su análisis comparativo.

Tabla 1. Extracto de la base de datos con las métricas musicales de las 1050 canciones

ID	Song title	Artist name	Album name	Decada	Year...6	Danceability	Energy
1	Brown Eyed Girl	Van Morrison	Blowin' Your Mind!	1960	1967	0.491	0.5830
2	Bad Moon Rising	Creedence Clearwater Revival	Green River (Expanded Edition)	1960	1969	0.508	0.7740
3	For What It's Worth	Buffalo Springfield	Buffalo Springfield	1960	1966	0.653	0.5190
4	Twist And Shout - Remastered 2009	The Beatles	Please Please Me (Remastered)	1960	1963	0.482	0.8490
5	Be My Baby	The Ronettes	Presenting the Fabulous Ronettes Featuring Veronica	1960	1964	0.512	0.7710
6	Mrs. Robinson - From "The Graduate" Soundtrack	Simon & Garfunkel	Bookends	1960	1968	0.606	0.4570
7	Somethin' Stupid	Frank Sinatra	The World We Knew	1960	1967	0.257	0.3380
8	I'd Rather Go Blind	Etta James	Tell Mama	1960	1968	0.477	0.4330
9	Oh, Pretty Woman	Roy Orbison	Oh, Pretty Woman	1960	1962	0.619	0.6030
10	Gimme Shelter	The Rolling Stones	Let It Bleed	1960	1969	0.634	0.6300
11	Happy Together	The Turtles	Happy Together	1960	1967	0.584	0.3670
12	People Are Strange	The Doors	Strange Days	1960	1967	0.699	0.4670
13	Respect	Aretha Franklin	I Never Loved a Man the Way I Love You	1960	1967	0.805	0.5580
14	For Once In My Life	Stevie Wonder	For Once In My Life	1960	1968	0.524	0.5190
15	I'm a Believer - 2006 Remaster	The Monkees	More of The Monkees (Deluxe Edition)	1960	1967	0.526	0.7750
16	Whole Lotta Love - 1990 Remaster	Led Zeppelin	Led Zeppelin II (1994 Remaster)	1960	1969	0.412	0.9020
17	Can't Take My Eyes off You	Frankie Valli	Solo	1960	1967	0.581	0.7800

Fuente: Elaboración propia a partir de los datos del estudio

Para completar nuestra base de datos, solo falta añadir una variable esencial: la letra de las 1050 canciones seleccionadas. Utilizando la función search_song del paquete de Geniusr se accede a la API y se obtiene el URL de cada una de las canciones (Henderson, 2022). Para obtener las letras de las 1050 canciones seleccionadas se realida el web scraping mediante la función extraer_texto. La función verifica la validez de cada URL, lee el contenido HTML de la página correspondiente (*Figura 5*) y extrae el texto

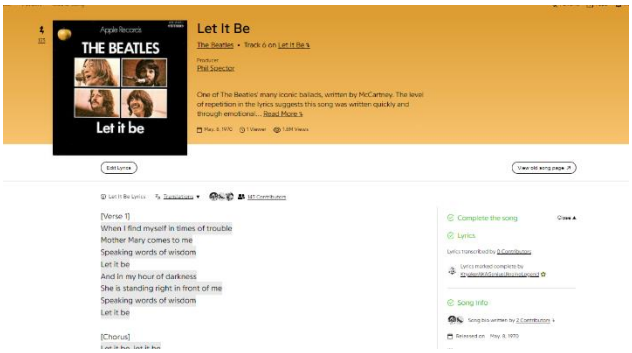
del nodo HTML especificado (usando la etiqueta “div. Lyrics__Container-sc-1ynbvzw-1.kUgSbL”). El texto extraído, que corresponde a las letras de las canciones, se almacena posteriormente en la variable corpus. La función utilizada es:

```
extraer_texto <- function(url) {
  if (is.na(url)) {
    return("")
  }

  contenido_html <- read_html(url)
  texto_extraido <- contenido_html %>%
    html_nodes("div.Lyrics__Container-sc-1ynbvzw-1.kUgSbL") %>%
    html_text()

  return(texto_extraido)
}
```

Figura 5. Extracto de la página web de Genius – Letra de la canción “Let it be” de The Beatles



Fuente: página web de Genius.com

De este modo, se genera una base de datos de 1050x7, aunque sólo se incluye la variable “corpus”, que contiene las letras de las canciones de las 150 canciones de cada década (*Tabla 2*). Antes de integrar la variable “corpus” con las demás obtenidas de Spotify, es crucial prepararla adecuadamente para análisis futuros. Los pasos de limpieza realizados son los siguientes:

- Se eliminan aquellos elementos que no añadan valor y que únicamente complican el tratamiento de datos, distorsionando así el resultado final: elementos gramaticales como los signos de puntuación, las interjecciones o expresiones (“huh”, “hah”, “eh”, etc.) y las denominadas “etiquetas de sección” del contexto musical (“Verso 1”, “Coro”, “Estribillo”, etc.)

- Se convierten las mayúsculas en minúsculas y se reemplazan todas las abreviaturas por sus palabras completas para que no haya pérdida de contenido ni significación (por ejemplo, reemplazar “goin’” por “going”).
- Para asegurar la integridad de la información, se traducen todas las canciones para que se encuentren en el mismo idioma, inglés. Por ello, dado la diversidad lingüística que surge en la década de los 2010 y 2020, a través del traductor “Deepl”, se realizan traducciones literales al inglés de letras como las de “Danza Kuduro”, “Titi me preguntó”, así como de las canciones del grupo coreano Blackpink.

Tabla 2. Extracto de la base de datos de las letras de las 1050 canciones

ID	Song title	Artist name	Decada	Year	URL_lyrics	corpus
1	Brown Eyed Girl	Van Morrison	1960	1967	https://genius.com/Van-morrison-brown-eyed-girl-lyrics	hey where did we go days when the rains came down in th...
2	Bad Moon Rising	Creedence Clearwater Revival	1960	1969	https://genius.com/Creedence-clearwater-revival-bad-moo...	see the bad moon arising see trouble on the way see earth...
3	For What It's Worth	Buffalo Springfield	1960	1966	https://genius.com/Buffalo-springfield-for-what-its-worth-ly...	there' something happening here but what it is aingt eactly ...
4	Twist And Shout - Remastered 2009	The Beatles	1960	1963	https://genius.com/The-beatles-twist-and-shout-lyrics	well shake it up baby now shake it up baby twist and sho...
5	Be My Baby	The Ronettes	1960	1964	https://genius.com/The-ronettes-be-my-baby-lyrics	the night we met knew needed you so and if had the chan...
6	Mrs. Robinson - From "The Graduate" Soundtrack	Simon & Garfunkel	1960	1968	https://genius.com/Simon-and-garfunkel-mrs-robinson-lyrics	and here' to you mrs robinson jesus loves you more than y...
7	Somethin' Stupid	Frank Sinatra	1960	1967	https://genius.com/Frank-sinatra-and-nancy-sinatra-someth...	know stand in line until you think you have the time to spe...
8	I'd Rather Go Blind	Etta James	1960	1968	https://genius.com/Etta-james-id-rather-go-blind-lyrics	something told me it was over when saw you and her talkin...
9	Oh, Pretty Woman	Roy Orbison	1960	1962	https://genius.com/Roy-orbison-oh-pretty-woman-lyrics	pretty woman walking down the street pretty woman the k...
10	Gimme Shelter	The Rolling Stones	1960	1969	https://genius.com/The-rolling-stones-gimme-shelter-lyrics	storm is threatening my very life today if don' get some she...
11	Happy Together	The Turtles	1960	1967	https://genius.com/The-turtles-happy-together-lyrics	imagine me and you do think about you day and night it'...
12	People Are Strange	The Doors	1960	1967	https://genius.com/The-doors-people-are-strange-lyrics	people are strange when you're stranger faces look ugly wh...

Fuente: Elaboración propia a partir de los datos del estudio

Tras preparar las dos bases de datos de manera independiente, finalmente se fusionan para crear la base de datos definitiva (*Tabla 3*). Esta incluye las 1.050 canciones más populares de las últimas siete décadas (150 canciones por década) y 23 variables, incluidas las métricas musicales y las letras de las canciones. El análisis se dividirá en dos partes: la primera examinará la evolución de las variables musicales de Spotify y la segunda se centrará en las letras de las canciones (variable "corpus"), sobre las que se aplicará o el modelo LDA para identificar tópicos.

Tabla 3. Estructura de la base de datos final que recoge las variables musicales y las letras de las 1050 canciones

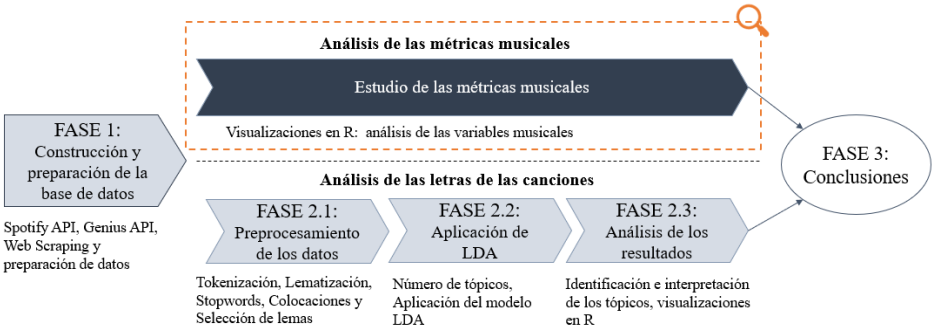
ID	Song title	Artist name	Album name	Decade	Year	Danceability	Energy	Loudness	Mode	Speechiness	Acousticness	Instrumentalness	Valence	Tempo	Mode name	Duration Ms	Explicit	Popularity	Release date	Duration	Loudness_Normalizada	corpus
1	Brown Eyed Girl	Van Morrison	Blowin' Your Mind!	1960	1967	0.491	0.5830	-10.364	1	0.0376	0.165000	0.00e+00	0.908	150.566	1	185306	0	85	1967-09	3.053100	0.4918597638	hey where did we go days when the rains came down...
2	Bad Moon Rising	Creedence Clearwater Revival	Green River (Expanded Edition)	1960	1969	0.508	0.7740	-5.973	1	0.0325	0.259000	4.67e-06	0.942	178.946	1	140600	0	84	1969-08-03	2.360000	0.8381717175	see the bad moon arising see trouble on the way see...
3	I've What It's Worth	Buffalo Springfield	Buffalo Springfield	1960	1966	0.653	0.5190	-10.164	1	0.0497	0.406000	2.09e-02	0.822	98.883	1	153800	0	79	1966-12-05	2.565550	0.6675177786	there something happening here but what it is aingt w...
4	Twist And Shout - Remastered 2009	The Beatles	Please Please Me (Remastered)	1960	1963	0.482	0.8490	-9.198	1	0.0452	0.845000	7.74e-06	0.937	124.637	1	155226	0	79	1963-03-22	2.587100	0.7787578869	well shake it up baby now shake it up baby twist an...
5	Be My Baby	The Ronettes	Presenting the Fabulous Ronettes Featuring Veronica	1960	1964	0.512	0.7710	-7.006	1	0.0402	0.156000	0.00e+00	0.803	129.699	1	160906	0	79	1964-11-01	2.681767	1.0000000000	the night we met knew needed you so and if had the...
6	Mrs. Robinson - From "The Graduate" Soundtrack	Simon & Garfunkel	Bookends	1960	1968	0.606	0.4570	-14.035	0	0.0497	0.713000	2.53e-05	0.813	102.033	0	244026	0	79	1968-04-03	4.067100	0.1849354651	and here to you mrs robinson jesus loves you more t...
7	Somethin' Stupid	Frank Sinatra	The World We Knew	1960	1967	0.257	0.3380	-12.902	1	0.0357	0.736000	0.00e+00	0.544	207.356	1	162493	0	79	1967-08	2.708217	0.3252869103	know stand in line until you think you have the time t...
8	I'd Rather Go Blind	Etta James	Tell Mama	1960	1968	0.477	0.4330	-6.473	0	0.0247	0.689000	1.95e-04	0.611	82.520	0	156853	0	78	1968-04-18	2.614083	0.9132115521	something told me it was over when saw you and her ...
9	Oh, Pretty Woman	Roy Orbison	Oh, Pretty Woman	1960	1962	0.619	0.6030	-9.481	1	0.0342	0.712000	0.00e+00	0.958	127.433	1	178933	0	77	1962	2.982217	0.5752423832	pretty woman walking down the street pretty woman ...
10	Gimme Shelter	The Rolling Stones	Let It Bleed	1960	1969	0.634	0.6300	-8.277	0	0.0310	0.447000	2.90e-02	0.489	118.628	0	270773	0	77	1969-12-05	4.512883	0.7332224164	storm is threatening my very life today if don' get som...
11	Happy Together	The Turtles	Happy Together	1960	1967	0.584	0.3670	-9.638	0	0.0328	0.550000	1.38e-05	0.588	120.575	0	176293	0	77	1967	2.938217	0.6091715925	imagine me and you do think about you day and nig...
12	People Are Strange	The Doors	Strange Days	1960	1967	0.609	0.4670	-8.578	0	0.0375	0.685000	0.00e+00	0.784	119.287	0	130173	0	76	1967-09-25	2.169550	0.7031789961	people are strange when you're stranger faces look ug...

Fuente: Elaboración propia a partir de los datos del estudio

Capítulo 5. Análisis de las características musicales / técnicas de las canciones identificadas por Spotify como las canciones más exitosas de su década

Con la base de datos construida, comienza la primera parte del análisis: las características técnicas / métricas musicales (*Figura 6*).

Figura 6. Procedimiento del Trabajo de Investigación: Análisis de las métricas musicales



***Fuente**: Elaboración propia*

A lo largo de estos años, la técnica y la estructura de las canciones han experimentado muchos cambios. Ni todas las canciones siguen una misma estructura técnica ni todas atribuyen la misma importancia a todas las variables. Los artistas han ido adaptando sus canciones a las necesidades de su público objetivo. Y, por consiguiente, y de manera predecible, las variables han evolucionado: algunas de las variables musicales

han perdido su importancia mientras que otras se han establecido como determinantes para el éxito de las canciones.

I. Estudio sobre la evolución de las variables musicales a lo largo de las últimas siete décadas (2016-2020)

A lo largo de las últimas décadas, las variables musicales han mostrado una alta volatilidad y múltiples fluctuaciones, sin seguir una tendencia estable y continua (*Figura 7*).

Considerando las tendencias generales y sin tener en cuenta las fluctuaciones de momentos específicos, las variables de Instrumentalidad ("Instrumentalness") y Locuacidad ("Speechniess") siempre han mantenido valores próximos a 0, indicando que los artistas, no los instrumentos, son los principales intérpretes y que las canciones no son exclusivamente habladas como los podcasts o entrevistas. En cambio, las variables de Bailabilidad ("Danceability"), Energía ("Energy"), Sonoridad ("Loudness") y Positividad ("Valence") siempre se han mantenido altas, con valores entre 0.5 y 1. Al fin y al cabo, la música fue creada originalmente para animar y entretener, por lo que es normal que siempre haya buscado ser bailable, enérgica, sonora y positiva. Desde 1960 hasta la fecha, la positividad ha disminuido, mientras que la energía ha aumentado. La bailabilidad y sonoridad de las canciones se han mantenido similares a los valores iniciales. A diferencia de otras variables, la acústica ha variado drásticamente, alcanzando valores medios de 0.6 en los años 60 y descendiendo a 0.1 en 2010, reflejando así los altibajos en la prevalencia de instrumentos acústicos en la música popular.

