Análisis de datos estructurados aplicado a un dataset con la evolución musical en México en un periodo de 71 años (1950-2021), según el ranking Billboard Hot 100.

Proyecto Final - Tópicos Selectos de Ciencia de Datos

Arely Hilda Luis Tiburcio
Tecnologías para la Información en Ciencias
UNAM Campus Morelia
arelyluis@comunidad.unam.mx

ABSTRACT

El siguiente documento realiza un análisis estructurado (grafos) a un conjunto de datos que contiene la evolución musical en un periodo de 71 años (1950-2021), tomando como base el Billboard Hot 100 en México. Para la realización del análisis se tomaron dos atributos en particular: 'Artista/Banda' y 'Año', se realizó una comparación entre estos dos atributos para poder observar la forma en que los artistas o agrupaciones tienen conexiones a lo largo de las decádas de estudio.

A continuación se muestran los hallazgos encontrados así como una descripción de parámetros generales de la red.

KEYWORDS

grafo, red, conexiones, centralidad

ACM Reference Format:
Arely Hilda Luis Tiburcio. 2023. Análisis de datos estructurados aplicado a un dataset con la evolución musical en México en un periodo de 71 años (1950-2021), según el ranking Billboard Hot 100.: Proyecto Final - Tópicos Selectos de Ciencia de Datos. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 DESCRIPCIÓN Y PRE-PROCESAMIENTO DE DATOS

El conjunto de datos a analizar fue creado por mi autoría como parte de mi proyecto de titulación de licenciatura. El conjunto datos se creó con el fin de poder realizar tareas de Procesamiento del Lenguaje Natural, no obstante, dicho dataset puede ser aplicado para realizar distintas tareas interés.

En este documento se busca hacer un análisis, tomando dos atributos: 'Artista/Banda' y 'Año'. La motivación detrás de elegir dichos atributos es la de poder visualizar la relación que los artistas o bandas tienen conforme las décadas. Además de poder aplicar distintos tipos de métricas para obtener información relevante de la red.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, July 2017, Washington, DC, USA
© 2023 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 VISUALIZACIÓN

La forma más simple de poder obtener información de un grafo es mediante la visualización. Es así que como primer tarea se tomaron los dos atributos de interés y mediante el uso de la librería networkx, se creo un grafo simple, creando conexiones entre las canciones con su año de aparición en el ranking.

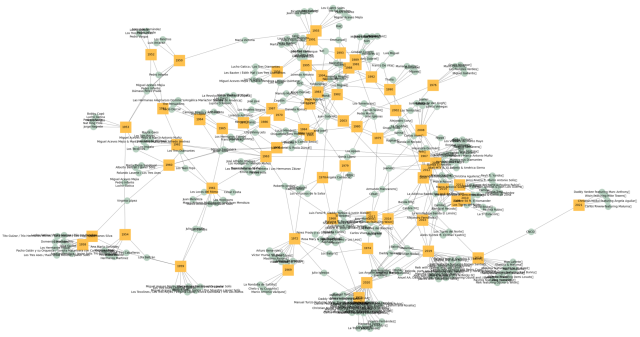

Figure 1: Grafo de la componente conexa mayor de la red.

La red presenta 3 componentes conexas, dos de ellas son muy pequeñas, por lo que se optó por mostrar la componente conexa mayor, la cual contiene la mayor parte de los datos del dataset. Se observan la formación de distintos sectores dónde la cantidad de nodos es mayor.

3 PROPIEDADES GENERALES

Entre las propiedades generales de la red podemos encontrar la siguiente información. La red presenta un tamaño de 505, la cantidad total de nodos es de 391, la cantidad de aristas es de 505 y la red tiene 3 componentes conexas, de las cuales dos son muy pequeñas y solo una es la que abarca la mayor cantidad de los datos.

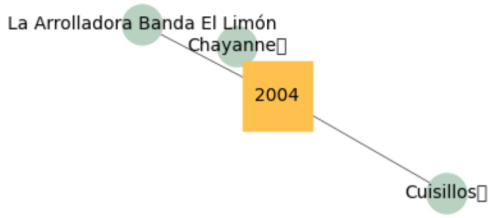

Figure 2: Grafo una de las tres componentes conexas de la red. El nodo del año 2004 presenta una centralidad de 3.

4 CENTRALIDADES

En esta sección se calcularon las medidad de centralidad de la red, dichas medidad fueron: centralidad de grado, centralidad de intermediación, centralidad de cercanía, Kaltz y la excentricidad. A coninuación se muestran algunos de los resultados obtenidos.

Como se ha visto a lo largo del curso, la centralidad de grado es la encaragda de medir el número de enlaces o conexiones que tiene un nodo con los demás nodos pertenecientes a un grafo. Al aplicar esta medida a la red y obtener los 5 mayores resultados podemos observar que el nodo con perteneciente al año 2020 es el que tiene un mayor número de conexiones, seguido por sus dos años anteriores 2019 y 2018. En general los resultados son esperados, ya que en el conjunto de datos existe una mayor cantidad de información en años recientes, lo que resulta interesante es el caso del nodo del año 1961 el cual se encuentra en el último año del top.

Table 1: Top 5 de centralidades de grado

Año	Centralidad de grado
2020	0.06923076923076923
2019	0.05641025641025641
2018	0.038461538461538464
2016	0.03333333333333333
1961	0.03076923076923077

En cuanto a la centralidad de intermediación, se define como una medida que cuantifica la frecuencia o el número de veces que un nodo actúa o sirve de puente dentro de una ruta corta entre dos nodos determinados. Al ver los resultados de los 5 nodos con mayor centralidad de intermediación, podemos notar que el nodo perteneciente al artista Juan Gabriel tiene un valor bastante elevado, esto nos indica a lo largo de la evolución musical de las canciones en México, el artista Juan Gabriel ha pasado por distintas décadas teniendo canciones pertenecientes al ranking ha tenido exitos en distintas décadas, por lo que existen muchas conexiones que sirven como puente para conectar con otros distintos nodos. Del mismo modo los nodos correspondientes a los años 1967 y 1980, así como los de los artistas Raphael y Luis Miguel tienen un alto grado de centralidad de intermediación.

