Universidad Nacional de Colombia

Facultad de Ciencias
Departamento de Estadística

Proyecto Analisis de Redes Sociales

Autores:

Alejandro Urrego-Lopez
aurrego@unal.edu.co

Cesar Prieto S.
ceprieto@unal.edu.co

6 de julio de 2024

Análisis de Redes Bipartitas en Series de Anime: Impacto de las Descripciones en la Conexión Anime-Anime mediante Bigramas, Clustering de Tópicos y Modelos de Grafos Aleatorios Exponenciales

Resumen

En este artículo se analiza una red que refleja las relaciones entre usuarios y animes, examinándose cómo las descripciones de estos animes influyen en la formación de grandes comunidades de usuarios. En particular, se introduce una nueva variable que cuantifica la frecuencia con la que las palabras de una descripción aparecen en ciertos clústeres de palabras. Estos clústeres se generan a partir del análisis de bigramas derivados de todas las descripciones en la base de datos. Se busca entender mejor la dinámica de estas comunidades y cómo el contenido textual puede afectar la cohesión y estructura de la red social de aficionados al anime. Se concluye que los hallazgos pueden tener implicaciones significativas para el diseño de sistemas de recomendación y la mejora de la experiencia del usuario en plataformas de anime.

Palabras Claves: *Procesamiento de lenguaje natural, Bigramas, Grafos, Red bipartita, Partición de redes , Modelos de grafos aleatorios exponenciales (ERGM).*

Índice

Índice de tablas

Índice de figuras

1. Introduccion

El análisis de redes bipartitas es un método valioso para estudiar las relaciones entre dos conjuntos de nodos, como usuarios y productos. Este enfoque es especialmente útil para plataformas que gestionan interacciones usuario-producto, permitiendo explorar estas relaciones a través de modelos estadísticos y aprovechando toda la información nodal disponible.

Las descripciones de productos son una información crucial que puede ser analizada mediante técnicas de procesamiento de texto, como el análisis de bigramas, que estudia las palabras que aparecen juntas en los textos. Al aplicar diferentes métodos de análisis de redes a estos datos, se pueden explorar las características y patrones que emergen de las descripciones de los productos.

Es evidente que las descripciones de los productos pueden impactar significativamente las redes usuario-producto. Para evaluar la influencia de una variable nodal, se puede analizar la red producto-producto generada a partir de la proyección de la matriz de adyacencia y aplicar modelos de grafos aleatorios exponenciales (ERGM). Los valores p obtenidos de estos modelos permiten observar la significancia de la influencia de estas descripciones.

1.1. Objetivo General

Evaluar la influencia de la frecuencia de palabras asociadas a temas en los grupos de palabras resultantes del análisis de bigramas en las descripciones de animes, en la formación de grandes comunidades de usuarios.

1.2. Justificación

La relevancia de este estudio radica en su potencial para mejorar los sistemas de recomendación en plataformas de entretenimiento, específicamente en el ámbito de los animes. Al entender cómo las descripciones de los productos (animes) influyen en las redes usuario-producto, se pueden desarrollar algoritmos de recomendación más precisos y personalizados. Además, se obtendrán indicios sobre cómo promocionar un anime mediante su descripción para atraer a una audiencia más amplia.

1.3. Alcance

Este estudio se basa en un conjunto de datos de *MyAnimeList*[1] que contiene información sobre 17,562 animes y las preferencias de 325,772 usuarios. Dado que el coste computacional de analizar una red bipartita compuesta por la totalidad de los usuarios y animes era considerablemente alto, se tomó la iniciativa de recurrir a un muestreo en la bases de datos de Animes. Este enfoque busca reducir el coste computacional y permitir la aplicación inicial de la metodología propuesta en una muestra más pequeña, con el objetivo de evaluar la viabilidad de su aplicación a una escala mayor o en su totalidad.

El muestreo se realizó utilizando un Muestreo Aleatorio Simple (MAS). Para la base de los animes, el muestreo tuvo en cuenta las categorías a las cuales pertenecen, realizando el muestreo dentro de cada categoría. Este enfoque evita tratar de manera equitativa categorías con más de 2000 animes listados y aquellas con solo 100, garantizando que no se excluya información valiosa para la red, para el análisis de texto propuesto se tomaron la totalidad de los Animes dado que esto no representaba un reto a nivel computacional.

Se excluyeron los animes que no contaban con una descripción adecuada, ya que la calidad y la integridad de las descripciones son cruciales para el análisis. Animes con descripciones incompletas o inexactas podrían introducir sesgos y afectar la precisión de los resultados. El análisis se limitará a observar la influencia de las descripciones textuales en la formación de grandes comunidades de usuarios, examinando cómo las palabras y frases utilizadas en estas descripciones pueden contribuir a la creación de clústeres de usuarios con intereses similares. Este enfoque permitirá evaluar la relevancia de las descripciones en la dinámica de agrupación y entender mejor los factores textuales que potencian la cohesión dentro de la red de usuarios.

[1] `https://github.com/Hernan4444/MyAnimeList-Database`

2. Estado del arte

En el contexto de los sistemas de recomendación, el análisis de redes bipartitas es una herramienta fundamental. Frecuentemente, se utiliza la matriz de adyacencia como entrada para algoritmos de aprendizaje no supervisado, permitiendo modelar las interacciones entre usuario y producto de manera efectiva. Este enfoque ha sido destacado en diversos estudios debido a su facilidad de implementación y eficacia en la representación de relaciones complejas [**lee 2015**; **chen 2013**].