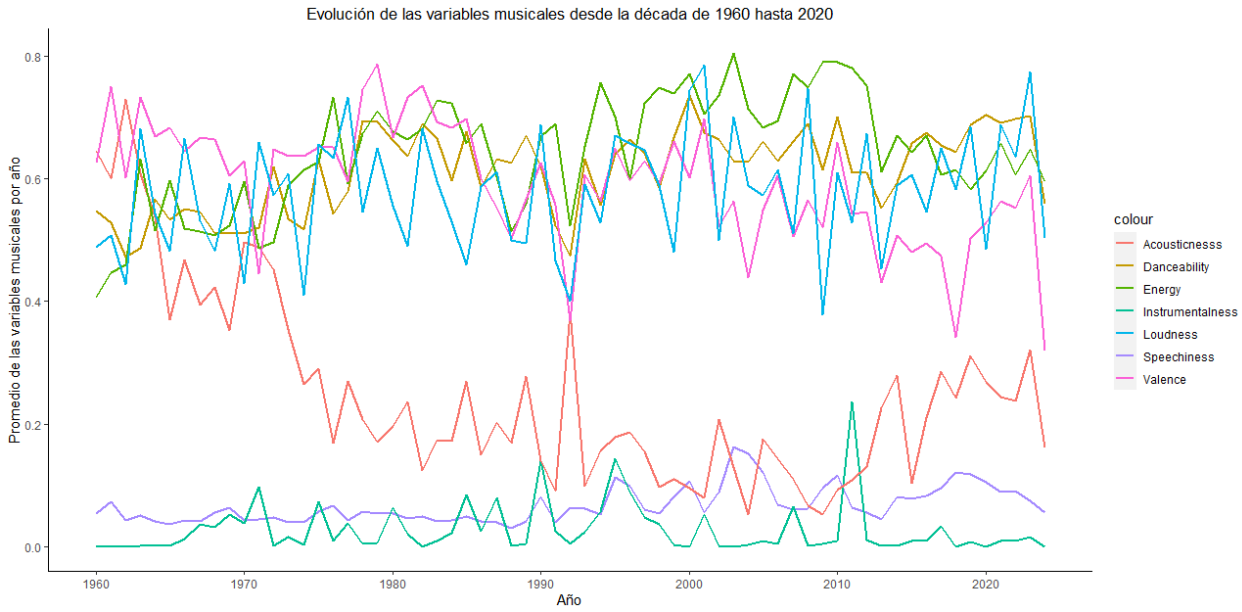
Un análisis más detallado de las variaciones ocurridas en años concretos nos va a permitir indagar sobre la posible relación entre la música y los diversos elementos de la sociedad.

- Los años 70 fueron años de gran diversidad musical. En las listas de éxitos se encuentran desde bandas icónicas de Hard Rock como The Rolling Stones o Led Zeppelin, hasta Pop suave con artistas como Elton John, e incluso el reggae con Bob Marley. Por lo general, este periodo se caracterizó por canciones enérgicas, bailables y cargadas de positividad. La música disco de los años 70, caracterizada por arreglos orquestales con instrumentos acústicos como cuerdas, vientos y metales, explica el pico observado ese año (Arizaga,

2023). Sin embargo, la diversidad musical provoca fluctuaciones pronunciadas que complican discernir tendencias claras en las variables.

- El panorama musical de los años 80 se caracterizó por el lanzamiento del canal de televisión MTV en EE. UU., en 1981. Se introdujeron los videopclips, un evento que marcó un antes y después en la forma de consumir la música. Esto dio lugar a canciones con alta sonoridad, optimistas y vibrantes para captar la atención del público (Tapia, 2021).
- Los años 90 destacan por la llegada de la música electrónica, y en España, por la popular Ruta del Bakalao. Fueron años de canciones enérgicas, ruidosas y vibrantes, y con un predominio de los instrumentos electrónicos. El pico de 1992 se debe a una peculiaridad de ese año, única en toda la década y es que cuatro de las ocho canciones que alcanzaron las listas de éxitos presentaban niveles acústicos medios entre 0.6 y 0.8. Como por ejemplo, "I Will always love you" de Whitney Houston o "Everybody hurts" de R.E.M.
- Desde la década de los 2000 hasta día de hoy, la sociedad se ha enfrentado a retos significativos como atentados, crisis económicas o pandemias. Durante estos años, variables como la energía y la bailabilidad se han mantenido elevadas, actuando como medio de escape ante la adversidad. Sin embargo, en las dos últimas décadas (2010 y 2020), aunque la bailabilidad se ha mantenido elevada, la energía ha disminuido y la sonoridad ha experimentado numerosas variaciones. En cambio, la positividad ha experimentado más fluctuaciones; mientras algunos artistas utilizaban sus canciones como forma de protesta y reivindicación, otros la utilizaban para evadirse de los problemas preferían temas alegres y positivos.

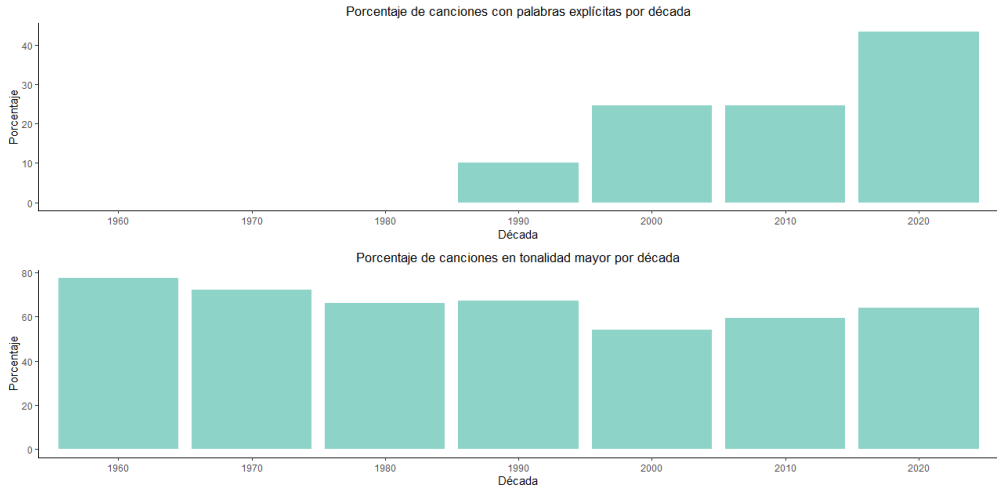
Figura 7. Evolución de las variables continuas musicales a lo largo de las últimas décadas



Fuente: Elaboración propia a partir de los datos del estudio

En cuanto a las variables dicotómicas (*Figura 8*), se observa de manera evidente cómo, a lo largo de los años, el porcentaje de canciones con contenido explícito, definido como el uso de palabras ofensivas e inapropiadas, se ha incrementado significativamente. mientras que, en los años 60, 70 y 80 este tipo de contenido era prácticamente inexistente, actualmente un 40% de las canciones incluye letras explícitas. Este fenómeno coincide con el auge de la música urbana, un género que se caracteriza por incorporar expresiones cotidianas, por su carácter reivindicador e incluso despectivo y por sus estilos propios (Úbeda, 2020). Con relación al tono de las canciones, se observa una preferencia por la tonalidad mayor, con al menos el 50% de las canciones de cada década compuestas en esta tonalidad.

Figura 8. Evolución en el porcentaje de canciones de cada década con contenido explícito y tono mayor



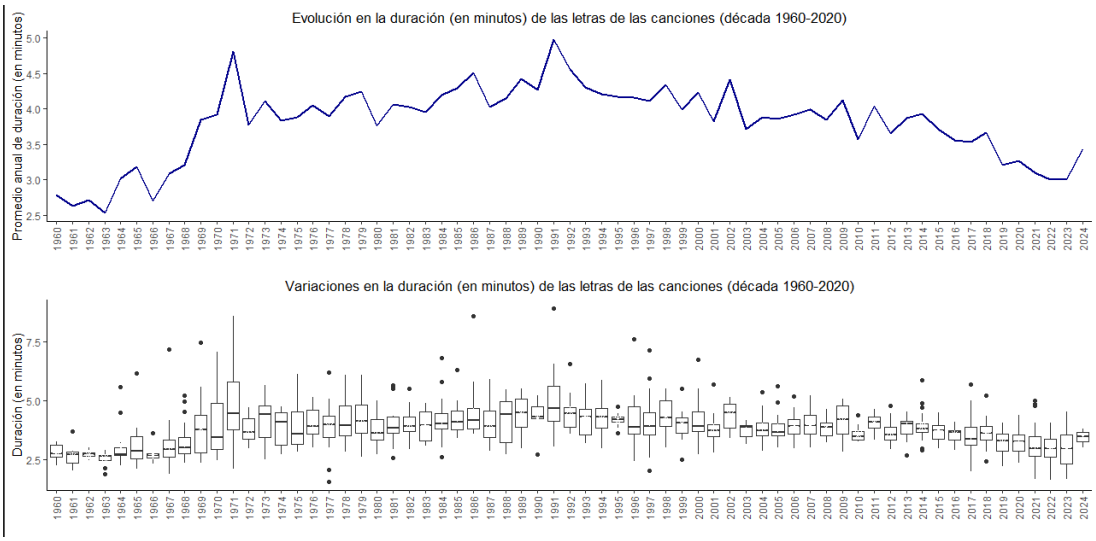
Fuente: Elaboración propia a partir de los datos del estudio

II. Estudios sobre la evolución en la duración de las canciones a lo largo de las últimas siete décadas

La duración de las canciones también ha ido variando a lo largo de las décadas (*Figura 9*). Al comienzo de los años 60, la duración media de las canciones no superaba los 3 minutos. Sin embargo, los artistas comenzaron a extender gradualmente la longitud de sus temas, alcanzando un promedio de 5 minutos por canción en 1971. De muestra de canciones exitosas que se lanzaron en 1971, aproximadamente la mitad de las canciones tienen una duración por encima de los 5 minutos. Desde entonces, la duración ha ido disminuyendo progresivamente, registrándose en 2023 canciones de aproximadamente 2,5 minutos de duración. El pico en la duración promedio de las canciones en 1991 se debe específicamente a "November Rain", una canción que tiene una duración de 8 minutos. Es decir, se trata de un “outlier” (se puede observar en el “box plot”), un caso aislado que no muestra un cambio movimiento general o de género musical. Actualmente vivimos en la era del “consumo rápido e instantáneo”, una época caracterizada por la decreciente capacidad de atención de los oyentes, su necesidad de conseguir respuestas inmediatas, su impaciencia y su facilidad para cambiar de estímulo. Un estudio de Microsoft confirma que el tiempo de atención y toma de decisiones del ser humano se ha reducido de 12 segundos (promedio de la década de los noventa) a 8 segundos (en 2015) y que se prevé que baje a 5 segundos en los próximos años (Gausby, 2015). Esta necesidad de obtener resultados “inmediatos” por parte de la sociedad, junto con las nuevas estrategias de facturación basadas en las reproducciones en las plataformas de streaming,

ha llevado a los artistas a reducir la duración de sus canciones casi en 1,5 minutos en comparación con 1992 y a adelantar el estribillo. Así disminuyendo el riesgo de despiste y ganando rentabilidad (Piera, 2023). Esta tendencia también se observa en redes sociales como Instagram y TikTok, donde los vídeos tienden a ser cada vez más cortos por la dificultad de retener la atención de los oyentes.

Figura 9. Evolución en la duración (en minutos) de las canciones desde la década de 1960 hasta 2020



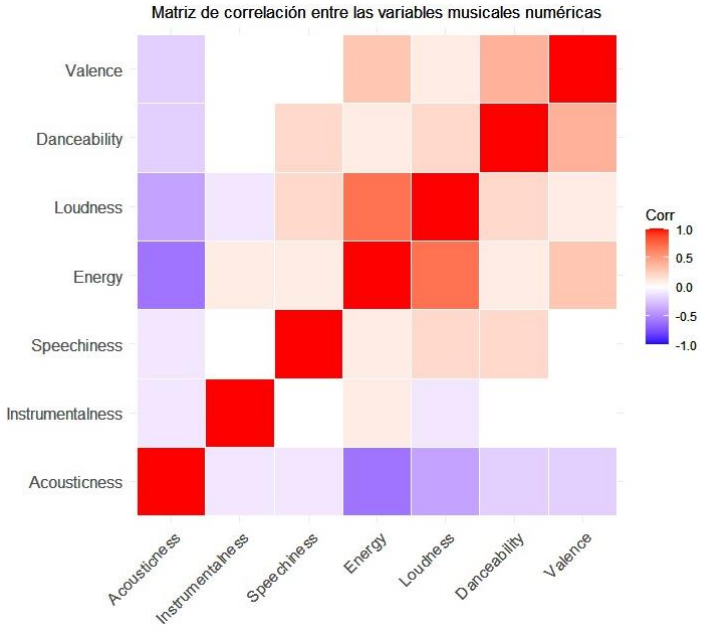
Fuente: Elaboración propia a partir de los datos del estudio

III. Estudio de las relaciones existentes entre las variables musicales a través de una matriz de correlación y gráficas de dispersión

Más allá de la variabilidad de las características musicales a lo largo de estos años, se ha querido completar y reforzar el estudio investigando la relación entre las variables musicales. Para realizar este análisis, se ha utilizado una matriz de correlación, centrando la atención en las variables numéricas con rango [0,1] y prescindiendo de las variables "Explicit" y "Mode". A raíz de los coeficientes que se obtienen como resultado de la matriz de correlación (*Figura 10*), se puede entender la relación que existe entre las variables. De las gráficas se pueden sacar dos conclusiones: si los valores se aproximan a -1, indica una correlación negativa y toma un color azul, mostrando que las dos variables se relacionan de manera inversa y se mueven en direcciones opuestas; si los valores se acercan a 1, adoptan un tono más rojo y señalan una correlación positiva, donde las variables se mueven en la misma dirección. Si la matriz muestra un colo blanco, significa que no hay relación alguna como es el caso de Instrumentalidad y Speechniness o

Speechiness con Valence. En pocas palabras, cuanto más cerca a 1 o a -1, más fuerte es la relación lineal entre las variables.

Figura 10. Matriz de correlación entre las variables numéricas



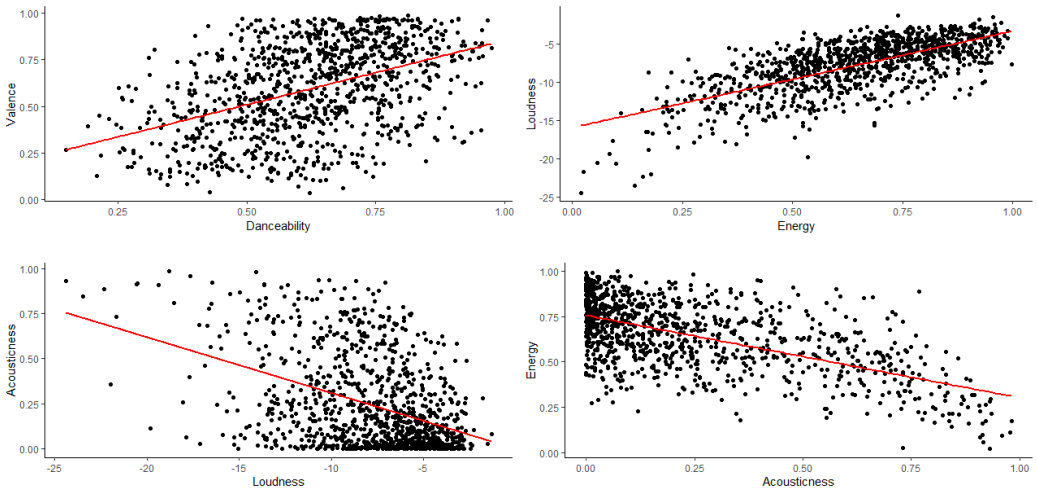
Fuente: Elaboración propia a partir de los datos del estudio

La matriz de correlación revela que la mayoría de las variables están relacionadas, ya sea positiva o negativamente. Algunas de ellas incluso muestran una correlación fuerte, como en los siguientes casos *(Figura 11):*

- La variable sonoridad (“Loudness”) presenta una fuerte correlación positiva con la variable energía (“Energy”), indicando que ambas tienden a moverse en la misma dirección; es decir, una mayor sonoridad en una canción suele estar acompañada de más energía, densidad y rapidez. Asimismo, las variables Positividad ("Valence") y Bailabilidad ("Danceability") también muestran una relación positiva, sugiriendo que las canciones más bailables tienden a ser positivas y alegres.

- En cambio, la variable acústica (“Accousticness”) tiene una relación negativa fuerte tanto con la variable de Sonoridad (“Loudness”) como Energía (“Energy”), es decir, se mueven en direcciones opuestos. Esto tiene su sentido pues normalmente las canciones más acústicas, suelen ser canciones tranquilas, con ritmo lento y por consiguiente, con bajo nivel de energía y de volumen.

Figura 11. Gráficos de dispersión para reflejar la correlación entre variables concretas



Fuente: Elaboración propia a partir de los datos del estudio

IV. Relación entre la popularidad de las canciones y las variables musicales ¿Qué variables favorecen una mayor popularidad para la canción?