Table 2: Top 5 de nodos centralidad de intermediación

Nodo	Centralidad de intermediación
Juan Gabriel	0.4044553412610729
1967	0.3298192736537483
1980	0.32326521678100434
Raphael	0.3096395397636601
Luis Miguel	0.2852211021213675

La centralidad de cercanía es con la que podemos determinar las rutas más cortas o más eficientes para llegar de un nodo a otro. Nuevamente el nodo correspondiente al artista Juan Gabriel es un nodo muy importante en la red, ya que funciona muy bien como punto para poder crear caminos cortos entre nodos. Es un nodo altamente conectado. También es importante destacar nodos como el de Luis Miguel y el de los nodos pertenecientes a los años 1994, 1995 y 1982.

Table 3: Top 5 de nodos centralidad de cercanía

Nodo	Centralidad de cercanía
Juan Gabriel	0.1602580490096468
1994	0.15717745132705432
Luis Miguel	0.15704619667041378
1995	0.1558746958249652
1982	0.15408670259141377

PageRankfunciona contando la cantidad y la calidad de los enlaces a una página para determinar una estimación aproximada de la importancia del sitio web. La suposición subyacente es que es probable que los sitios web más importantes reciban más enlaces de otros sitios web. En el caso aplicado a la red observamos una leve diferencia a comparación de la centralidad de grado. Siento los tres primeros sitios los mismos, variando solamente los dos últimos puestos que toman los de los nodos de los añoss 1972 y 1967.

Table 4: Top 5 de nodos usando PakeRank

Año	PageRank
2020	0.02656531708013413
2019	0.019826993821230116
2018	0.015303204387396109
1972	0.01179428241745801
1967	0.011694881377899393

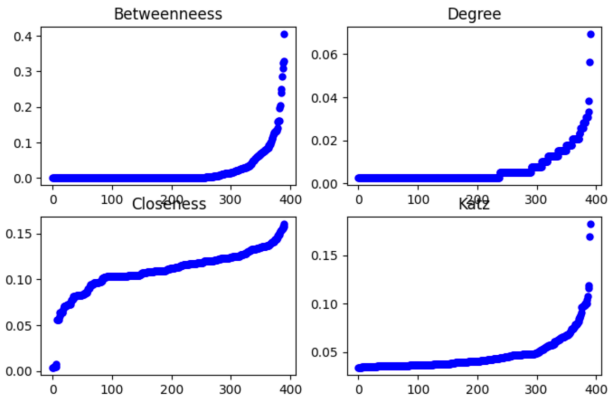

Figure 3: Distribución de las métricas de centralidad.

5 REDES ALEATORIAS

En esta sección se creo una red aleatoría con los mismos parámetros que la red original. Se creo por medio del modelo de Erdős–Rényi. En este modelo se tiene que un nuevo nodo se enlaza con igual probabilidad con el resto de la red, es decir posee una independencia estadística con el resto de nodos de la red. A partir de la creación de esta red aleatoría se compararon distintos aspectos y métricas para analizar y observar si existe una diferencia significativa entre una red creada artificialmente y una red de la vida real.

A continuación se muestran los experimentos realizados:

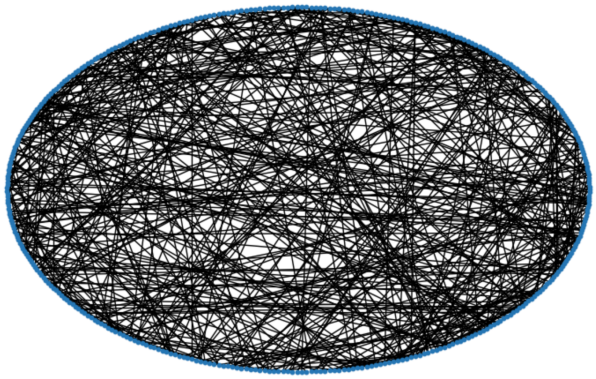

Figure 4: Grafo de una red aleatoria creada usando el modelo de Erdős–Rényi. Con 391 nodos y una probabilidad de conexión de 0.006

Podemos notar que el radio en la red original es mucho mayor en comparación a red aleatoría, es decir, en la red aleatoría el camino mayor para llegar de un nodo a otro es de 16, mientras que en la red original 26.

Table 5: Diametro de las redes

Red Aleaoria	Red Original
16	26

En cuanto a la longitud promedio de la ruta más corta, podemos notar que la red aleatoría tiene la longitud más pequeña con aproximadamente 6, mientras que en la red original es de 9.

Table 6: Longitud promedio de la ruta más corta

Red Aleaoria	Red Original
6.4957105495494	9.16676185813751

Por último con el coeficiente de clustering cuantifica qué tanto está de agrupado (o interconectado) con sus vecinos. En este caso en particular, podemos notar que en la red aleatoria el coeficiente de clustering es alto, sin embargo para la red original el valor es 0. Analizandolo más a detalle y dando contexto a los datos con los que estamos trabajando, se concluye que este valor tiene sentido, ya que los nodos de canciones nunca van a estar nonectados con otro nodo canción, sino que con los nodos de año.

Table 7: Coeficiente de Clustering

Red Aleaoria	Red Original
6.567241940735917	0.0

6 BIBLIOGRAFÍA

Networkx:https://networkx.org/documentation/latest/tutorial.html