Paralelamente, el uso de los n-gramas en algoritmos de aprendizaje supervisado ha demostrado ser útil para tareas de clasificación basadas en texto. Por ejemplo, en el análisis del nivel de Burnout en académicos universitarios, se ha utilizado el reconocimiento del lenguaje natural y la clasificación automática basada en n-gramas para identificar patrones de estrés [**abbe 2018**]. Este enfoque se centra en la relación entre los n-gramas y una variable específica, proporcionando información valiosa pero limitada para el análisis de redes más complejas.

Este artículo propone una metodología que combina ambos enfoques: el análisis de enlaces en una red discretizada resultante de la proyección de una matriz bipartita , y el agrupamiento de bigramas para crear una nueva variable asociada. Esta combinación permite una mejor compresión de las relaciones en la red y mejora la efectividad de los sistemas de recomendación [**Bedi 2018**; **Shimomura 2021**].

El análisis de redes sociales, como se observa en la caracterización del discurso de posesión presidencial y la identificación de comunidades políticas en Colombia, ha demostrado ser una herramienta poderosa para identificar términos frecuentes y examinar la cohesión en los discursos [**Luque 2024**]. De manera similar, la modelación de datos de calificación esparcidos como un grafo bipartito ponderado, utilizando medidas de similitud basadas en la entropía, ha mostrado mejorar la calidad de las recomendaciones al aprovechar las propiedades del grafo [**Bedi 2018**]. .

Al integrar estos enfoques, este artículo presenta una metodología innovadora que promete avances significativos en la precisión y aplicabilidad de los sistemas de recomendación, ofreciendo una perspectiva más holística de las interacciones entre usuarios y productos. Esta metodología no solo mejora la comprensión de las relaciones en la red, sino que también potencia la efectividad de los sistemas de recomendación, combinando el análisis de texto y redes complejas para proporcionar recomendaciones más precisas y relevantes [**lee 2015**; **chen 2013**; **Bedi 2018**; **Shimomura 2021**].

3. Metodología

3.1. Marco teórico

Muestreo Aleatorio Simple (MAS)

El Muestreo Aleatorio Simple (MAS) es una técnica de selección de muestras que garantiza que cada elemento de la población tenga la misma probabilidad de ser seleccionado. Es fundamental en la investigación estadística para obtener muestras representativas de una población más amplia.

El Proceso para el Muestreo Aleatorio Simple se describe a continuación:

1. **Definición de la Población:** Se define claramente la población objetivo de interés. La población puede ser cualquier grupo de elementos que compartan una característica común, como personas, objetos, eventos, etc.

2. **Determinación del Tamaño de la Muestra:** Para realizar un MAS en una población finita, se utiliza la siguiente fórmula para determinar el tamaño de la muestra n:

$$n = \frac{Z^2 \cdot p \cdot q \cdot N}{(N-1) \cdot E^2 + Z^2 \cdot p \cdot q}$$

Donde:

- n: Tamaño de la muestra.
- N: Tamaño total de la población.
- Z: Valor crítico de la distribución normal estándar que corresponde al nivel de confianza deseado.
- p: Proporción estimada de la población que tiene la característica específica de interés.
- $q = 1 - p$: Proporción estimada de la población que no tiene la característica específica.
- E: Margen de error permitido en la estimación de la proporción p.

Procesamiento de lenguaje natural: El procesamiento del lenguaje natural (PLN) combina la lingüística computacional, que se basa en reglas para modelar el lenguaje humano, con modelos estadísticos y de machine learning. Esto permite que computadoras y dispositivos digitales puedan reconocer, comprender y generar texto y voz.

Bigramas: Transformación de un texto en unidades conformadas por dos palabras consecutivas para su análisis.

Grafo: Un grafo $G = (V, E)$ es una estructura que consiste de un conjunto de *vértices* (nodos) V y de un conjunto de *aristas* (enlaces) E, donde los elementos de E son parejas de la forma $e = \{u, v\}$, con $u, v \in V$.

Red bipartita: Una red bipartita, también conocida como red de dos modos en la literatura sociológica, es una red con dos tipos de nodos y conexiones que solo se dan entre nodos de diferentes tipos. Las redes bipartitas se usan comúnmente para representar la pertenencia de un conjunto de personas u objetos a grupos específicos. Las personas se representan con un conjunto de nodos, los grupos con otro, y las conexiones unen a las personas con los grupos a los que pertenecen.

Partición de redes: La partición de redes, también conocida como detección de comunidades, es una técnica no supervisada para identificar subconjuntos de vértices que son "homogéneos" en función de sus patrones de relación.

Los algoritmos de agrupamiento de grafos crean una partición $C = \{C_1, \ldots, C_K\}$ del conjunto de vértices V de un grafo $G = (V, E)$. Esta partición se realiza de manera que el número de aristas que conectan vértices de C_k con vértices de C_ℓ sea relativamente pequeño en comparación con el número de aristas que conectan vértices dentro de C_k.