Una vez establecido que las variables musicales tienen algún tipo de relación, se continúa la investigación explorando si existen factores y características musicales que contribuyen al éxito de una canción en las listas principales de Spotify. Partiendo de la base que nuestra base de datos está formada por canciones con una popularidad ya mayor del 60%, se han creado dos grupos extremos utilizando el promedio de esta variable de popularidad. El grupo superior incluye 129 canciones cuyo valor de popularidad está por encima de la media más una desviación típica, mientras que el grupo inferior incluye 112 canciones cuyo valor de popularidad está por debajo de la media menos una desviación típica. El objetivo de esta tabla descriptiva (*Figura 12*) es comparar el comportamiento medio de cada una de las variables en esos dos grupos extremos, viendo si hay diferencias notorias en alguna variable entre el grupo de mayor popularidad y el de menor popularidad, para así identificar aquellas variables que podrían estar más relacionadas con el éxito. Para las variables dicotómicas "Mode" e "Explicit", su media se corresponde con la proporción de canciones con tonalidad mayor y con contenido explícito, respectivamente.

Comparando las medias de cada una de las variables entre ambos grupos y con el promedio total, se observa que las canciones más populares destacan por ser enérgicas, con alta sonoridad y contenido explícito. En estas variables, la media del grupo superior es mayor que el promedio total. Sobre todo, resalta la variable de energía y sonoridad, que difieren significativamente de los valores del grupo de menor popularidad. Además, se distinguen por su tonalidad menor y poca instrumentalidad (incluidos los instrumentos acústicos) dando más importancia a las voces.

Por ello, estas diferencias notables en la media de las variables sugieren que variables como la energía, el volumen y el contenido explícito pueden influir más que otras en el éxito de una canción y estar relacionadas con la popularidad. No obstante, resulta difícil determinar qué características van a garantizar al cien por cien el éxito de una canción y su entrada en las listas de éxito de Spotify ya que entran en juego las preferencias del oyente en ese momento específico. Utilizando una metodología estadística inferencial, se podría profundizar en este análisis.

Figura 12. Tabla descriptiva comparando el comportamiento medio de cada de las variables entre el grupo con popularidad más alta y baja

	All	Popularidad > Popularidad media + 1 desv.	Popularidad < Popularidad media - 1 desv.
Sample	1050	129	112
Variable musical	**Media total**	**Media del grupo superior**	**Media del grupo inferior**
Danceability	0,624	0,608	0,547
Energy	0,646	0,653	0,499
Loudness	0,591	0,604	0,506
Mode	0,657	0,628	0,741
Speechiness	0,069	0,066	0,053
Acousticness	0,242	0,242	0,476
Valence	0,594	0,498	0,648
Instrumentalness	0,024	0,012	0,012
Duration	3,764	3,686	3,088
Explicit	0,147	0,194	0,045

Fuente: Elaboración propia a partir de los datos del estudio

V. Estudio sobre el sexo de los artistas que han alcanzado las listas de éxitos de cada década

Este último análisis busca examinar la evolución de la diversidad en términos de sexo y composición musical entre los artistas que han alcanzado una posición destacada en las

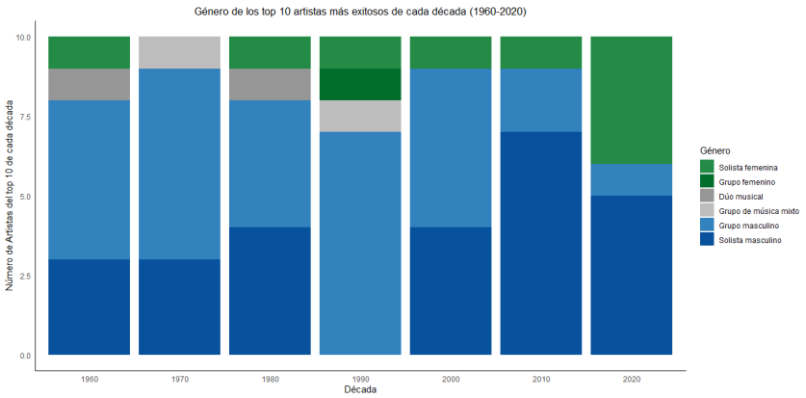
listas de éxito de Spotify. Para ello, se han filtrado las 150 canciones de cada década por la variable de popularidad, identificando a los artistas de las 10 canciones más reconocidas de cada período. Dado que el estudio se limita a los diez artistas principales de cada década, los resultados no son definitivos y deben interpretarse con precaución, aunque proporcionan indicativos valiosos. A partir de esta gráfica, se pueden obtener varias conclusiones.

A simple vista, es evidente que la industria musical ha estado históricamente dominada por el género masculino (*Figura 13*). Tanto en formatos solistas como en grupos, desde los años 1960 hasta 2010, más del 75% de los artistas en el top 10 han sido hombres. Sin embargo, esta tendencia ha comenzado a cambiar debido al creciente empoderamiento femenino. A lo largo de los años, las mujeres han ido ganando presencia y reconocimiento en la industria musical; por ejemplo, en los años 90 solo una artista femenina figuraba en el ranking, mientras que en la actualidad casi el 50% de los artistas en el top son mujeres.

Los grupos musicales han tenido un papel fundamental en cada década, ocupando aproximadamente la mitad de los lugares en los rankings. Destacablemente, en los años 90, solo una artista femenina alcanzó el top 10, mientras que el resto eran grupos. Sin embargo, desde esa década, la influencia de las bandas ha disminuido, mientras que la presencia de artistas solistas ha aumentado. En la actualidad, el 90% de los artistas en los rankings son solistas, distribuidos casi equitativamente entre hombres y mujeres.

Al comparar la nacionalidad de los artistas, se observa una mayor diversidad en las últimas décadas. En las décadas 2010 y 2020, la globalización y las nuevas tecnologías han facilitado la aparición de artistas de diversas nacionalidades, distintas a la estadounidense, que ha sido predominante hasta ahora. En los últimos años, artistas como Tom Odell (británico), Lewis Capaldi (escocés), The Weeknd (canadiense), Shakira (venezolana) y Jung Hook (surcoreano) han alcanzado reconocimiento mundial con canciones muy populares. Esto explica la mayor variedad en los idiomas de las canciones de las últimas décadas.

Figura 13. Evolución en el sexo y tipo de agrupación musical de los top 10 artistas de cada década

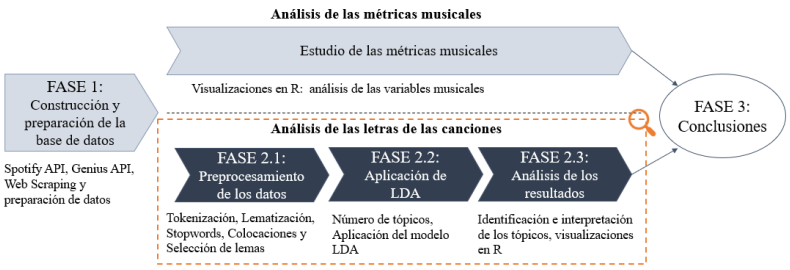


Fuente: Elaboración propia a partir de los datos del estudio

Capítulo 6. Análisis de las letras de las canciones a través de la técnica de Topic Modeling utilizando Latent Dirichlet Allocation (LDA)

Tras analizar las características musicales, se procede a la segunda parte del estudio: el análisis de las letras de las canciones. De la base de datos final de dimensión 1050x23, para este análisis concreto, se utiliza sólo la última variable, “corpus”, que recoge las letras de las canciones preparadas para el estudio, y el año y década de las canciones omitiendo así el resto de las variables. Este capítulo explora en profundidad el "corpus", aplicando la técnica de Topic Modeling (LDA) para identificar temas relevantes y analizar su evolución a lo largo de las décadas. Esta parte del análisis de letras de las canciones (fase 2.1, 2.2 y 2.3 de la figura 14), se ha basado en el procedimiento presentado en Velilla (2023).

Figura 14. Procedimiento del Trabajo de Investigación: Análisis de las letras de las canciones



Fuente: Elaboración propia

6.1 Aplicación del modelo LDA

El modelo LDA se aplica a las letras de las canciones para la identificación de tópicos y la prevalencia de los mismos. Para llevar a cabo el análisis es necesario realizar un preprocesamiento de los datos antes de aplicar el algoritmo LDA, tal y como se observa en la figura 14 (fases 2.1 y 2.2).

Fase 2.1. Preprocesamiento de los datos

Para aplicar el modelo LDA, primero se deben procesar los datos de texto, que, en nuestro trabajo, se corresponde con la variable "corpus". Se trata de una fase crucial, pues todas las palabras que se recogen en la base de datos conforman el "bag of words" que es utilizado en el modelo LDA. Esta fase tiene como objetivo eliminar todas aquellas palabras que son irrelevantes dando mayor valor e importancia a las palabras más significativas (Trenquier, 2018).

En primer lugar, es necesario someter las letras de las canciones al proceso de tokenización que consiste en delimitar las palabras del texto para convertirlas en elementos de una lista. Para nuestro estudio, cada canción conforma un documento, lo que da un total de 1050 documentos, y cada documento se fragmenta en tokens, unigramas totalmente independientes sin considerar el orden y las posibles relaciones contextuales. Esto da lugar a lo que se conoce como enfoque de "Bag of Wards" (Maier at al., 2018). Para el proceso de tokenización, se utiliza la función "udpipe_annotate", que devuelve un dataframe con todos los tokens de las 1.050 canciones, un total de aproximadamente 372.000 tokens, asociados al documento al que pertenece, a su categoría gramatical y a su lema. Gracias al paquete "udpipe", se realiza el proceso de Part-of-Speech Tagging sobre el corpus, el cual identifica la categoría gramatical de cada elemento léxico (Schweinberger, 2023). Esto nos permite conocer la naturaleza gramatical de los tokens, fundamental para el posterior preprocesamiento de los datos. Además, la función "udpipe_annotate", basándose en la categoría gramatical identificada, devuelve el lema de cada token, resolviendo así el problema de que una misma palabra tenga múltiples formas de representación. Al final, los lemas son la forma base de las palabras en el diccionario, y la categoría gramatical ayuda a encontrar la entrada correcta en el mismo y, por tanto, a identificar la correcta raíz de cada token (Porras, s.f.).

Una vez tokenizado y obtenido el lema de cada uno de los tokens, se realiza un filtrado para eliminar los tokens y lemas que no aportan valor y que son irrelevantes ya sea por su poca frecuencia en el texto o, por el contrario, por su presencia de manera generalizada (Maier et al., 2018). Para el preprocesamiento de los datos, se realizan los siguientes pasos:

1. Se filtran los tokens y lemas por su categoría gramatical, eliminando ciertas categorías como pronombres, verbos, interjecciones, conjunciones, artículos, números, símbolos, entre otros. Por tanto, manteniendo únicamente para el análisis los sustantivos, adjetivos y adverbios.

2. Se eliminan los stopwords y se incorpora una lista manual de palabras que son frecuentes en el contexto musical pero que no aportan valor al análisis. Por ejemplo, se eliminan palabras como "yeah", "oh", "baby", así como términos como “music” o “song” que al ser tan comunes pierden importancia en el texto.

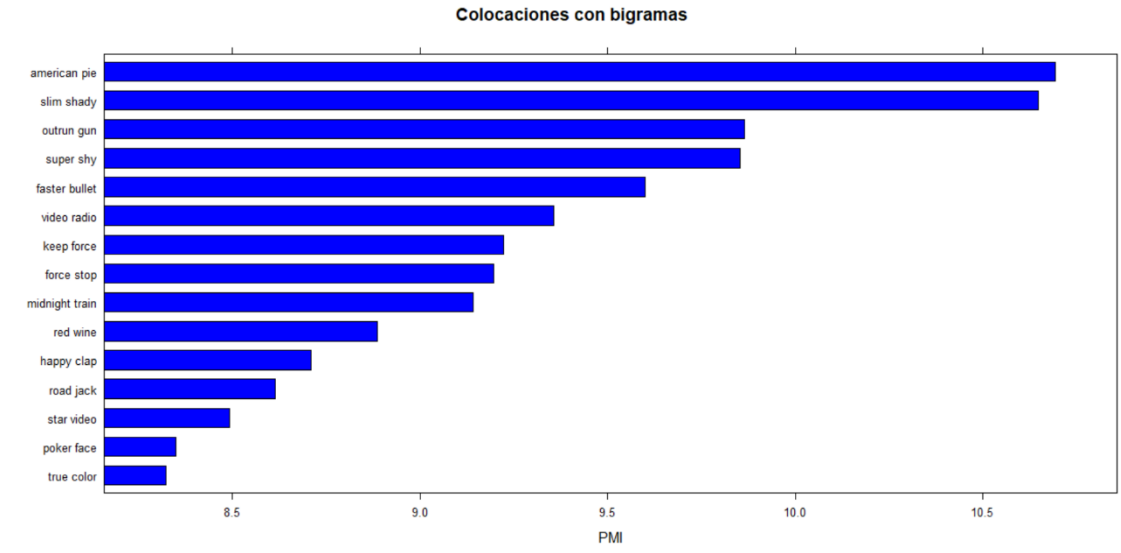
3. Se filtran los lemas para trabajar únicamente con aquellos que tengan una longitud igual o superior a 3 caracteres asegurándonos así de estar utilizando lemas con significados relevantes. Se consideran los lemas de 3 caracteres porque, aunque algunos, como "til" o “bit”, aportan poco valor, hay otros, como "boy", "now", "kiss" o “gun” que no podemos eliminar.

4. Como paso final del preprocesamiento de datos, es crucial reconocer que ciertas palabras tienden a estar intrínsecamente relacionadas y aparecen juntas de manera consistente en el texto. El análisis de colocaciones y bigramas implica estudiar las posibles relaciones semánticas entre cada palabra (en nuestro caso, los lemas) y aquellas que la preceden o siguen en el contexto. A la hora de clasificar las palabras, es importante considerar no solo su significado, sino también su co-ocurrencia con otras palabras (Church y Hanks, 1990). Para evaluar la relación interna que existe entre los lemas, se utilizan las métricas de asociación y en este caso concreto, la métrica de PMI (“Pointwise mutual information”). Esta compara la frecuencia de las palabras "x" e "y" cuando aparecen juntas en el corpus con su frecuencia esperada si fueran tratadas como palabras independientes (Petrovic et al., 2006). En conformidad con investigaciones previas, y siguiendo el enfoque presentado en Velilla (2023), un PMI superior a 3 es suficiente para considerar una colocación como significativa. En nuestro estudio, se seleccionan

las colocaciones con un PMI mínimo de 3, para no restar valor a ninguna de ellas. Además, se filtran las colocaciones que consisten en la repetición de las mismas palabras, una característica común en las canciones pero que no añade valor significativo.

Analizando las colocaciones con mayor PMI (*Figura 15*), se observa que muchas de ellas coinciden con los títulos de algunas de las canciones que fueron éxitos en su década, como es el caso de "American Pie", "Slim Shady", "Midnight Train", "Road Jack" o "Poker Face". Los artistas tienden a titular sus canciones con las palabras más repetitivas de sus letras, especialmente de sus estribillos, para facilitar así su identificación, recuerdo y posterior búsqueda en las plataformas.

Figura 15. Análisis de las colocaciones – las 15 primeras colocaciones con mayor PMI



Fuente: Elaboración propia a partir de los datos del estudio y adaptado del código propuesto en Velilla (2023)

Tras el preprocesamiento de datos y el estudio de las colocaciones, se crea una nueva variable, denominada "term", que recoge todos los lemas preprocesados de las categorías gramaticales sustantivo, adjetivo y adverbio, así como las colocaciones con un PMI superior a 3, fusionándose en un mismo término para evitar duplicaciones y conseguir un estudio más preciso. Esto se puede ver reflejado en la Tabla 4, en el término "road jack" que surge como fusión de dos sustantivos: "road" y "jack". Para la aplicación del modelo LDA, ya no se habla de tokens o lemas, sino de términos, haciendo referencia a la variable "term".

