Big Data

Propuesta de Trabajo Final Borrador

Albareti Renato, López Federico y Mettola Franco Fecha de entrega: 1 de mayo de 2024

1

IDEA 2: Ritmo y Recesión: el efecto de variables macroeconómicas en las preferencias musicales

Introducción:

La pregunta de investigación de este trabajo sería: de qué manera impacta la situación económica del país en el consumo musical de las personas. Vamos a trabajar con la idea de que la música popular es un reflejo del bienestar de las personas y tratar de analizar si, por ejemplo, cuando la economía está peor la gente busca escuchar música más alegre y análisis similares. Sería un interesante aporte a la literatura sobre la interrelación entre la cultura y la economía. El trabajo va a tener dos partes: la primera consiste en la clasificación de las canciones usando clusters y otra usando regresión logística multinomial y regresión lineal.

Base de datos:

*Variables macroeconómicas:*

Con la nueva API del Banco Central, Yahoo Finance y la EPH obtendríamos datos de estas variables: nivel de desempleo, tasa de inflación, tasa de interés, el índice MERVAL.

*Datos musicales:*

Para obtener las canciones que fueron más populares en cada año vamos a tener que utilizar distintas fuentes, ya que, a diferencia de otros países como España o Mexico, solo recientemente se compila un ranking general de las canciones más escuchadas. Estos son los datos que vamos a buscar usando web-scraping y la API de Spotify ya que algunas radios los compilaron en esa aplicación:

desde el 1985 tenemos los Ranking de la Rock and Pop

desde 1991 tenemos el rankings de la radio FM 105.5 (LOS40)

desde 1999 está el Ranking de MTV

desde 2018, Billboard publica el top 100 Argentina

desde 2021 CAPIF publica las canciones mas escuchadas de Argentina

2

Para muchos años hay overlapping y vamos a tratar de combinar los datos de estas distintas fuentes para obtener las 50 canciones mas populares de cada año. De esta manera tendríamos 39 años de datos para utilizar en la segunda etapa. La base de datos sería un diccionario con cada año (key) asociado una lista de canciones las cuales vamos a referenciar con el ’id’ de esa canción en Spotify para que sea mas facil referenciar diferentes datos de esas canciones

Con eso a armar algunas bases de datos: Para la primera vamos a usar la API de Spotify para obtener las caracteristicas de las canciones. En esta base de datos cada observación es una cancion y cada cancion va a tener asociadas estas columnas: *{* .acousticness": , "danceability":, "duration ms": , .energy": , ¨ınstrumentalness": , "key": , "liveness": , "loudness": , "mode": , "speechiness": , "tempo": , "valence":,

"time signature": (esas son las features principales, pero hay muchas mas que descri ben el contenido y la estructura musical más en detalle que también podemos usar) La más importante de estas que vamos a usar es una que cuantifica el ’mood’ (que tan negativa o positiva una canción) con la variable ‘valence‘ que está entre 0 y 1. La idea es que un valor más alto sería una canción más alegre. Con esta vamos a construir una variable ‘average valence‘ por cada a˜no.

Otra base de datos la vamos a armar con web-scraping de páginas web con letras de canciones como genius.com, letras.com y musixmatch.com. Esto nos da más libertad para explorar distintos aspectos de las canciones. Vamos a usar métodos de Natural Language Processing y la librería ‘nltk‘ y ‘vader‘ para clasificar las letras como positivas o negativas.

COMMENT: Este ejercicio puede estar muy interesante para testear cual indicador es mas representativo de la demanda por musica. Las medidas de ritmo mencionada antes versus las medidas de sentimiento de las canciones. En un marco conceptual micro, con tres "types" de demanda, algunos ponderan mas el ritmo, otro mas la letras y otros ponderan por igual ambos.

Podríamos realizar otros análisis con las letras por ejemplo, también construir un índice de agresividad o de obscenidad y de esa manera ver si, quizás, en los momentos de crisis las canciones más escuchadas expresan más o menos sentimiento de enojo, por ejemplo.

Metodología:

*Primera etapa*

Vamos a usar los datos que obtuvimos del análisis de Spotify para realizar un análisis de clusters sobre los datos de análisis musical de las canciones. La API tiene una variable ’genres’ que contiene varios géneros en los que clasifica las canciones. Por ejemplo: *{* ”genres”: [.alternative”, ”samba”]*}*

La idea es tratar de reducir la gran cantidad de géneros que devuelve Spotify usando PCA y Clusters, asi que dadas todas las features: timbre, max vol, etc.. cuales son todos los géneros más comunes en los 6 o 7 clusters que identificamos (hay que decidir cuántos clusters usar).

COMMENT: Acá no me quedó del todo claro si quieren reducir la cantidad de géneros para luego tener una variable dependiente con menos categorías. O si quieren generar nuevos clusters y que eso luego sea usado como v. dependiente

*Segunda etapa*

Vamos a usar dos o tres modelos de logit multinomial y de regresión lineal. COMMENT: Como?

Para seleccionar el modelo de serie de tiempo vamos a elegir rezagos de las variables dependientes económicas usando el criterio AIC y BIC. Un modelo usaría como variable dependiente los generos que obtuvimos

COMMENT:si el tipo de genero musical es la variable de respuesta, no queda claro que eso sea un indicador del "aninmo cultural" en distintos momentos del ciclo economico.

(como son variables categóricas con varias categorías correspondería usar logit multinomial). Esto nos permitiría ver si ciertos generos son más o menos populares en una u otra condicion económica. A este análisis debería incluirsele un control por el efecto del tiempo en el cambio en la popularidad relativa de los géneros

Otro modelo usaría regresión de series tiempo de las varaibles dependeintes y la variable valence que mide la positividad o negatividad de las canciones. Luego otro modelo con algun otro indice que hayamos creado usando sentiment analysis de las letras de las canciones.

COMMENT Estas propuestas son las que efectivamente darian evidencia de la hipotesis empirica mencionada en la motivacion.