Modelos de grafos aleatorios exponenciales: Los modelos de grafos aleatorios exponenciales (ERGMs), o modelos p^*, se especifican de manera análoga a los modelos lineales generalizados (GLMs). Estos modelos describen la probabilidad condicional de una matriz de adyacencia aleatoria $Y = [Y_{ij}]$ para una red binaria simple no dirigida como:

$$p(y \mid \theta) = \frac{1}{\kappa(\theta)} \exp\{\theta^\top g(y)\}$$

donde:

- $y = [y_{ij}]$ representa una realización de la matriz de adyacencia.
- $g(y) = [g_1(y), \ldots, g_K(y)]^\top$ es un vector de estadísticos de y (variables endógenas) y/o funciones conocidas de y y atributos nodales x (variables exógenas).
- $\theta = [\theta_1, \ldots, \theta_K]^\top$ es un vector de parámetros desconocidos.
- $\kappa(\theta) = \sum_y \exp\{\theta^\top g(y)\}$ es la constante de normalización.

Los coeficientes θ representan la magnitud y dirección de los efectos de $g(y)$ sobre la probabilidad de observar la red.

La probabilidad de observar una arista entre dos vértices i y j, condicionada al resto de la red, se expresa en términos logit como:

$$\text{logit } \Pr(y_{ij} = 1 \mid y_{-(i,j)}) = \theta^\top \delta_{ij}(y)$$

donde:

- $y_{-(i,j)}$ es la matriz y excluyendo la entrada y_{ij}.
- $\delta_{ij}(y)$ es la estadística de cambio que se calcula como $g(y)$ cuando $y_{ij} = 1$ menos $g(y)$ cuando $y_{ij} = 0$, manteniendo constantes los demás valores de y.

Los coeficientes θ se interpretan como la contribución a la probabilidad (en escala logit) de observar una arista particular, condicionado a que todas las otras conexiones se mantengan constantes.

3.2. Metodo

El procedimiento realizado durante esta investigación es el siguiente:

Se utilizaron los software de R y Python para analizar las bases de datos, con la librería `dplyr` de R y `pandas` de Python. Se obtuvo la correspondiente matriz de aristas de los usuarios y los animes, y los bigramas de todas las descripciones. Posteriormente, con la librería `igraph`, se analizó el grafo inducido por estos. Una vez obtenidos los bigramas y el grafo inducido por estos, se selecciona un umbral para dejar los bigramas que aparecen con mayor frecuencia. Posteriormente, se hace una partición de la red con diferentes métodos para elegir el que nos dé una modularidad mayor. Luego, se cuenta la frecuencia de cada descripción de anime con la que alguna palabra aparece en las particiones para crear una variable asociada a cada anime. También se utiliza inteligencia artificial para analizar cada grupo de palabras y asignarle algún significado que pueda deslumbrar los tópicos más importantes tratados en las descripciones.

Para la matriz de aristas, se sigue el mismo procedimiento, pero esta vez se hace la respectiva proyección para analizar la red anime-anime. Después de esto, se binariza la red, es decir, se quitan los pesos y se dejan los enlaces que representen al menos un 75 % de la audiencia de alguno de los dos animes. El principal motivo de esto es para dejar las comunidades más grandes correspondientes a cada uno de estos. Luego, se asignan las variables obtenidas con el análisis de las descripciones mediante bigramas para finalmente ver su significancia en la probabilidad de que se cree una comunidad grande entre dos animes.

4. Aplicación

4.1. Descripción de los Datos

4.1.1. Descripción de animes:

Para analizar las descripciones de los animes, se concatenaron todas las descripciones en la base de datos, separándolas por un carácter especial para evitar la creación de bigramas inexistentes. Posteriormente, se realizó la respectiva tokenización, donde cada token corresponde a una palabra de las descripciones concatenadas, y se eliminaron las palabras vacías. Finalmente, se normalizó el texto eliminando los números, sin eliminar los acentos debido a que las descripciones están en inglés.

Tokens mas frecuentes:

Los tokens más frecuentes que aprecen en la tabla 1 ofrecen una visión sobre los temas y estructuras comunes

	token	frecuencia
1	world	3147
2	one	3127
3	school	2601
4	new	2537
5	life	2142
6	ann	2036
7	however	1967
8	girl	1883
9	story	1878
10	series	1836

Tabla 1: Frecuencias de tokens

en las descripciones de animes. Palabras como “world”, “school”, “life”, y “girl” reflejan los contextos y personajes recurrentes, mientras que términos como “new”, “one”, y “however” revelan elementos narrativos típicos de las tramas.