Tabla 4. Extracto del "corpus limpio tokenizado" una vez preprocesado los datos y con las colocaciones incluidas – nueva variable "term"

doc_id	paragraph_id	sentence_id	sentence	token	lemma	upos	term
doc38	1	1	hit the road jack and don' ya come back no more no more n...	money	money	NOUN	money
doc38	1	1	hit the road jack and don' ya come back no more no more n...	good	good	ADJ	good
doc38	1	1	hit the road jack and don' ya come back no more no more n...	well	well	ADJ	well
doc38	1	1	hit the road jack and don' ya come back no more no more n...	guess	guess	NOUN	guess
doc38	1	1	hit the road jack and don' ya come back no more no more n...	right	right	ADV	right
doc38	1	1	hit the road jack and don' ya come back no more no more n...	road	road	NOUN	road jack
doc38	1	1	hit the road jack and don' ya come back no more no more n...	jack	jack	NOUN	*NA*

Fuente: Elaboración propia a partir de los datos del estudio

Fase 2.2. Determinación del número de tópicos y aplicación del modelo LDA

Para la aplicación de la técnica de Topic Modeling (LDA) se utiliza el paquete "topicmodels" y la función LDA que contiene los siguientes argumentos (Grun y Hornik, 2023).

```
modelo <- LDA(dtm, method = "Gibbs", k =i, control = list(alpha = 50/i, delta=0.1, seed=58),
initialize="random")c
```

Entrando más en detalle en cada uno de los argumentos:

Por un lado, el "Document Term Matrix" (DTM). Este se refiere a la matriz que recoge la frecuencia con la que los términos del corpus aparecen en el conjunto de documentos (Nguyen, 2014). Se ejecuta la función "dfm-trim" de la librería "quanteda" sobre el DTM para eliminar los términos que aparecen en menos del 0,5% de los documentos o en más del 99% de los documentos al ser considerados irrelevantes y ruidoso (Maier et al., 2018). De esta forma, la matriz acaba recogiendo todos los términos importantes, procesados y preparados para implementación del modelo LDA.

Por otro lado, el argumento "método" que define el método utilizado para la aplicación del modelo que en nuestro caso es "Gribbs Sampling", el argumento "control" que permite establecer los hiperparámetros así como la semilla para asegurar que cada ejecución del código genera resultados consistentes y el argumento "initializie" que hace alusión al método que se utiliza para la primera asignación de términos. Siguiendo el

enfoque sugerido por Velilla (2023), se selecciona una asignación inicial aleatoria, conocida como "random" en R.

Y, por último, el número de tópicos. Como ya se ha comentado en el apartado 3.2, el modelo LDA no decide el número de tópicos a extraer, sino que es un dato que se debe introducir antes de ejecutar el algoritmo. La decisión sobre el número de tópicos tiene un papel fundamental en el desarrollo de la investigación; un error en el número de temas puede conducir a problemas de sobreajuste o infra ajuste en el modelo. Mientras que el sobreajuste aparece cuando el número de temas seleccionados es excesivamente alto, el infra ajuste cuando el número es muy bajo (Niekler y Wiedemann, 2017). Ambas son igualmente problemáticas. El sobreajuste puede conducir a un análisis del contenido semántico excesivamente detallado en el que, al dividir el contenido en tantos bloques, hay muchos solapamientos y similitudes entre lo que dificultan cualquier interpretación de resultados (Niekler y Wiedemann, 2017). Por otro lado, en el escenario del infra ajuste, ocurre todo lo contrario. El análisis puede ser tan general que existe el riesgo de perder información relevante.

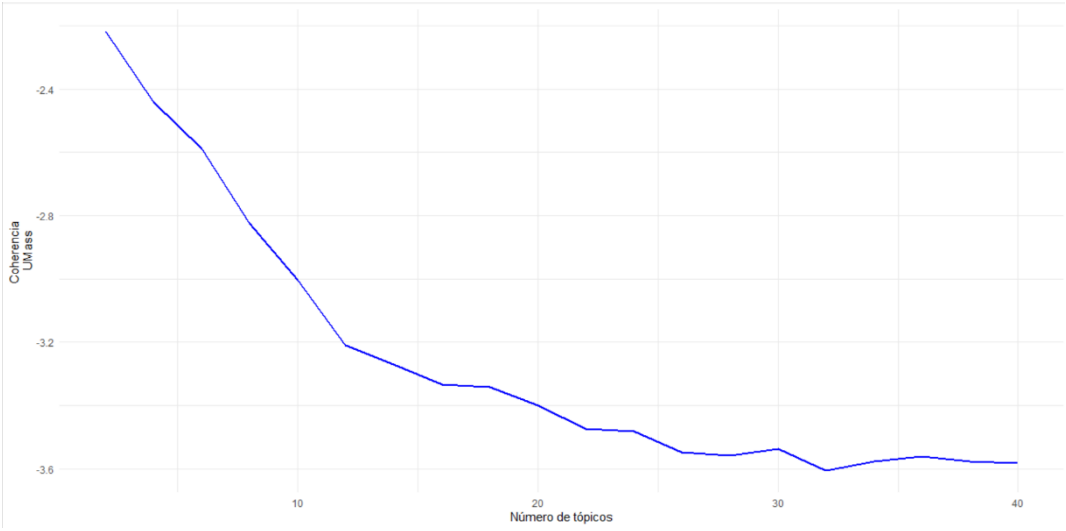
Con el propósito de evitar alguno de estos dos problemas, se utilizan las métricas de coherencia para determinar el número óptimo de tópicos a extraer. Dado que las medidas de coherencia evalúan las similitudes semánticas que existen de cada tópico, es posible diferenciar entre aquellos temas que resultan interpretables y aquellos, que según Stevens et al. (2012), son “artefactos de inferencia estadística”. En otras palabras, a mayor similitud semántica dentro de un tópico, mayor coherencia e interpretabilidad. Por esta razón, a la hora de seleccionar el número de tópicos se busca aquel que genera la mejor coherencia del modelo con todos sus tópicos. Para nuestra investigación, se utiliza la métrica de UMass, creada por Mimno et al. (2011), que utiliza el corpus original en lugar de un corpus externo (es decir, datos no utilizados durante el Topic Modeling como por ejemplo Wikipedia). La ecuación considera tanto la frecuencia de que dos palabras, “i” y “j”, aparezcan en un mismo documento, reflejado como D(vi, vj), como la frecuencia de que una de ellas (“j”) aparezca en los documentos, D(vj) (Stevens et al., 2012) y se expresa de esta manera:

$$score(v_i, v_j, \epsilon) = \log \frac{D(v_i, v_j) + \epsilon}{D(v_j)}$$

Cuanto mayor sea la frecuencia de que dos palabras aparezcan juntas y menor la frecuencia de que una de ellas se encuentre en el documento, mayor será el nivel de coherencia, traduciéndose en valores más próximo a cero por el logaritmo. Sin embargo, es importante tener en cuenta que, al ser una probabilidad condicionada, el numerador D (vi, vj) no puede ser mayor que el denominador D(vj) (Rosner et al, 2014). En función de los resultados devueltos por la métrica, se determina el número adecuado de tópicos a utilizar.

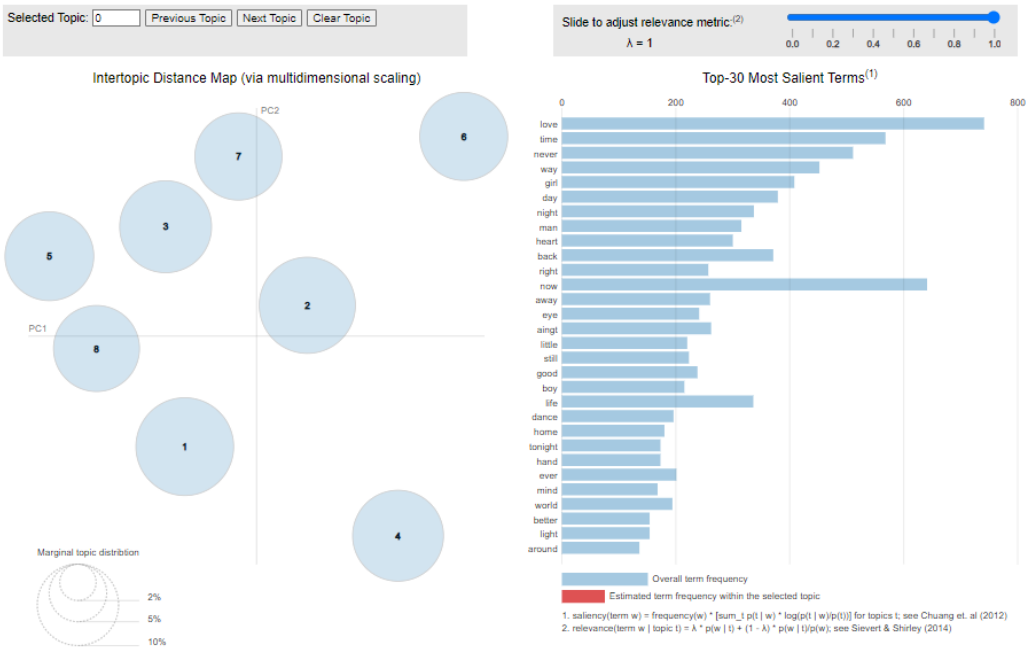
En nuestro caso concreto para evaluar la coherencia del modelo, el algoritmo LDA se ejecuta cinco veces para cada valor de K (número de tópicos), que varía de 2 a 40. Se calcula la media de coherencia de estas cinco ejecuciones para cada valor K y luego se evalúan y comparan todas las coherencias conseguidas y se determina el número óptimo de tópicos, que es aquel con el que se consigue una mayor coherencia, es decir, una coherencia más próxima a cero (Grun y Hornik, 2023; Velilla, 2023). Observando la figura 16, los posibles escenarios a estudiar son 4, 6 y 8 pues son los que presentan una pequeña alteración, y a partir de los cuales los valores disminuyen mucho. Al determinar el número de tópicos, no sólo debe considerarse el resultado generado por la métrica de coherencia, sino que también es fundamental hacer uso del juicio propio. Aunque una menor coherencia puede ser indicativa del número óptimo de tópicos, es esencial asegurarse de que tenga sentido interpretativo. En función del "Intertopic Distance Map" (*Figura 17*), se determina que el número óptimo de tópicos es 8, debido a su proximidad a cero en coherencia y por la ausencia de solapamientos entre tópicos. Además, ocho tópicos proporcionan un número razonable que facilita la interpretación de los resultados. El escenario de 6 tópicos presenta solapamientos entre temas, y el escenario de 4 tópicos, pese a no tener solapamientos, se considera un número reducido que puede limitar la interpretación del corpus. Este "Intertopic Distance Map" es un mapa de distancias que traza los círculos utilizando un algoritmo multidimensional y que permite visualizar los temas en un espacio bidimensional mostrando con el tamaño de los círculos el número de palabras que recoge cada tópico y con la distancia entre círculos el grado de en qué se comparten ciertas palabras (Sievert y Shirley, 2014).

Figura 16. Métrica de coherencia UMass para la selección del número óptimo de tópicos



Fuente: Elaboración propia a partir de los datos del estudio y adaptado del código propuesto en Velilla (2023)

Figura 17. El "Intertopic Distance Map" para el escenario de 8 tópicos

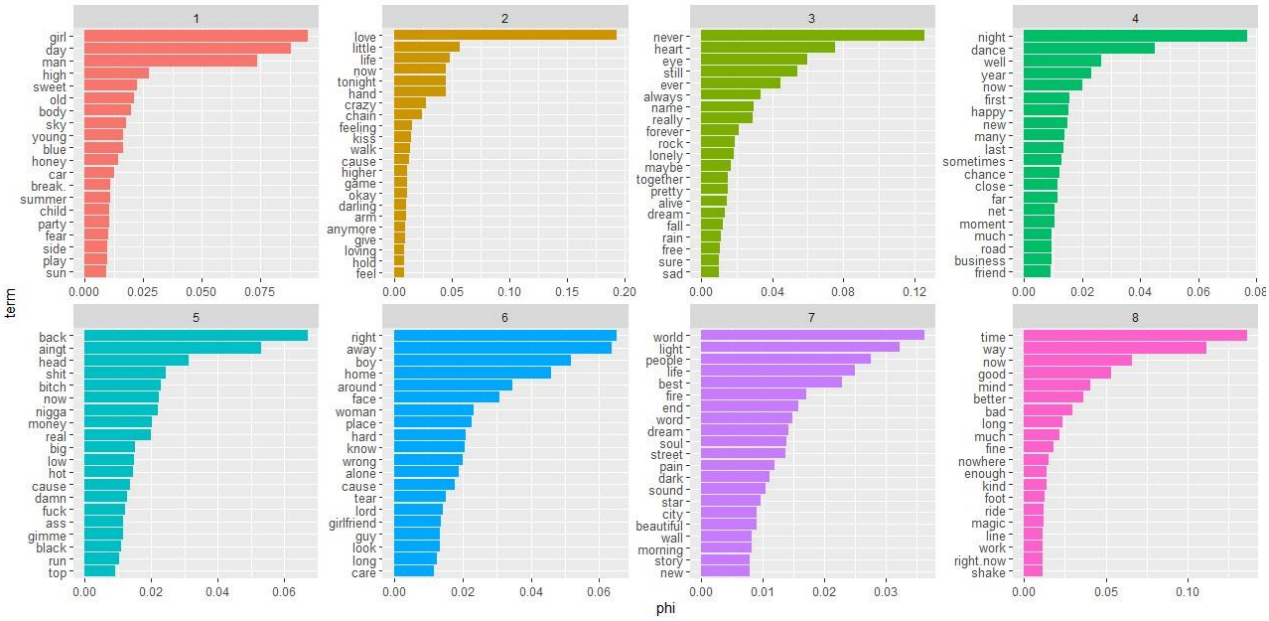


Fuente: Elaboración propia a partir de los datos del estudio y adaptado del código propuesto en Velilla (2023)

6.2 Identificación de los tópicos

El modelo LDA genera una lista con los 8 tópicos extraídos del corpus, cada uno compuesto por los términos más significativos. Ya en la fase 2.3 de la implementación del algoritmo, dado que el modelo no asigna nombres a los tópicos, somos nosotros quienes, basándonos en los términos que los constituyen, debemos asignarles un nombre que refleje adecuadamente su temática. Para seleccionar el nombre más apropiado, decidimos estudiar en detalle los 20 términos más representativos de cada tópico, es decir, los 20 términos con mayor peso en la distribución de palabras por tópico (*Figura 18*)

Figura 18. Peso (representado por las barras) de los 20 términos más representativos de cada tópico



Fuente: Elaboración propia a partir de los datos del estudio y adaptado del código propuesto en Velilla (2023)

La interpretación de los tópicos es la siguiente:

Tópico 1 – "Juventud y Libertad": Este tópico habla de la inocencia, de la libertad y del disfrute de los jóvenes ("girl", "man", "child", "young", "party"). Captura las experiencias diarias de los jóvenes desde lo cotidiano ("sky", "body", "car", "summer")

hasta a los altibajos emocionales (“sweet”, “break”) que enfrentan, con sus momentos dulces y desafíos.

Tópico 2 – “Amor Apasionado”: Describe un amor intenso y apasionado ("love", "kiss", "feeling", “loving”, “feel”, “darling”), que resalta las relaciones profundas y los intensos conflictos emocionales. Para algunos el amor puede llegar a ser divertido y alocado (“crazy”, “higher”) mientras que para otros puede ser considerado un juego o una condena (“game”, “chain”).

Tópico 3 – “Perseverancia y Lucha”: Aborda la resistencia personal, destacando la perseverancia y el esfuerzo colectivo (“still”, “always”, “together”), junto con la autorreflexión para identificar y superar adversidades. Mientras que términos como “forever”, “never” y “dream” hacen alusión a no rendirse y a ser constantes en el camino por alcanzar los objetivos, "rock", "alive" y “free” refuerzan la idea de fortaleza y vitalidad en medio de las circunstancias desafiantes.

Tópico 4 – “Diversión y Celebración”: Este tópico recoge el sentido de la diversión, de la felicidad (“happy”, “well”), de la festividad y de la apertura a nuevas oportunidades. Términos como "night", "dance" y “friend” indican celebraciones o eventos festivos, mientras que "new", "chance", "many" o “first” sugieren nuevos comienzos y posibilidades.

Tópico 5 – “Cruda Realidad Urbana”: Este tópico refleja la realidad de algunos entornos en los que se utilizan expresiones fuertes, de estilo directo y sin filtros, típico del lenguaje callejero y común en las letras de rap y hip-hop que abordan temas urbanos y desafíos sociales (“bitch”, “damn”, “fuck”, “shit”, “ass”). Incluso, menciona el creciente problema de discriminación en diversos contextos urbanos actuales (“nigga”, “money”, “black”).

Tópico 6 – “Inquietudes del día a día”: Aborda las complejidades y las preocupaciones cotidianas de las relaciones personales y sociales, así como la búsqueda de identidad y sentido de pertenencia. Refleja ese camino individual que hace cada uno por encontrar su lugar en el mundo ("home", "place", “alone”) y definir su identidad ("boy", "woman", “guy”), enfrentando desafíos, decisiones difíciles, y situaciones dolorosas ("right", "wrong", "hard", “tear”).

Tópico 7 – “Sueños y Aspiraciones”: Aborda los sueños, objetivos y aspiraciones que mantienen al ser humano motivado y lleno de vida, impulsándolo a explorar nuevas

experiencias y lograr una sensación de realización personal. Los términos "dream", "star" "light" y “new” simbolizan metas e inspiraciones; "soul", "fire" y “morning” reflejan la motivación interna; mientras que "street", "sound", “city” y "pain" vinculan estas aspiraciones con la realidad.

Tópico 8 – “Reflexiones temporales”: Refiere al paso del tiempo con términos como “time”, “now”, “long”, “enough”, “nowhere” y “right now”, enfatizando la irreversibilidad del tiempo y la importancia de disfrutar el presente, un tema recurrente en la música.

6.3 Análisis de los tópicos: su relevancia en el corpus y su evolución a lo largo de las décadas 2016-2020

Para profundizar y continuar con el estudio de los ocho tópicos seleccionados se realizan dos análisis: (1) mediante una representación gráfica de barras se analiza la importancia de cada tópico dentro del corpus final, y (2) se utiliza un gráfico de líneas para estudiar la evolución de los tópicos a lo largo de las décadas. Para completar los análisis anteriores y entender completamente la evolución musical desde todas las perspectivas, la última parte del estudio se centra en el contenido de las canciones. Se estudian los tópicos, su evolución y prevalencia a lo largo del tiempo para determinar si efectivamente las letras de las canciones son un reflejo de los cambios y condiciones sociales de cada década.

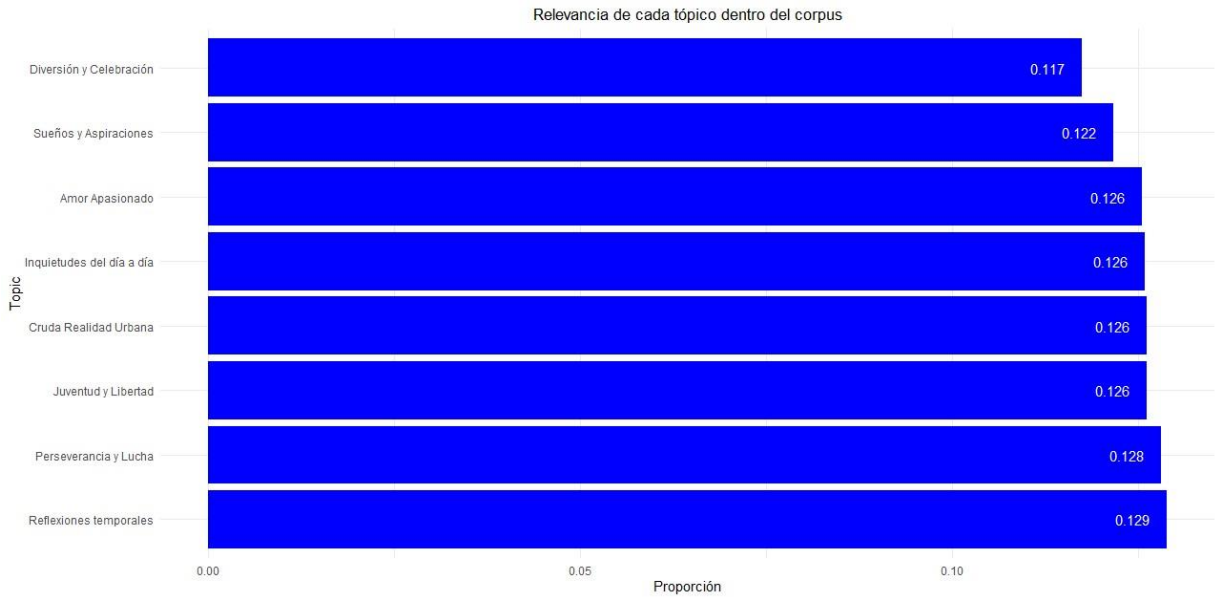
Por un lado, para determinar cuál de estos ocho tópicos tiene mayor relevancia en el corpus, es decir, dentro de las 1050 canciones, se utiliza el peso que LDA asigna a cada tópico dentro de cada documento. Es decir, se tiene en cuenta la distribución de tópicos dentro de cada documento. Aquellos tópicos que en media tienen un mayor peso en todo el corpus son considerados los más relevantes (*Figura 19*). Aunque la distribución de la importancia de los tópicos es bastante equilibrada, con probabilidades en torno al 12,5%, el tópico de "Reflexiones temporales" sobresale ligeramente, presentando una probabilidad de casi el 13% de aparecer en el corpus. En contraste, el tópico "Diversión y Celebración" se encuentra en la posición más baja del ranking, con una probabilidad del 11,7%.

El tópico de “Reflexiones temporales” destaca como el más relevante, lo cual tiene su sentido dada su universalidad, su transversalidad y su facilidad de integrarse con otros

temas. Este enfoque es frecuente en la música, no solo por la facilidad de uso de estos términos sino también porque a muchos artistas les gusta explorar sobre la nostalgia y la fugacidad del tiempo. Además, el tópico de “Perseverancia y Lucha” es común en las canciones, mientras que los otros tópicos, con una probabilidad del 12,6%, se acercan al peso medio.

En resumen, no hay una gran diferencia entre los pesos de los ocho tópicos, tan solo 1,2 puntos porcentuales entre el más y el menos relevante. Después de todo, la música es una vía de escape para los artistas, un medio por el cual expresan sus pensamientos y sentimientos, todo lo que les ocupa la mente, lo que conlleva una gran diversidad temática en las canciones. Por este motivo, cada uno de los tópicos identificados tiene una probabilidad similar de representación en el corpus.

Figura 19. Relevancia de cada tópico dentro del corpus



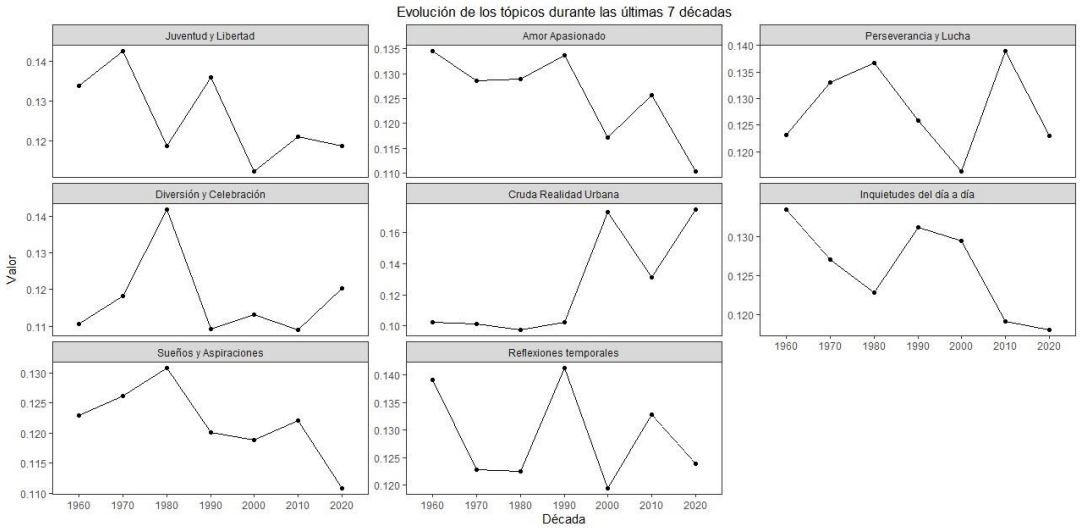
Fuente: Elaboración propia a partir de los datos del estudio y adaptado del código propuesto en Velilla (2023)

Para concluir el análisis de las letras de las canciones, se estudia cómo los ocho tópicos seleccionados han ido cambiando a lo largo de las últimas décadas. Esto permite identificar los temas que han sido cruciales en ciertas décadas, aquellos que han ganado o perdido relevancia con el tiempo, y determinar los más destacados en la actualidad.

Con solo observar la figura 20, se obtienen varias conclusiones, destacando principalmente que ningún tópico ha mantenido su relevancia a través de las últimas siete décadas. Esto tiene sentido, dado que cada década es diferente y los comportamientos

sociales y los eventos significativos cambian, influyendo así en la evolución de los temas tratados en las canciones.

Figura 20. Evolución del peso de los tópicos a lo largo de las últimas siete décadas (2016-2020)



Fuente: Elaboración propia a partir de los datos del estudio y adaptado del código propuesto en Velilla (2023)

Otras observaciones que pueden apreciarse en la figura 20 son:

- Contrario a la creencia popular, las canciones actuales ya no enfatizan tanto el tema del "Amor". Esta situación contrasta con las décadas de 1960 a 1990, cuando las relaciones amorosas dominaban las letras de las canciones. Desde entonces, la importancia del tema ha disminuido, teniendo una menor relevancia hoy en día. Una posible explicación para este cambio es que no es que las canciones modernas ya no tratan el amor, sino que la propia definición del concepto ha experimentado notables cambios. Quizá esto se puede deber a que, en la actualidad, la gente tiende a evitar compromisos a largo plazo sintiendo miedo y temor hacia relaciones serias. Muchas personas prefieren no comprometerse de por vida, ser más independientes y buscar la gratificación de manera más inmediata.
- Coincidiendo con los resultados del capítulo 5, el tópico de "Cruda Realidad Urbana", que contiene mayoritariamente términos explícitos, ha ido ganando relevancia desde la década de los 90. Las canciones modernas ahora adoptan tonos más agresivos, irrespetuosos y discriminatorios. En una sociedad que presume de

su amplia libertad de expresión, los límites parecen haber desaparecido. Ahora, los artistas pueden hablar de cualquier tema en sus canciones, incluso si implican confrontaciones directas o faltas de respeto con otras personas, sin tener grandes consecuencias.

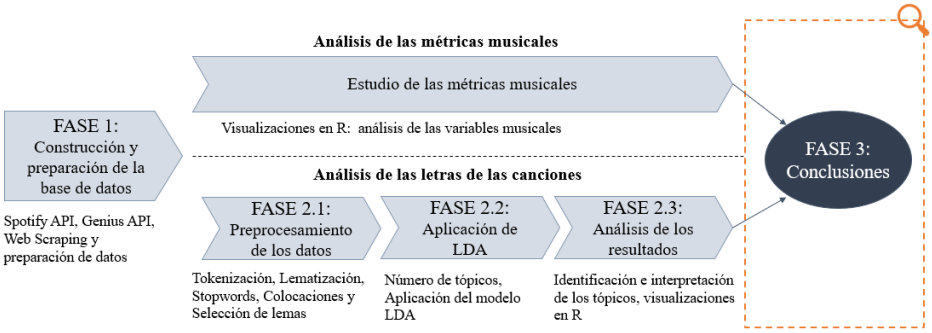
- Durante los años 80, considerados por muchos como la época dorada de la música, temas como "Diversión y Celebración" y "Sueños y Aspiraciones" alcanzaron su punto álgido. Esta era estuvo dominada por grandes figuras del pop y el rock, cuya energía y habilidad para animar al público marcaron un hito. Hoy en día, el tema de "Diversión y Celebración" ha resurgido, coincidiendo con el éxito de las canciones de reguetón y la salida de la pandemia. Al final, cuando al ser humano se le prohíbe algo, en el momento en el que vuelve a recuperar la libertad, regresa con más ganas e ilusión (Alonso, 2022). Después de muchos meses confinados, la gente necesitaba socializar, disfrutar y divertirse con amigos y familia; los eventos y las discotecas volvieron con el ritmo bailable del reguetón.
- En la década de 2010, el tema de "Perseverancia y Lucha" cobró especial relevancia. Este período estuvo marcado por desafíos significativos, como crisis económicas que afectaron profundamente a la sociedad. En respuesta, emergieron canciones que fomentaban la resiliencia y el esfuerzo como medios para superar adversidades, posicionando a la música como un refugio para encontrar felicidad y alegría.
- El tema "Juventud y Libertad" alcanzó su apogeo en los años 70, coincidiendo con la popularidad de la música disco. Este género, que animaba a la gente a liberarse bailando, llenaba las discotecas de jóvenes deseosos de divertirse y evadir la monotonía de sus vidas, representando una forma de liberación (Moral, 2019).

En conclusión, aunque no hay diferencias notables en el peso de los tópicos en el corpus global (todos próximos al peso medio del 12,5%), si se estudia cada década por separado, se pueden observar diferencias más notables y llegar a más conclusiones.

Capítulo 7. Discusión de resultados, conclusiones y futuras líneas de investigación

Este trabajo de investigación concluye recopilando todos los resultados obtenidos tanto en el análisis de las métricas musicales como en el análisis de las letras de las canciones (*Figura 21*). A continuación, se discuten los resultados y las conclusiones en el apartado 7.1, y se detallan las limitaciones y las posibles extensiones del trabajo en el apartado 7.2

Figura 21. Procedimiento del Trabajo de Investigación: Fase 3



Fuente: Elaboración propia

7.1 Discusión de los resultados y conclusiones

Después de analizar las letras de las canciones desde dos enfoques distintos -técnico, mediante el estudio de sus métricas musicales, y cualitativo, al examinar su contenido lírico— se han obtenido los argumentos y conclusiones necesarios para dar respuesta al objetivo principal de este trabajo de investigación y a las preguntas planteadas en el apartado 1.2.

Gracias al uso de las APIs de Spotify y Genius, y a las técnicas de Web Scraping y Topic Modeling en R, se ha logrado recopilar suficiente información como para construir una base de datos sólida y coherente y estudiar sobre la evolución de las canciones a lo largo de las últimas siete décadas (2016-2020). Algunas conclusiones coinciden con investigaciones previas, otras difieren y en algunos casos específicos, se han realizado análisis adicionales para completar el estudio de investigación.

En relación con las métricas musicales, se han obtenido las siguientes conclusiones:

En primer lugar, coincidiendo con los resultados obtenidos por Interiano et al. (2018), se destaca que no existen reglas específicas que aseguren el éxito de ninguna canción. Definir las características que contribuyen al éxito de una canción representa un desafío significativo, debido a la singularidad de cada pieza musical y a la diversidad de factores que intervienen. Aunque es cierto que, a través de nuestro análisis, se han identificado ciertas variables como la energía, la sonoridad y el contenido explícito que sí muestran estar relacionadas con la popularidad de las canciones. Sin embargo, estas no garantizan de manera absoluta alcanzar el éxito. Quizá estos resultados pueden apuntar a que la clave no radica, como sugiere Interiano et al., en crear canciones con gran capacidad distintiva y alejadas de las tendencias generales, sino más bien en producir canciones que respondan a las necesidades sociales y culturales del momento. Por ejemplo, en la música disco de los años 70, aunque uno quisiera diferenciarse del resto, si se quería alcanzar el éxito, las canciones debían ser bailables, rítmicas y con estribillos pegajosos. Este análisis podría ser ampliado mediante el uso de una técnica de estadística inferencial.

En segundo lugar, y como era de esperar, el contenido explícito en las canciones ha crecido considerablemente en los últimos años. Como se menciona en el capítulo 5, aproximadamente el 40% de las canciones de los últimos diez años contienen lenguaje inapropiado y ofensivo. Actualmente, nuestra sociedad valora la autenticidad y la libertad de expresión. Sin embargo, nos encontramos en un ambiente social tenso y polarizado, donde la música se transforma en un canal para la crítica y el desahogo social, lo que ha incrementado la presencia de letras agresivas e inapropiadas en las canciones. Como ya señalaron Fisher y Greitemeyer (2006), este fenómeno impacta significativamente en la sociedad, generando sentimientos de venganza y crispación que pueden llegar a perturbar el orden social y fomentar actitudes negativas entre hombres y mujeres. Esto podría estar relacionado con el incremento generalizado del índice de criminalidad. Mientras que España ha registrado un aumento del 3% en el año 2023 con respecto a los datos de 2022, Estados Unidos de un 1% (Numbeo, 2023).