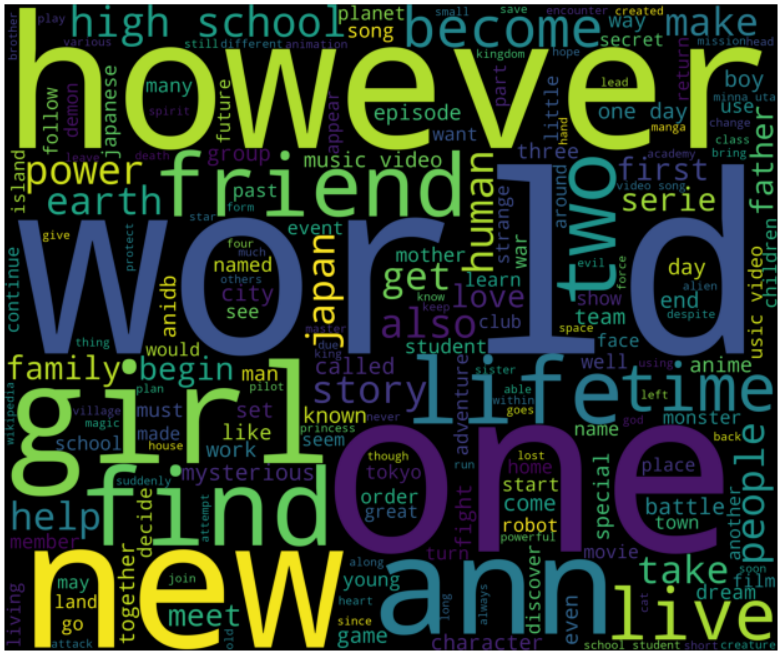


Figura 1: Nube de palabras

Con la nube de palabras nos da una mejor idea de los tokens mas frecuentes a lo largo de las descripciones además permite identificar rápidamente los temas y elementos más comunes en las descripciones de animes. Las palabras más destacadas reflejan una combinación de contextos escolares, relaciones personales, mundos ficticios, y elementos de aventura y fantasía.

Bigramas: Para realizar los bigramas se tomaron dos palabras consecutivas. Se eliminaron aquellos que contenían el carácter especial que separaba las descripciones y la palabra “source”, que aparecía en varias descripciones indicando quién había escrito o de dónde se había obtenido la reseña.

Bigramas mas frecuentes: El análisis de estos bigramas revelan patrones y temas comunes en las descripciones

Bigrama	Frecuencia
high school	1012
one day	719
music video	636
video song	397
minna uta	302
school student	280
nhks minna	255
uta program	243
second season	242
tv series	241
years ago	210
mal news	209

Tabla 2: Bigramas más frecuentes

de animes. Los entornos escolares, la música, los contextos históricos, y la naturaleza serializada de las series son elementos destacados.

Red de Bigramas Para construir la red, se seleccionaron los bigramas con una frecuencia superior a un umbral de 20. Luego, esta red se hizo no dirigida y se analizó con Python y `igrpah`.

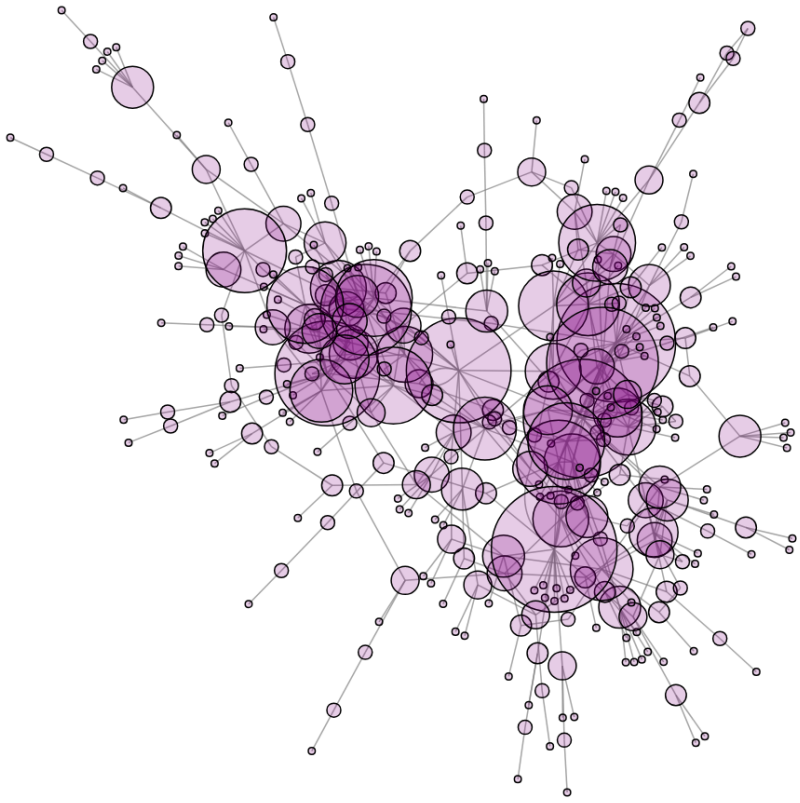


Figura 2: Red de bigramas componente conexa

Métrica	Valor
Distancia media	5.295
Grado medio	2.326
Grado desviación	0.043
Número clan	3
Densidad	0.004
Transitividad	0.048
Asortatividad	0.043

Tabla 3: Métricas de la red de bigramas

La componente conexa de la red de bigramas revela una estructura compleja con varios nodos clave que actúan como puntos centrales en la red. El grado y la presencia de clanes indican que existen temas y palabras recurrentes en las descripciones de los animes. La baja densidad y transitividad sugieren una estructura dispersa pero conectada, con ciertos nodos actuando como hubs en la red.

Grupos de palabras: Para hacer la partición de la red y encontrar sus respectivos grupos de palabras, se emplearon varios algoritmos de agrupación jerárquica implementados en `igraph`, y se escogió el que tuviera una mayor modularidad.