Junto a estos aspectos se añade la significativa disminución en la duración de las canciones. Según los resultados de este estudio y coincidiendo con Gausby (2015) y Piera (2023), se observa una tendencia descendente en la duración de las canciones en los

últimos años. Esta parece estar relacionada con los cambios en los comportamientos de la sociedad contemporánea. Influenciados por la cantidad de estímulos disponibles a través de internet y otras tecnologías, los jóvenes han perdido toda capacidad de atención y concentración. Son más inquietos, impacientes, y apuestan por un consumo rápido. Los artistas musicales, al igual que otros como los creadores de contenido o publicistas, para capitalizar estos cambios, no han tenido que ajustarse a estas nuevas necesidades reduciendo así la duración de sus vídeos y simplificando los mensajes para que sean breves y claros.

Y, por último, los cambios sociales contemporáneos como la globalización, la mayor conectividad o la reivindicación de la mujer han dejado su huella en la música. La globalización ha transformado la industria musical, introduciendo una mayor variedad de artistas (en términos de composiciones, género y nacionalidad) y permitiendo que músicos, sin el respaldo de grandes discográficas, alcancen las listas de éxitos. En los últimos años, se ha visto cómo canciones en diversos idiomas, desde el español con artistas como Karol G hasta el coreano con grupos como Blackpink, han logrado popularidad internacional. Paralelamente, la reivindicación de la mujer, un tema recurrente en agendas políticas y medios de comunicación, también ha encontrado eco en la música. La presencia de solistas femeninas ha ganado terreno, a menudo superando a las bandas en popularidad. En la lista de los artistas más destacados de la década de 2020, la representación entre hombres y mujeres es prácticamente equitativa.

En cuanto al análisis del contenido lírico de las canciones, la técnica de Topic Modeling ha permitido estudiar la evolución y prevalencia de ocho tópicos a lo largo de las últimas décadas y obtener conclusiones, mostrando así los primeros indicios de la relación existente entre el contenido de las canciones (los tópicos) y la sociedad. Los tópicos no mantienen la misma relevancia a lo largo del tiempo, sino que evolucionan, a la vez que lo hace la sociedad. Como es el caso del tópico de “Cruda Realidad Urbana”, que en estas últimas décadas ha experimentado una tendencia ascendente, cambio que se justifica con el aumento de la radicalización, la libertad de expresión, y el uso de contenido explícito y lenguaje vulgar (como palabrotas) entre las generaciones actuales. Asimismo, el tópico “Diversión y Celebración” también ha experimentado un incremento en esta última década, coincidiendo con la popularidad del reguetón y posiblemente debido al fin de la pandemia, lo que ha motivado a la gente a salir y disfrutar.

En pocas palabras, la música, aunque a veces pueda parecer insignificante, está relacionada con la sociedad, se configura según la evolución de la sociedad, convirtiéndose así en una fuente de aprendizaje y de conocimientos. A partir de los análisis realizados, es razonable afirmar que la música sí que refleja los cambios de la sociedad. Como respuesta directa a la evolución de la sociedad, a los cambios en sus comportamientos, a la globalización y al aumento de la conectividad, la música ha ido cambiando: la duración de las canciones ha disminuido, el contenido inapropiado en las letras ha aumentado, los tópicos de las canciones han evolucionado y ahora existe una mayor diversidad tanto en los artistas como en los idiomas de las canciones. Estudiar la música y en concreto las canciones, tiene mucha utilidad ya que como bien decían Dodds y Danforth (2010), beneficia no solo a los músicos, sino también a quienes están fuera de la industria. Por ejemplo, comprender por qué la gente hoy día prefiere canciones de corta duración puede ofrecer pistas valiosas para agencias de comunicación y marketing sobre cómo diseñar sus campañas.

7.2 Limitaciones y futuras líneas de investigación

A la hora de desarrollar ese trabajo de investigación, se han identificado ciertas limitaciones relacionadas con la fuente de información utilizada. Por un lado, la API de Spotify ofrece numerosas variables, pero no proporciona información sobre el género musical de las canciones o del artista (por ejemplo, pop, rock, jazz). Esto ha impedido identificar con precisión los géneros más representativos de cada década y complementar los resultados obtenidos con otros análisis adicionales.

Por otro lado, las canciones elegidas para nuestro estudio son aquellas que el equipo de expertos interno de Spotify ha identificado como las más relevantes sin considerar la opinión de los oyentes en otras plataformas. Por lo que es probable que las listas varíen ligeramente en otras aplicaciones. Estas limitaciones podrían abordarse ampliando el alcance de nuestro estudio. Se podría cruzar y complementar los datos proporcionados por Spotify con información adicional de otras plataformas digitales como Apple Music para crear una base de datos más completa y fiable. Y se podría incluir el género musical de las canciones y otras variables (como por ejemplo la nacionalidad de los cantantes) para ampliar el estudio.

Además, en cuanto al análisis de las letras de las canciones, es necesario hacer notar que los artistas suelen utilizar expresiones y modificar palabras para que encajen con el ritmo, lo cual dificulta la categorización precisa de las letras, así como la identificación e interpretación de los tópicos. Una forma de mejorar este aspecto en futuras investigaciones sería la utilización de una muestra mayor de canciones para cada década.

En este trabajo, se ha puesto el foco en el análisis de las características técnicas de las canciones y en el contenido (o tópicos) de las mismas, dando una visión preliminar de cómo los cambios en la música pueden responder a cambios en la sociedad. No obstante, con el objeto de corroborar y complementar los resultados y conclusiones obtenidas, se propone como futura línea de investigación la incorporación de un análisis más profundo y riguroso desde la perspectiva de la psicología y la sociología.

Declaración sobre el uso de la Inteligencia Artificial

Por la presente, yo, Marta Ybarra Fernández-Iriondo, estudiante de E2 + Analytics de la Universidad Pontificia de Comillas, al presentar mi Trabajo Fin de Grado titulado "Estudio de la evolución de la música y su relación con los cambios en la sociedad a través de técnicas de análisis de datos", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa Chat GPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

- Corrector de estilo literario y de lenguaje: para mejorar la calidad lingüística y estilística del texto.

- Traductor: para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado Chat GPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 19 de junio de 2024

Firma:

Bibliografía

Alonso, R. (2017). La música es lo único que puede salvar a una sociedad cargada de dolor Y llantos. *elDiario.es*. https://www.eldiario.es/cantabria/cultura/musica-salvar-sociedad-cargada-llantos_128_3139682.html#:~:text=El%20cantautor%20gallego%20Andr%C3%A9s%20Su%C3%A1rez%20(Ferrol%2C%201983)%2C%20con,la%20poes%C3%ADa%20y%20la%20palabra%E2%80%9D.

Antal , D. (2022). Spotifyr: R wrapper for the "Spotify" web api. https://cloud.r-project.org/web/packages/spotifyr/spotifyr.pdf

Alonso, M. (2022) Días de Mucha Fiesta: Después de la pandemia llega la fiebre por salir, ¿Aguantaremos?. *Mujer Hoy*. https://www.mujerhoy.com/actualidad/fiesta-excesos-opulencia-despues-de-la-pandemia-20220319122441-nt.html (Accessed: 04 June 2024)

Arizaga, M. (2023). La música disco, UN Género Emblemático de los 70. *LaCarne Magazine.* https://lacarnemagazine.com/la-musica-disco-genero-emblematico-los 70/#:~:text=Este%20tipo%20de%20m%C3%BAsica%20se,y%20sus%20arreglos%20orquestales%20exuberantes.&text=Algunos%20de%20los%20rasgos%20distintivos,altamente%20adecuada%20para%20el%20baile.

Bhattacharjee, S., Gopal, R. D., Lertwachara, K., Marsden, J. R., & Telang, R. (2007). The effect of digital sharing technologies on music markets: A survival analysis of albums on ranking charts. Management Science, 53(9), 1359-1374.

Blei, D., Carin, L., & Dunson, D. (2010). Probabilistic topic models: A focus on graphical model design and applications to document and image analysis. *IEEE signal processing magazine*, *27*(6), 55-65. https://www.researchgate.net/publication/264630088_Probabilistic_Topic_Models_A_focus_on_graphical_model_design_and_applications_to_document_and_image_analysis

Blei, D. (2012). "Probabilistic topic models." Communications of the ACM 55 (4): 77. https://dl.acm.org/doi/pdf/10.1145/2133806.2133826

Bail, C. [Summer Institute in Computational Social Science] (2020). An Introduction to Topic Modeling [Video]. YouTube**.** https://www.youtube.com/watch?v=IUAHUEy1V0Q

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Muller, S., . . . Lowe, W. (2023). Package "quanteda". CRAN Repository. https://cran.r-project.org/web/packages/quanteda/quanteda.pdf

Church, K., & Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1), 22-29. https://aclanthology.org/J90-1003.pdf

DalleMule, L., & Davenport, T. H. (2017). What's Your Data Strategy? *Harvard Business Review*. https://hbr.org/2017/05/whats-your-data-strategy

Ding, Y., & Yan, S. (2015). Topic Optimization Method Based on Pointwise Mutual Information. In *Neural Information Processing: 22nd International Conference, ICONIP 2015, Istanbul, Turkey, November 9-12, 2015, Proceedings Part III 22* (pp. 148-155). Springer International Publishing.

Dodds, P. S., & Danforth, C. M. (2010). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. Journal of happiness studies, 11, 441-456

Fischer, P., & Greitemeyer, T. (2006). Music and aggression: The impact of sexual-aggressive song lyrics on aggression-related thoughts, emotions, and behavior toward the same and the opposite sex. Personality and Social Psychology Bulletin, 32(9), 1165-1176

Gausby, A. (2015) Attention spans. *Microsoft Canada* https://dl.motamem.org/microsoft-attention-spans-research-report.pdf

Grun, B., & Hornik, K. (2023). Package ‘topicmodels’. Repositorio CRAN. https://cran.rproject.org/web/packages/topicmodels/topicmodels.pdf

Henderson, E. (2022). Geniusr: Tools for working with the “genius” API. https://cran.r-project.org/web/packages/geniusr/geniusr.pdf

Interiano, M., Kazemi, K., Wang, L., Yang, J., Yu, Z., & Komarova, N. L. (2018). Musical trends and predictability of success in contemporary songs in and out of the top charts. *Royal Society open science*, 5(5), 171274

IFPI. (2022). *Engaging with Music 2022* https://www.ifpi.org/wp-content/uploads/2022/11/Engaging-with-Music-2022_full-report-1.pdf

IFPI (2023) Global Music Report 2023. IFPI GLOBAL MUSIC REPORT 2023. https://globalmusicreport.ifpi.org/

Krotov, V., & Silva, L. (2018). Legality and ethics of web scraping. *Emergent Research Forum (ERF)*

Lander, J. P. (2014). R for Everyone: Advanced Analytics and Graphics. Boston, MA: Addison-Wesley.

Leighton, M. (2023) ¿Cuáles son los distintos tipos de listas de reproducción de spotify?, *Groover Blog.* https://blog.groover.co/es/consejos-para-musicos/tipos-de-listas-de-reproduccion-de-spotify-es/#3_Listas_de_reproduccion_editoriales_El_objetivo_final_de_los_artistas_emergentes

Laoh, E., Surjandari, I., & Febirautami, L. R. (2018). Indonesians' Song Lyrics Topic Modelling Using Latent Dirichlet Allocation. In *2018 5th International Conference on Information Science and Control Engineering (ICISCE)* (pp. 270-274). IEEE

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262-272).

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., . . . Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. Communication Methods and Measures: Special Issue on Computational Methods, 12, 93-118

Meindertsma, P. (2019). Changes in lyrical and hit diversity of popular US songs 1956-2016. Digital humanities quarterly, 13(4)

Moral, S. (2019). Disco inferno: 40 Años del Fin de la Música Disco: Música. LOS40. https://los40.com/los40/2019/07/29/los40classic/1564392545_472554.html

Mena Roa, M. (2023). Usuarios activos y de pago de Spotify. *Statista.* https://es.statista.com/grafico/19793/usuarios-activos-y-de-pago-de-spotify/

MusicBrainz. (s.f.). *Acerca de MusicBrainz*. https://musicbrainz.org/doc/About

Mehltrettervatter, J., & Sohnius, M.-L. (n.d.). Chapter 8 Google News API: Apis for Social Scientists: A Collaborative Review. https://bookdown.org/paul/apis_for_social_scientists/google-news-api.html

Nguyen, E. (2014). Chapter 4-Text mining and Network Analysis of Digital Libraries. *Data Mining Applications with R (by Zhao Y., Cen Y.), Academic Press, New York*, 514.

Niekler, A., & Wiedemann, G. (2017). Tutorial 6: Topic Models. https://nballier.github.io/tm4ss.github.io/Tutorial_6_Topic_Models.html#3_topic_distributions

Numbeo. (2023). Clasificaciones de criminalidad por país. https://es.numbeo.com/criminalidad/clasificaciones-por-pa%C3%ADs?title=2023

Petrovic, S., Snajder, J., Dalbelo-Basic, B., & Kolar, M. (2006). Comparison of collocation extraction measures for document indexing. *In 28th International Conference on Information Technology Interfaces, 2006.* (pp. 451-456). IEEE.

Piera, M. (2023). Canciones, cada vez más cortas y con estribillos al inicio: la fórmula para sobrevivir en el streaming. *La Vanguardia.* https://www.lavanguardia.com/vivo/20231125/9403098/canciones-vez-mas-cortas-estribillos-inicio-formula-sobrevivir-streaming.html

Porras, H. (s.f.). *Análisis de comportamiento en redes sociales usando Procesamiento del Lenguaje Natural.* RPubs. https://rpubs.com/hugoporras/nlp_capitulo3_2

Rosner, F., Hinneburg, A., Röder, M., Nettling, M., & Both, A. (2014). Evaluating topic coherence measures. https://doi.org/10.48550/arXiv.1403.6397

Stafford, S. A. (2010). Music in the digital age: The emergence of digital music and its repercussions on the music industry. *The Elon Journal of Undergraduate Research in Communications*, 1(2), 112-120.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 952-961).

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63-70).

Sirisuriya, D. S. (2015). A comparative study on web scraping.

Sohan, S. M., Anslow, C., & Maurer, F. (2015). A case study of web API evolution. In *2015 IEEE World Congress on Services* (pp. 245-252). IEEE.

Schweinberger, M. (2023). *Part-of-Speech Tagging and Dependency Parsing with R*. Brisbane: The University of Queensland. https://slcladal.github.io/postag.html

Sust, L., Stachl, C., Kudchadker, G., Bühner, M., & Schoedel, R. (2023). Personality Computing With Naturalistic Music Listening Behavior: Comparing Audio and Lyrics Preferences. Collabra: Psychology, 9(1), 75214. https://doi.org/10.1525/collabra.75214

Spotify. (s.f.). Spotify Web API Documentation. https://developer.spotify.com/documentation/web-api/

Trecenti, J., & Wundervald, B. (2019). Music Data Analysis in R. IV International Seminar on Statistics with R. DOI:10.13140/RG.2.2.16222.48966

Tong, Z., & Zhang, H. (2016). A text mining research based on LDA topic modelling. In *International conference on computer science, engineering and information technology* (pp. 212-279).