Algoritmo	Modularidad
edge betweenness	1.000
infomap	0.657
label propagation	0.652
leading eigenvector	0.690

Tabla 4: Modularidad de diferentes algoritmos de agrupación

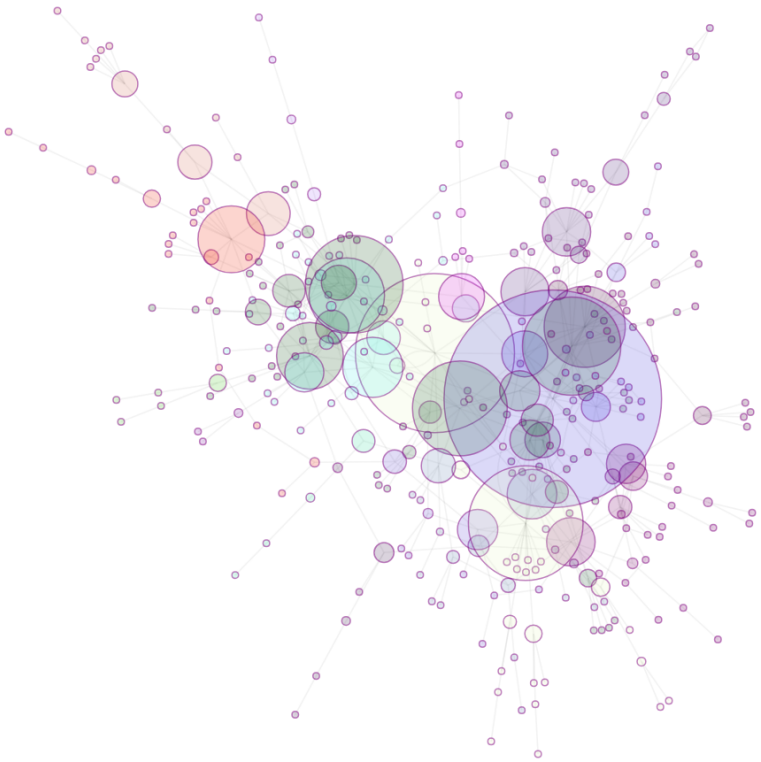


Figura 3: Grupos de palabras obtenidos con el metodo edge betweenness

Se obtienen 18 grupos de palabras. Con estos grupos se crean 18 variables en la base de datos Ponderada, las cuales cuentan la frecuencia con la que aparece una palabra de la descripción en cada uno de los 18 grupos de palabras, para utilizarlas posteriormente en el análisis de la red bipartita mediante el ERGM. Adicionalmente, con ayuda de inteligencia artificial, se analizan las palabras por grupo y se intenta dar un significado a cada grupo, obteniendo que:

Cluster	**Tópico Principal**
Cluster 0	Lanzamientos de episodios y medios
Cluster 1	Relaciones y experiencias personales
Cluster 2	Música y medíos digitales
Cluster 3	Temporalidad y distancias
Cluster 4	Personajes y relaciones familiares
Cluster 5	Vídeos promocionales y juegos
Cluster 6	Narrativas épicas y mundos fantásticos
Cluster 7	Viajes y misiones
Cluster 8	Vida escolar y cotidiana
Cluster 9	Adaptaciones y formatos de medios
Cluster 10	Amistad y relaciones escolares
Cluster 11	Diseño y desarrollo de personajes
Cluster 12	Participación y decisiones
Cluster 13	Información pendiente
Cluster 14	Contextos históricos y de ciencia ficción
Cluster 15	Técnicas de animación y estudios
Cluster 16	Libros infantiles y dramas
Cluster 17	Bandas y música tradicional

Tabla 5: Temas o Tópicos Principales por Clústeres

Se puede encontrar un análisis más detallado por cluster en el apéndice de este documento, donde se detallan las palabras claves y se hace una extinción de la posible temática tratada (véase el Apéndice A).