Trenquier, H. (2018). Improving Semantic Quality of Topic Models for Forensic Investigation. *University of Amsterdam*, 2017-2018. https://www.os3.nl/_media/2017-2018/courses/rp2/p76_report.pdf

Tapia, K. (2021). MTV cumple 40 años: cómo ha evolucionado el canal que cambió la música. *Digital Trends en Español*. https://es.digitaltrends.com/entretenimiento/mtv-40-anos/

Úbeda, L. (2020). ¿A qué nos referimos cuando hablamos de 'música urbana'? *LOS40*. https://los40.com/los40/2020/04/13/los40urban/1586778819_804441.html

Varnum, M. E., Krems, J. A., Morris, C., Wormley, A., & Grossmann, I. (2021). Why are song lyrics becoming simpler? A time series analysis of lyrical complexity in six decades of American popular music. *PloS one*, 16(1), e0244576

Velilla Cadahia, M. D. R. (2023). Guía práctica de implementación de topic modeling en R: analizando artículos periodísticos sobre el metaverso-Velilla Cadahía, María del Rocío. http://hdl.handle.net/11531/68840

Anexos

Anexo I. Breve descripción de las variables musicales proporcionadas por la API de Spotify

Variables proporcionadas	Variables renombradas	Rango	Definición
Danceability	Capacidad de baile / Bailabilidad	Número [Flotante]: De 0 a 1 Un valor de 1 indica una alta bailabilidad	Describe cómo de adecuada es una canción para bailar, basándose en una combinación de elementos musicales que incluyen tempo, estabilidad rítmica, fuerza del ritmo y regularidad en general
Energy	Energía	Número [Flotante]: De 0 a 1 Un valor de 1 indica que la canción es densa, rápida, ruidosa y muy energética	Medida perceptiva de la intensidad y la actividad Las características perceptivas que contribuyen a este atributo incluyen el rango dinámico, la sonoridad percibida, el timbre, la velocidad de aparición y la entropía general
Loudness	Volumen	Número [Flotante] Decibelios (dB), que suelen oscilar entre -60 dB y 0dB	Volumen de una canción, es decir, valores que reflejan la sonoridad de la canción. Cualidad de un sonido que es el correlato psicológico primario de la fuerza física (amplitud)
Mode	Tonalidad	Número [Entero] Valor “major” indica “mayor” [1] mientas que “minor”, “menor” [0]	La tonalidad (mayor o menor) de una canción. El tipo de escala a partir de la cual se deriva un contenido melódico.

Speechiness	Locuacidad	Número [flotante] De 0 a 1 Cuánto más cerca al valor 1, más exclusivamente hablado es la grabación (programas de entrevistas, Podcast) Valores > 0,66 indican canciones que están compuestas en su totalidad por palabras habladas. Valores > 0,33 o < 0,66 indican canciones que pueden contener tanto música como voz. Valores < 0,33 representan probablemente música y otros elementos no verbales	Detecta la presencia de las palabras habladas en una canción
Acousticness	Acústica	Número [flotante]: De 0 a 1 Un valor de 1 indica que una canción es puramente acústica	Describe si la canción utiliza principalmente instrumentos acústicos o electrónicos/eléctricos. Es decir, cómo de acústica es.

Instrumentalness	Instrumentalidad	Número [flotante]: De 0 a 1 Cuánto más cerca este el valor a 1, más probable es que la canción no contenga voces. Los valores superiores a 0,5 representan canciones instrumentales, pero la confianza es mayor a medida que se acerca al 1	Describe hasta qué punto el cantante no es el intérprete principal de la canción. Los sonidos "ooh" y "aah" se consideran instrumentales en este contexto.
Valence	Positividad	Número [flotante]: De 0 a 1 Cuánto más próximo a 1, más alegres, animadas y positivas son las canciones. Cuánto más cercanas a 0, más negativas y tristes son	Describe la positividad musical transmitida por una canción
Duration	Duración	Número [Entero]: Milisegundos	Longitud de la canción
Popularity	Popularidad	Numérico	La popularidad de la canción.
Explicit	Explicito	Dicotómica El valor TRUE [1] indica que una canción contiene letras explícitas mientras que FALSE [0] que no.	Si la canción contiene o no letras explícitas

Anexo II. Código de R.

1. Análisis de las características musicales / técnicas de las canciones

```
install.packages("glue")
install.packages("spotifyr")
install.packages("pacman")
install.packages("kableExtra")
install.packages("geniusr")
install.packages("magrittr")
install.packages("reshape2")
install.packages("paletteer")

pacman::p_load('spotifyr',
               'tidyverse',
               'plotly',
               'ggimage',
               'kableExtra',
               'httpuv',
               'httr')
library(lubridate)
library(dplyr)
library(stats)
library(kableExtra)
library(tidytext)
library(spotifyr)
library(ggplot2)
library(rvest)
library(purrr)
library(reshape2)
```

```
# Here you can store the credentials as follows:

Sys.setenv(SPOTIFY_CLIENT_ID="ddb4379c217740a3838b1cac37e39d0e")

Sys.setenv(SPOTIFY_CLIENT_SECRET="45893f8aea6e41b88e995d69278ff8a7")

#Sys.setenv(SPOTIFY_REDIRECT_URI="http://localhost:3000/callback")

access_token <- get_spotify_access_token()


#Guardar los audio feautures para cada década

top_1960s <- get_playlist_audio_features(playlist_uris =
'37i9dQZF1DXaKIA8E7WcJj')

song_artist_1960 <- map_chr(top_1960s$track.artists, function(x) x$name[1])

song_artist_1960 <- c(song_artist_1960)

top_1960s$track.artists <- song_artist_1960

… Este mismo paso para cada una de las décadas, cada década tiene su propia lista en
Spotify y, por tanto, su propio playlist_uris

library(lubridate)

#Combinamos todas las bases de datos para crear una SOLA

#Limpiar y ajustar la base de datos GLOBAL

datos_totales <- rbind(top_1960s, top_1970s, top_1980s, top_1990s, top_2000s,
top_2010s, top_2020s)

columnas_mantener <- c("playlist_name", "danceability", "energy", "key", "loudness",
"mode", "speechiness", "acousticness",  "instrumentalness", "liveness", "valence",
"tempo","time_signature", "mode_name", "track.artists", "track.available_markets",
"track.duration_ms", "track.explicit", "track.name", "track.popularity",
"track.album.album_type", "track.album.name", "track.album.release_date")

datos_finales <- datos_totales[, c(columnas_mantener)]

colnames(datos_finales) <- c("Playlist name", "Danceability", "Energy", "Key",
"Loudness", "Mode", "Speechiness", "Acousticness", "Instrumentalness", "Liveness",
"Valence", "Tempo","Time Signature", "Mode name", "Artist name", "Available
markets","Duration Ms", "Explicit", "Song title", "Popularity", "Type of album",
"Album name", "Release date")

#Transformación de algunas variables para el posterior análisis

datos_finales$Explicit [datos_finales$Explicit == "TRUE"] <- 1

datos_finales$Explicit [datos_finales$Explicit == "FALSE"] <- 0

datos_finales$Explicit <- as.integer(datos_finales$Explicit)

datos_finales$Mode [datos_finales$Mode == "major"] <- 1

datos_finales$Mode [datos_finales$Mode == "minor"] <- 0
```

```
datos_finales$Mode <- as.integer(datos_finales$Mode)

datos_finales$Year <- substr(datos_finales$`Release date`, 1, 4)

datos_finales$Year <- as.numeric(datos_finales$Year)

datos_finales$Decada <- floor(datos_finales$Year / 10) * 10

datos_finales$Duration<- as.numeric(datos_finales$`Duration Ms`)/ 60000


datos_finales <- datos_finales %>%

  group_by(Year) %>%

  mutate(Loudness_Normalizada = (Loudness - min(Loudness)) / (max(Loudness) -
min(Loudness)))

write_xlsx(datos_finales, "C:\\Users\\marta\\OneDrive - Universidad Pontificia
Comillas\\TFG ANALYTICS\\database_Spotify.xlsx")

database_Spotify <- read_excel("C:\\Users\\marta\\OneDrive - Universidad Pontificia
Comillas\\TFG ANALYTICS\\Version Final\\database_Spotify.xlsx")


#Análisis de las métricas musicales

variables_estudiar <- c("Danceability", "Energy", "Loudness", "Mode", "Speechiness",
"Acousticness", "Valence", "Instrumentalness", "Duration Ms", "Explicit")


#Análisis 1 - Evolución de las variables musicales (numéricas)

resumen_por_año <- database_Spotify  %>% group_by(Year) %>%

  summarize(promedio_Speech = mean(Speechiness), promedio_Instrument =
mean(Instrumentalness),

            promedio_acustico= mean(Acousticness), promedio_danceability=
mean(Danceability),

            promedio_energy= mean(Energy), promedio_valence= mean(Valence),
promedio_liveness= mean(Liveness), promedio_loudness=
mean(Loudness_Normalizada))

resumen_por_año$Year <- as.numeric(resumen_por_año$Year)

resumen_por_año$Decada <- floor(resumen_por_año$Year / 10) * 10


ggplot(resumen_por_año, aes(x = Year, group = 1)) +

  geom_line(aes(y = promedio_Speech, color = "Speechiness"), size = 1) +

  geom_line(aes(y = promedio_Instrument, color = "Instrumentalness"), size = 1) +

  geom_line(aes(y = promedio_acustico, color = "Acousticnesss"), size = 1) +
```

```
  geom_line(aes(y = promedio_danceability, color = "Danceability"), size = 1) +
  geom_line(aes(y = promedio_energy, color = "Energy"), size = 1) +
  geom_line(aes(y = promedio_valence, color = "Valence"), size = 1) +
  geom_line(aes(y = promedio_loudness, color = "Loudness"), size = 1) +
  labs(x = "Año",
       y = "Promedio de las variables musicales por año",
       title = "Evolución de las variables musicales desde la década de 1960 hasta 2020")
+
  scale_fill_brewer(palette = "Set3") +
  scale_x_continuous(breaks = seq(min(resumen_por_año$Year),
max(resumen_por_año$Year), by = 10),  # Establece las marcas del eje X cada 10 años
                     labels = paste0(seq(min(resumen_por_año$Year),
max(resumen_por_año$Year), by = 10), "")) +  # Etiquetas de décadas
  theme(plot.title = element_text(hjust = 0.5),
        panel.background = element_rect(fill = "white", color = NA),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(color = "black"))


#Análisis 2 - Evolución de las variables musicales (dicotómicas)
Dicotomicas <- database_Spotify %>%
  select(Year, Decada, Explicit, Mode)


library(dplyr)
percentage_data <- Dicotomicas %>%
  group_by(Decada) %>%
  summarize(Percentage_Explicit = mean(Explicit) * 100,
            Percentage_Mode = mean(Mode) * 100)


p1 <- ggplot()+
  geom_bar(data = percentage_data, aes(x = Decada, y = Percentage_Explicit, fill =
"Explicit"), stat = "identity", position = "dodge") +
  scale_fill_brewer(palette = "Set3") +
```

```
  labs(x = "Década", y = "Porcentaje", title = "Porcentaje de canciones con palabras explícitas por década")+
  theme(plot.title = element_text(hjust=0.5),
        panel.background = element_rect(fill = "white", color = NA),  # Fondo blanco sin recuadros
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(color = "black")) + guides(fill = FALSE)

p2 <- ggplot()+
  geom_bar(data = percentage_data, aes(x = Decada, y = Percentage_Mode, fill = "Mode"), stat = "identity", position = "dodge") +
  scale_fill_brewer(palette = "Set3") +
  labs(x = "Década", y = "Porcentaje", title = "Porcentaje de canciones en tonalidad mayor por década")+
  theme(plot.title = element_text(hjust=0.5),
        panel.background = element_rect(fill = "white", color = NA),  # Fondo blanco sin recuadros
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(color = "black")) + guides(fill = FALSE)

library("gridExtra")
grid.arrange(p1, p2, nrow = 2)

#Análisis 3 - Evolución de la variable: duración de las canciones
duracion <- database_Spotify %>%
  select(Year, Decada, Duration)
g2 <- ggplot(duracion, aes(x = Year, y = Duration, fill = Duration)) +
  geom_boxplot() +
  stat_summary(fun = "mean", geom = "point", shape = 8,
               size = 2, color = "white") +
  labs(title = "Variaciones en la duración (en minutos) de las letras de las canciones (década 1960-2020)", x = " ", y = "Duración (en minutos)") +
```

```
  theme(plot.title = element_text(hjust=0.5),
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1),
        panel.background = element_rect(fill = "white", color = NA),  # Fondo blanco sin recuadros
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(color = "black"))


resumen_duracion <- duracion %>%
  group_by(Year) %>%
  summarize(promedio = mean(Duration))
g1 <-ggplot(resumen_duracion, aes(x = Year, group = 1)) +
  geom_line(aes(y = promedio), color = "darkblue", size = 1) +
  labs(x = " ", y = "Promedio anual de duración (en minutos)", title = "Evolución en la duración (en minutos) de las letras de las canciones (década 1960-2020)") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
        panel.background = element_rect(fill = "white", color = NA),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(color = "black"))
grid.arrange(g1, g2, nrow = 2)


#Análisis 4 - Matriz de correlación
install.packages("ggcorrplot")
library(ggcorrplot)


#Quitamos la dicotómicas y popularidad (no) --> queremos ver la relación que guardan las variables
variables_estudiar1 <- c("Danceability", "Energy", "Loudness", "Speechiness", "Acousticness", "Valence", "Instrumentalness")


variables_corr <- database_Spotify[c(variables_estudiar1)]
```

```
matriz_correlacion <- round(cor(variables_corr),1)
p.mat <- cor_pmat(variables_corr)

ggcorrplot(matriz_correlacion, hc.order = TRUE, outline.col = "white") +
  labs(title = "Matriz de correlación entre las variables musicales numéricas") +
  theme(plot.title = element_text(hjust=0.5))

#Comprobar correlaciones
p1 <- ggplot(data = database_Spotify, mapping = aes(x = Danceability, y = Valence)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, formula = y ~ x, color = "red") +
  scale_fill_brewer(palette = "Set3") +
  labs(x = "Danceability", y = "Valence", title = " ")+
  theme(plot.title = element_text(hjust=0.5),
        panel.background = element_rect(fill = "white", color = NA),  # Fondo blanco sin
recuadros
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(color = "black"))
p2 <- ggplot(data = database_Spotify, mapping = aes(x = Energy, y = Loudness)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, formula = y ~ x, color = "red") +
  scale_fill_brewer(palette = "Set3") +
  labs(x = "Energy", y = "Loudness", title = " ")+
  theme(plot.title = element_text(hjust=0.5),
        panel.background = element_rect(fill = "white", color = NA),  # Fondo blanco sin
recuadros
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(color = "black"))

p3 <- ggplot(data = database_Spotify, mapping = aes(x = Loudness, y = Acousticness))
+ geom_point() +
```

```
  geom_smooth(method = "lm", se = FALSE, formula = y ~ x, color = "red") +
  scale_fill_brewer(palette = "Set3") +
  labs(x = "Loudness", y = "Acousticness", title = " ")+
  theme(plot.title = element_text(hjust=0.5),
        panel.background = element_rect(fill = "white", color = NA),  # Fondo blanco sin recuadros
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(color = "black"))
p4 <- ggplot(data = database_Spotify, mapping = aes(x = Acousticness, y = Energy)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, formula = y ~ x, color = "red") +
  scale_fill_brewer(palette = "Set3") +
  labs(x = "Acousticness", y = "Energy", title = " ")+
  theme(plot.title = element_text(hjust=0.5),
        panel.background = element_rect(fill = "white", color = NA),  # Fondo blanco sin recuadros
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(color = "black"))
library(gridExtra)
library(ggplot2)
grid.arrange(p1, p2, p3, p4, ncol = 2)



# Análisis 5: Popularidad VS resto de variables
media_popularidad <- mean(database_Spotify$Popularity)
desv_est_popularidad <- sd(database_Spotify$Popularity)
limite_inferior <- media_popularidad - desv_est_popularidad
limite_superior <- media_popularidad + desv_est_popularidad
grupo_superior <- database_Spotify[database_Spotify$Popularity >= limite_superior, ]
grupo_inferior <- database_Spotify[database_Spotify$Popularity <= limite_inferior, ]
```

```
media_total <- colMeans(database_Spotify[, c("Danceability", "Energy",
"Loudness_Normalizada", "Mode", "Speechiness", "Acousticness",
                                             "Valence", "Instrumentalness",
                                             "Duration", "Explicit")])


options(scipen = 999)
# Media de cada variable para el grupo superior e inferior
media_superior <- colMeans(grupo_superior[, c("Danceability", "Energy",
"Loudness_Normalizada", "Mode", "Speechiness", "Acousticness",
                                              "Valence", "Instrumentalness",
                                              "Duration", "Explicit")])
media_inferior <- colMeans(grupo_inferior[, c("Danceability", "Energy",
"Loudness_Normalizada", "Mode", "Speechiness", "Acousticness",
                                              "Tempo", "Valence", "Instrumentalness",
                                              "Duration", "Explicit")])
tabla_resultados <- data.frame(
  Variable = c("Danceability", "Energy", "Loudness_Normalizada", "Mode",
"Speechiness", "Acousticness",
               "Valence", "Instrumentalness",
               "Duration", "Explicit"),
  Media_Total = media_total,
  Media_Superior = media_superior,
  Media_Inferior = media_inferior
)


#Análisis 6 - artistas
artistas_populares <- popularidad %>%
  arrange(Decada, desc(Popularity_2))

# Filtrar las 10 canciones más populares de cada década
top_songs <- artistas_populares %>%
  group_by(Decada) %>%
  slice_head(n = 10)
```

```
top_artists <- top_songs %>%
  select(Decada, "Artist name", Popularity_2)
recuento_género <- top_artistas %>%
  group_by(Década, Género) %>%
  summarize(recuento = n())
recuento_género$Género <- factor(recuento_género$Género, levels = c("Solista femenina", "Grupo femenino","Dúo musical", "Grupo de música mixto",  "Grupo masculino", "Solista masculino"))
colores <- c("Grupo masculino" = "#3182bd",
             "Solista masculino" = "#08519c",
             "Grupo de música mixto" = "#bdbdbd",
             "Dúo musical" = "#969696",
             "Grupo femenino" = "#006d2c",
             "Solista femenina" = "#238b45")


ggplot(recuento_género, aes(x = Década, y = recuento, fill = Género)) +
  geom_bar(stat = "identity") +
  labs(x = "Década", y = "Número de Artistas del top 10 de cada década", title = "Género de los top 10 artistas más exitosos de cada década (1960-2020)") +
  scale_fill_manual(values = colores) +
  theme_minimal() +
  theme(plot.title = element_text(hjust=0.5),
        panel.background = element_rect(fill = "white", color = NA),  # Fondo blanco sin recuadros
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(color = "black"))
```