4.2. Red bipartita:

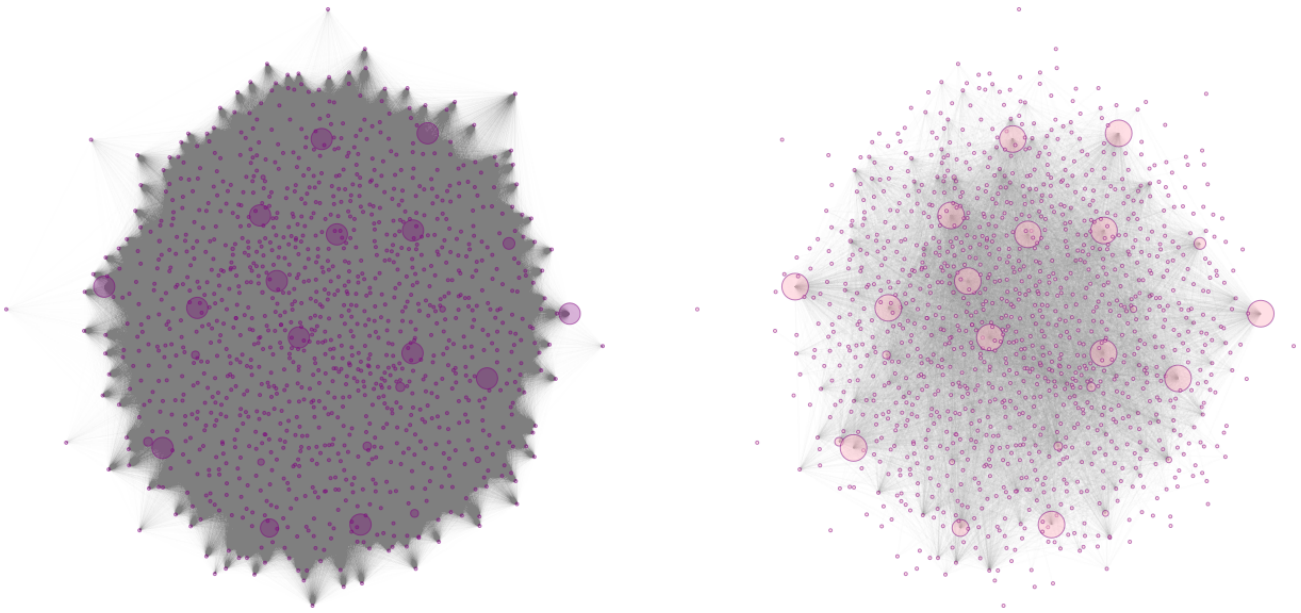
Para conformar la red bipartita en primer lugar se tuvo que recurrir a un muestreo en la base de Animes, el método para elección de la muestra propuesto se centro en realizar un muestreo aleatorio simple dentro de los datos de animes obtenidos para el análisis, el principal objetivo fue obtener una muestra más pequeña pero representativa y reducir el costo computacional del proyecto.

En primer lugar, se llevó a cabo la limpieza y preparación de los datos para el respectivo muestreo, incluyendo la búsqueda de valores faltantes o repetidos, así como la identificación de animes categorizados para una audiencia adulta. Posteriormente, se extrajeron aquellos géneros considerados inapropiados o irrelevantes para el análisis propuesto. Además, se observó la cantidad de apariciones de cada género y, con base en ello, se decidió trabajar únicamente con los géneros que presentaban una aparición mayor a 100 en el conteo. Esto no implicó la eliminación de los animes pertenecientes a esos géneros, ya que un anime puede pertenecer a más de un género simultáneamente; simplemente se excluyó la categorización en los géneros con menos de 100 apariciones.

Finalmente, para completar la preparación de los datos, se decidió descomponer la variable de género en indicadores para cada género, permitiendo así el muestreo dentro de cada uno y no en su totalidad, lo que garantizó la participación de todos los géneros en el análisis final. Posteriormente, se determinó el tamaño de muestra necesario para cada género, teniendo en cuenta la variabilidad de las puntuaciones obtenidas. Animes de cada género fueron seleccionados aleatoriamente según el tamaño de muestra calculado, asegurando que las muestras fueran estadísticamente significativas y equilibradas. El resultado de este procedimiento fue la obtención de una base de animes con una cantidad muy reducida dentro de cada género.

Esta base resultante del muestreo fue la que se utilizo en conjunto con la base de Usuarios para la creación de la red bipartita con la cual se continuo el desarrollo del proyecto

Ahora, la red bipartita cuenta con 8.710.328 enlaces, los cuales representan si un usuario vio alguno de los animes resultantes del muestreo. Posteriormente, se hace la proyección de la red, es decir, se toma la matriz de adyacencia A y se premultiplica por su transpuesta. Dado que A en sus filas contiene los usuarios que hay en la base de datos y en sus columnas los animes de la misma, se obtendrá una matriz de dimensiones 1264 x 1264 en la que cada entrada representa el número de usuarios que han visto estos dos animes. Por otro lado, para poder aplicar el ERGM, se binarizó la red con el criterio de convertir en uno cada entrada de la matriz de proyección si este número es mayor o igual al 75 % de los usuarios que vieron el anime. Esto es sobre todo para garantizar que varios usuarios han visto ambos animes a la vez



(a) Red anime- anime

(b) Red anime- anime binarizada

Figura 4: Red anime - anime decorada con centralidad de intermediacion comparación

Se puede apreciar que la red ponderada tiene una densidad mucho mayor que la red sin pesos, ya que, aunque el nivel de transparencia de los enlaces esté al máximo, se nota que hay una gran cantidad de los mismos.

Top 5 Animes según Métricas de Centralidad sin pesos

Name	Closeness	Betweenness	Eigenvector	Red
Mahou Shoujo Madoka Magica	1.00	-	1.00	
Bishoujo Senshi Sailor Moon	1.00	-	1.00	
Charlotte	1.00	-	1.00	
Shinmai Maou no Testament	1.00	69.58	-	
Boku no Hero Academia 4th Season	-	69.58	-	No binarizada
Digimon Tamers	-	69.58	-	
Pokemon	-	69.58	-	
Yakusoku no Neverland	-	-	1.00	
Kimi no Na wa.	-	-	1.00	
One Piece Movie 5: Norowareta Seiken	0.95	64,189.11	0.99	
Kara no Kyoukai 1: Fukan Fuukei	0.90	36,944.35	1.00	
Natsume Yuujinchou San	0.80	15,552.56	0.98	
Zombie-Loan	0.79	14,127.91	0.97	binarizada
Trigun	0.72	-	-	
One Piece: Episode of Nami	-	6,830.04	-	
Neon Genesis Evangelion	-	-	0.95	

Tabla 6: Top 5 animes según métricas de centralidad sin pesos

Top 5 Animes según Métricas de Centralidad con pesos

Name	Closeness	Betweenness	Eigenvector
Little Village People	0.99	402,989.42	-
Baolie Feiche II: Xing Neng Juexing	0.81	141,314.74	-
Baolie Feiche 3: Shou Shen Heti	0.81	141,314.74	-
Kouya no Kotobuki Hikoutai Kanzenban	0.59	16,990.44	-
Closers: Side Blacklambs	0.56	19,071.02	-
Death Note	-	-	1.00
Sen to Chihiro no Kamikakushi	-	-	0.78
Elfen Lied	-	-	0.78
Kimi no Na wa.	-	-	0.77
Kiseijuu: Sei no Kakuritsu	-	-	0.70

Tabla 7: Top 5 animes según métricas de centralidad con pesos

La comparación entre la red binarizada como la no binarizada muestra cómo las métricas de centralidad pueden variar significativamente dependiendo de la consideración del peso de las conexiones. Mientras que ciertos animes como “Elfen Lied” y “Kimi no Na wa.” mantienen una alta centralidad en ambas forma de medir la centralidad, otros como “Little Village People” y “Death Note” destacan más en las medidas de centralidad que tienen en cuenta el peso de los enlaces.