2. Extracción de las letras de las canciones de Genius

```
install.packages("glue")
install.packages("spotifyr")
install.packages("pacman")
install.packages("kableExtra")
```

```
install.packages("geniusr")

library(geniusr)

library(spotifyr)

library(tidytext)

library(dplyr)

library(stats)

library(kableExtra)

library(ggplot2)

library(rvest)

library(purrr)

library(readxl)

library(openxlsx)

database_Spotify <- read_excel("C:\\Users\\marta\\OneDrive - Universidad Pontificia Comillas\\TFG ANALYTICS\\Version Final\\database_Spotify.xlsx")

Sys.setenv(GENIUS_API_TOKEN = "0naYY08MAWAt2k4fUvCrXpZhstCewrDhZjS55HeMnKHOa5VRsj3X-eDjmFrNp1sm")

genius_token <- Sys.getenv("GENIUS_API_TOKEN")

#Para generar mi API TOKEN: https://genius.com/api-clients

#1. Coger las canciones de la base de datos

song_title <- database_Spotify$"Song title"

song_artist <- database_Spotify$"Artist name"


#2. Ya tenemos las canciones y los artistas, buscar con la canci?n su URL

num_canciones <- nrow(database_Spotify)

terminos_busqueda <- c(song_title)

resultados <- vector("character", length = num_canciones)

for (i in 1:num_canciones) {


  # Realizar la b?squeda para el t?rmino actual

SongID_URL <- search_song(search_term = terminos_busqueda[i], n_results = 3, access_token = Sys.getenv("GENIUS_API_TOKEN"))
```

```
  # Verificar si hay resultados antes de filtrar
  if (nrow(SongID_URL) > 0) {
    # Filtrar por artistas deseados

    URL <- SongID_URL %>%
      filter(grepl(paste(song_artist, collapse = "|"), SongID_URL$artist_name, ignore.case
= TRUE))

    resultados[i] <- as.character(URL[1, 3])

  } else {

    resultados[i] <- NA  # O alg?n valor que indique la ausencia de resultados
  }
}

song_URLs <- data.frame(resultados)
colnames(song_URLs) <- c("URL_lyrics")
song_URLs$ID <- 1:nrow(song_URLs)

datos_lyrics <- database_Spotify %>% select (ID, "Song title", "Artist name", Decada,
Year)
datos_lyrics <- merge(datos_lyrics,song_URLs, by = "ID", all = TRUE)

#3. Ya tenemos los URLS de las canciones, ahora hay que hacer webscrapping
extraer_texto <- function(url) {
  if (is.na(url)) {
    return("")
  }

  contenido_html <- read_html(url)
  texto_extraido <- contenido_html %>%
```

```
   html_nodes("div.Lyrics__Container-sc-1ynbvzw-1.kUgSbL") %>%  # Ajusta el selector seg?n la estructura de las p?ginas web
   html_text()


  return(texto_extraido)
}

# Aplicar la funci?n a cada URL en el dataframe
letra_decada60 <- datos_lyrics[datos_lyrics$Decada == 1960, ]
resultados_decada60 <- letra_decada60$URL_lyrics %>% map(extraer_texto)
resultados_decada60_limpio <- vector("character", length = 150)
library(stringr)
expresiones_eliminar <- c("oh", "dee", "doo","do", "hah", "huh", "eh", "la-la", "na-na", "hey", "wow", "yeah", "woo", "whoa", "Yeah",
                          "oops", "oh-oh", "ah", "uh-huh","Dee","Doo", "mm-hmm", "yay", "alright", "shh", "hey-hey", "oooh", "Oh","ooh","Ooh","oh", "yay", "huh", "hah", "eh", "hmm", "la-la", "doo-wop","Mmm", "ha", "Do", "Oi")
patron_eliminar <- paste0("\\b", paste(expresiones_eliminar, collapse = "\\b|\\b"), "\\b")
for (i in 1:150) {
  resultados_decada60_limpio [i] <- resultados_decada60[i] %>% paste(sep = " ", collapse=" ") %>% gsub("\\[[^\\]]*\\]", "",. , perl = TRUE) %>%
   gsub("([a-z])([A-Z])", "\\1 \\2", .) %>%
   gsub("[x]", "", .) %>%
   gsub("in'", "ing",.) %>%
   gsub("[^[:alnum:][:space:]']", " ", .) %>%
   gsub("([a-z])([A-Z])", "\\1 \\2", .) %>%
   gsub("\\b\\w{1}\\b", "", .) %>%  # Elimina palabras de una única letra
   str_replace_all(patron_eliminar, "") %>%
   str_trim()
}

letra_decada60$corpus <- as.character(resultados_decada60_limpio)
write.xlsx(letra_decada60, file = "C:\\Users\\marta\\OneDrive - Universidad Pontificia Comillas\\TFG ANALYTICS\\Version Final\\letra_decada60.xlsx")
```

….Este mismo paso se repite para la década de los 70, 80, 90, 00, 10 y 2020, y las guardamos en nuestro PC por si acaso

```
database_letras<- read_excel("C:\\Users\\marta\\OneDrive - Universidad Pontificia Comillas\\TFG ANALYTICS\\Version Final\\database_letras.xlsx")


database_letras <- database_letras %>%
  mutate(corpus = tolower(corpus))
database1 <- database_letras %>%
  select (ID, corpus)
database_master <- merge (database_Spotify, database1, by = "ID", all.y = TRUE)
```

3. Aplicación LDA y análisis de las letras de las canciones. El código para el análisis de las letras de las canciones está basado y adoptado de Velilla (2023)

```
library(lubridate)
library(dplyr)
library(stats)
library(kableExtra)
library(tidytext)
library(spotifyr)
library(ggplot2)
library(rvest)
library(purrr)
library(reshape2)
library(readxl)
library(openxlsx)
library(writexl)


database_master<- read_excel("C:\\Users\\marta\\OneDrive - Universidad Pontificia Comillas\\TFG ANALYTICS\\Version Final\\database_master.xlsx")
database_LDA <- database_master %>%
  select (ID,"Song title", Decada, Year, corpus)
#1. TOCKENIZACION
```

```
library(udpipe)

#Algoritmos capaces de realizar la tokenización de forma eficiente según el idioma del corpus

#Convertimos cada palabra en token, lema y UPOS(categoría gramatical)

ud_model_download <- udpipe_download_model(language = "english") #Traducido de forma manual todas las lyrics distintas al inglés

ud_model <- udpipe_load_model(ud_model_download$file_model)

corpus <-database_LDA$corpus

corpus_tokenizado <- udpipe_annotate(ud_model, x = corpus)

corpus_tokenizado_data_frame <- as.data.frame(corpus_tokenizado)

head(corpus_tokenizado_data_frame)


#Preprocesamiento del texto: eliminamos todo aquello que no aporte valor

#Basicamente nos quedamos con los sustantivos, los adejtivos, los adverbios y los verbos

unique(corpus_tokenizado_df$upos)

corpus_no_PUNCT_SYM_NUM <- subset(corpus_tokenizado_df, (upos %in%

                                        c("NOUN", "ADJ", "ADV")))

#Eliminamos stopwords:

library(quanteda)

library(stopwords)

library(tm)

stopwords <- stopwords("en")

stopwords<-c(stopwords, "eh", "song","music", "one", "uh", "ah", "ha", "yeah", "oh", "baby", "woo", "hey", "la", "na", "doo", "yeah yeah", "oh yeah", "ooh", "la la", "woo hoo", "oh oh", "na na", "yeah baby", "mmm", "mm", "huh", "whoa", "where",  "just", "my","fact", "when", "whatever", "behind", "along", "sha", "s", "da", "bout", "bit", "like", "since", "darl", "when", "who", "music", "may","don","what", "make", "get","ayy", "even", "thing",”eactly", "ono", "why","whoah", "woh", "whoa","woah")


corpus_nostopwords<-subset(corpus_no_PUNCT_SYM_NUM, !lemma %in% c(stopwords))

corpus_nostopwords<-subset(corpus_nostopwords, !token %in% c(stopwords))

corpus_limpio <- corpus_nostopwords %>% filter(nchar(lemma)>= 3)
```

```
#Colocaciones - pedimos los bigramas del corpus
bigramas <- keywords_collocation(corpus_limpio, term = "lemma", group =
                        c("doc_id", "sentence_id"), ngram_max = 2, n_min = 20)

bigramas<- bigramas %>% filter(nchar(left)>= 3)
bigramas<- bigramas %>% filter(nchar(right)>= 3)
bigramas <- bigramas %>%
  filter(left != right)
#Un PMI de 3 es suficiente para indicar que una colocación es buena
bigramas <- bigramas[bigramas$pmi >= 3,]
#Plotear los bigramas ordenándolos de mayor a menor
bigramas$key <- factor(bigramas$keyword, levels = rev(bigramas$keyword)) #Todos
los bigramas
library(lattice)
barchart(key ~ pmi, data = head(subset(bigramas, freq>3), 20), col = "blue",
        main = "Colocaciones con bigramas", xlab = "PMI")
#enriquecer y que no haya duplicación de colocaciones
corpus_limpio$term <- corpus_limpio$lemma
corpus_limpio$term<-txt_recode_ngram(corpus_limpio$lemma, compound
                                = bigramas$keyword, ngram = bigramas$ngram, sep = " ")

corpus_limpio <- subset(corpus_limpio, upos %in% c("NOUN", "ADJ", "ADV"))

library(tm)
library(quanteda)
dtf <- document_term_frequencies(corpus_limpio, document = "doc_id",  term =
"term")
dtm <- document_term_matrix(dtf)
dfm_quanteda<-as.dfm(dtm)

#Eliminamos aquellos términos que no aparecen casi en el documento y lo ponemos en
el correcto formato para topic modeling
```

```
library(topicmodels)
dfm_trimmed<-dfm_trim(dfm_quanteda, min_docfreq = 0.005, max_docfreq = 0.99,
                      docfreq_type="prop")
dtm=convert(dfm_trimmed, to="topicmodels")
dim(dtm)

#Determinar el número de tópicos
library(text2vec)
library(ggplot2)
range<-seq(2, 40, by=2)
tcm=crossprod((as.matrix(dtm)))
coherence_logratio=data.frame()

for(i in range){
  resultsloop_logratio=data.frame()
  for (j in seq(1,5,by=1)){ #sino 3
    topicmodeling <- LDA(dtm, method="Gibbs", k=i,iter=200, burnin=100,
                         control=list(alpha = 50/i, delta=0.1), initialize="random")
    words_topic<-terms(topicmodeling, k=10)
    words_topic_matrix<-as.matrix(words_topic)
    resultsloop_logratio[j,1]=mean(coherence(words_topic_matrix, tcm,
                                   metrics = c("mean_logratio"), smooth=1, n_doc_tcm =
nrow(tcm)))}
  coherence_logratio[i,1]=i
  coherence_logratio[i,2]=mean(resultsloop_logratio$V1)}

ggplot() + geom_line(data= coherence_logratio, aes(x = V1, y = V2), color =
                       'blue', stat="identity", size=1)+labs(x="Número de tópicos",
y="Coherencia
UMass")+theme_minimal()

#Una vez determinado el numero de tópico  podemos empezar LDA
```

```
library(reshape2)
library(tidytext)
library(dplyr)
library(stats)
library(ggplot2)

i=8
modelo <- LDA(dtm, method = "Gibbs", k =i, control = list(alpha = 50/i,
                                        delta=0.1, seed=58), initialize="random")
#Crear los ocho tópicos con la 20 main terms
phi <- data.frame(as.matrix(posterior(modelo)$terms)) #distribución de las palabras dentro de cad tópico
phi$row_name <- (rownames(phi))
topicos <- melt(phi)
colnames(topicos) <- c("topic", "term", "phi")
topicos$topic<-as.numeric(topicos$topic)

top_terms <- topicos %>%
  group_by(topic) %>%
  top_n(20,phi) %>%
  ungroup() %>%
  arrange(topic,-phi)

plot_topic <- top_terms %>%
  mutate(term = reorder_within(term, phi, topic)) %>%
  ggplot(aes(term, phi, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free", ncol = 4) +
  coord_flip() +
  scale_x_reordered() +  theme(axis.text.y = element_text(size = 10))

plot_topic
```

```
#Visualizacion


library(LDAvis)
library(slam)


phi <- as.matrix(posterior(modelo)$terms)
theta <- as.matrix(posterior(modelo)$topics) #distribución de los tópicos por cada uno de los documentos
vocab <- colnames(phi)
doc.length <- as.vector(table(dtm$i))
term.frequency <- col_sums(dtm)
json_lda <- createJSON(phi = phi, theta = theta, vocab = vocab, doc.length =
                       doc.length, term.frequency = term.frequency)
library(servr)
serVis(json_lda)


#Decidir topic names


topicNames<-c("Juventud y Libertad", "Amor Apasionado", "Perseverancia y Lucha", "Diversión y Celebración", "Cruda Realidad Urbana", "Inquietudes del día a día", "Sueños y Aspiraciones", "Reflexiones temporales")


#Visualizacion para ver relevancia de tópicos.


topicproportions <- colSums(theta) / nrow(dtm)
names(topicproportions) <- topicNames
topicproportions<-sort(topicproportions, decreasing = TRUE)
proportions <- data.frame(topic = names(topicproportions), prop =
                          as.numeric(topicproportions))
proportions_sorted <- proportions[order(proportions$prop, decreasing = TRUE), ]


ggplot(proportions, aes(x=prop, y=topic))+
```

```
  geom_bar(stat="identity", fill="blue")+ geom_text(aes(label = round(prop,3)),
                                                    hjust = 1.5,color="white") + labs(title = "Relevancia de
cada tópico dentro del corpus", x = "Proporción", y = "Topic") +
  theme(plot.title = element_text(hjust=0.5))+ theme_minimal()


ggplot(proportions, aes(x = prop, y = factor(topic, levels = topic))) +
  geom_bar(stat = "identity", fill = "blue") +
  geom_text(aes(label = round(prop, 3)), hjust = 1.5, color = "white") +
  labs(title = "Relevancia de cada tópico dentro del corpus", x = "Proporción", y =
"Topic") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))


#Evolución de los tópicos
library(lubridate)
body <- data.frame(Decada= database_master$Decada, corpus =
database_master$corpus)
topic_proportion_per_decade <- aggregate(theta, by = list(Decada = body$Decada),
mean)
colnames(topic_proportion_per_decade)[2:(i+1)] <- topicNames


vizDataFrame <- melt(topic_proportion_per_decade, id.vars = "Decada")
ggplot(vizDataFrame, aes(x = Decada, y = value, group = variable)) +
  geom_line()+
  geom_point()+
  facet_wrap(~ variable, scales = "free_y", ncol=3) +
  theme_minimal()+
  theme_bw() +
  theme(panel.grid.major = element_blank() ,
        panel.grid.minor = element_blank()) +
  theme(plot.title = element_text(hjust=0.5))+
  labs(x= "Década", y ="Valor", title ="Evolución de los tópicos durante las últimas 7
décadas")
```