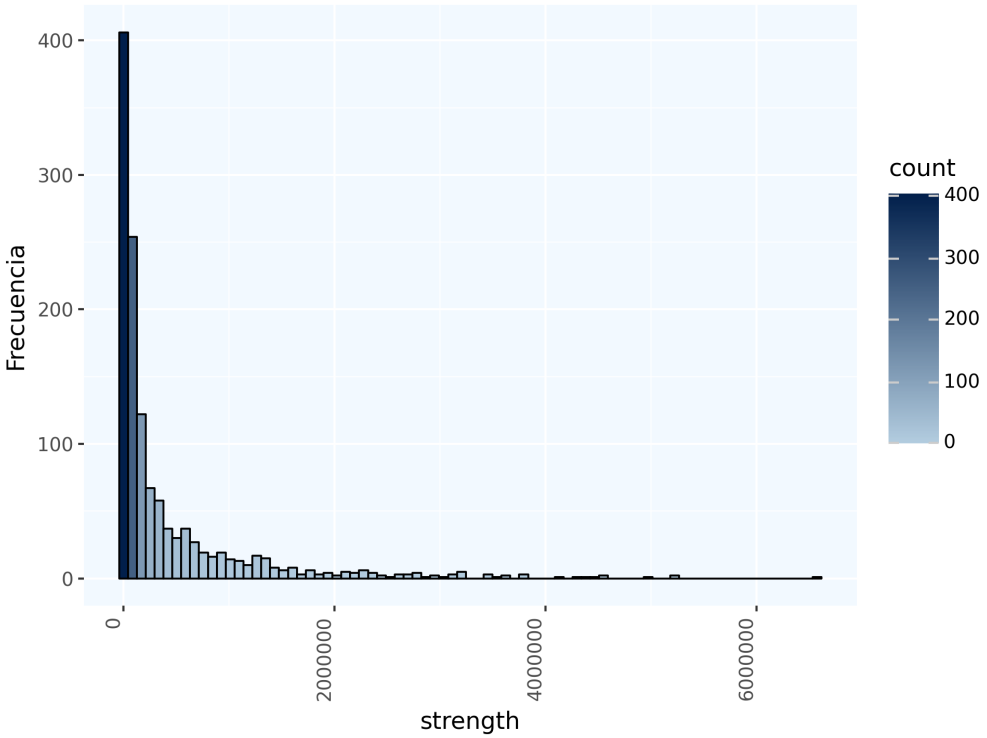


Figura 5: Fuerza red anime anime ponderada

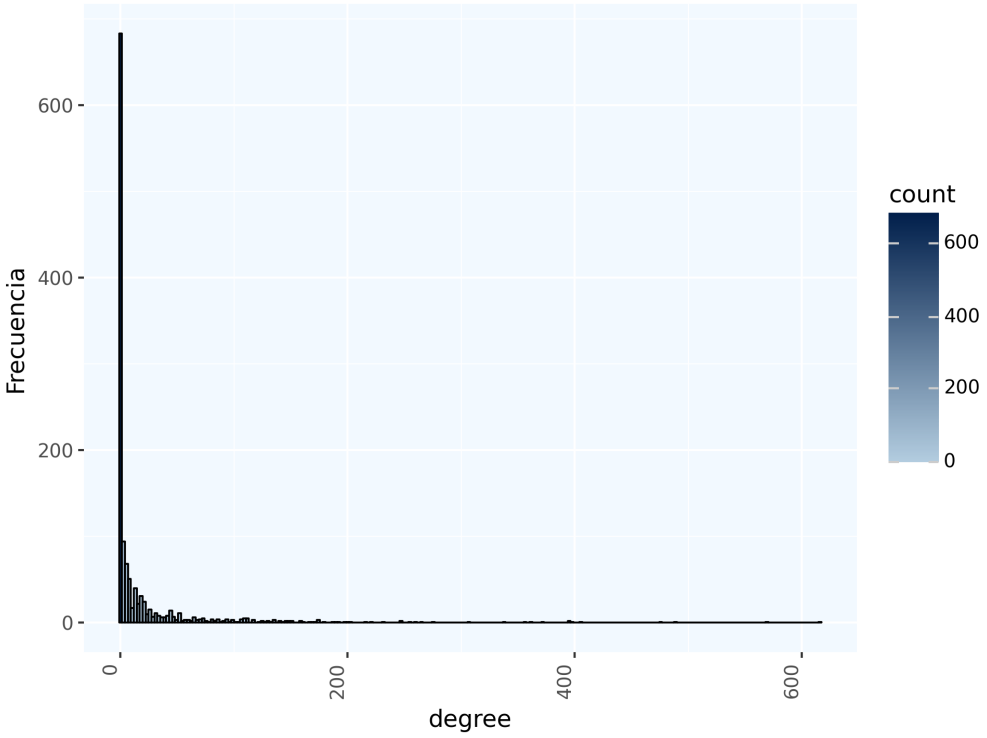


Figura 6: Grado red anime anime no ponderada

Ambos gráficos presentan distribuciones altamente sesgadas hacia la derecha, lo cual es común en muchas redes. Esto indica que la mayoría de los nodos tienen pocas conexiones o conexiones débiles, mientras que unos pocos nodos tienen muchas conexiones o conexiones muy fuertes.

Algoritmo	No binarizada	No binarizada sin pesos	binarizada
Fastgreedy	0.001	0.083	0.186
Leading Eigenvector	0.001	0.084	0.198

Tabla 8: Modularidad de la red de anime según distintos algoritmos

La red binarizada presenta una mejor definición de comunidades, especialmente con los algoritmos Fastgreedy y Leading Eigenvector,

4.3. Análisis de Datos y Resultados

Cluster	Estimado	Error Estándar	p-valor
edges	-1.6084768	0.0065582	<1e-04
cluster 0	-0.0082617	0.0012566	<1e-04
cluster 1	-0.0100252	0.0007925	<1e-04
cluster 2	0.0157385	0.0025446	<1e-04
cluster 3	-0.0136692	0.0016403	<1e-04
cluster 4	-0.0119378	0.0011925	<1e-04
cluster 5	-0.0103690	0.0010620	<1e-04
cluster 6	0.0023567	0.0008391	0.004977
cluster 7	-0.0192175	0.0012180	<1e-04
cluster 8	0.0106841	0.0009673	<1e-04
cluster 9	0.0074773	0.0008841	<1e-04
cluster 10	0.0048164	0.0013969	0.000565
cluster 11	0.0180805	0.0011363	<1e-04
cluster 12	0.0235229	0.0015399	<1e-04
cluster 13	0.0191817	0.0017804	<1e-04
cluster 14	-0.1334576	0.0042158	<1e-04
cluster 15	0.0910541	0.0039146	<1e-04
cluster 16	0.0154359	0.0016520	<1e-04
cluster 17	-0.0254279	0.0020949	<1e-04

Tabla 9: Influencia de los Clusters en la Formación de Enlaces

Al ajustar un modelo de grafos aleatorios exponenciales (ERGM) a la red analizada, se obtuvieron resultados significativos en cuanto a la influencia de ciertos grupos de palabras (clusters) en la formación de enlaces. Los parámetros estimados y sus respectivas pruebas de significancia estadística proporcionan una visión clara sobre qué clusters afectan positivamente o negativamente la creación de conexiones dentro de la red.

La significancia de los coeficientes de los clusters sugiere que la frecuencia de ciertos grupos de palabras tiene una notable influencia en la formación de enlaces en la red. Clusters como el 2, 8 y 15 contribuyen a la creación de enlaces, mientras que clusters como el 0, 1 y 14 disuaden la formación de estos. Esto apoya la importancia del análisis de bigramas realizado previamente, destacando cómo ciertos patrones de palabras facilitan o inhiben las conexiones dentro de la red. Finalmente, la bondad y ajuste del modelo capturan adecuadamente la naturaleza de los clusters definidos, haciendo que las conclusiones derivadas de este modelo sean apropiadas.

5. Discusión

Conclusiones

El presente estudio demuestra que el análisis de redes bipartitas, en combinación con técnicas de procesamiento de lenguaje natural, es una herramienta eficaz para comprender la influencia de las descripciones textuales en la formación de comunidades de usuarios de anime. Los modelos de grafos aleatorios exponenciales (ERGM) utilizados

permitieron identificar cómo ciertos clústeres de palabras impactan de manera significativa en la creación de enlaces entre animes.
Las conclusiones principales del análisis incluyen:

- Los grupos de palabras como el clúster 2 (Música y medios digitales), clúster 6 (Narrativas épicas y mundos fantásticos), clúster 8 (Vida escolar y cotidiana), clúster 9 (Adaptaciones), clúster 10 (Amistad y relaciones escolares), clúster 11 (Diseño y desarrollo de personajes), clúster 12 (Participación y decisiones), clúster 13 (Técnica de animación), clúster 15 y clúster 16 (Libros infantiles y dramas) mostraron tener una influencia significativa positiva. Esto indica que mientras más palabras relacionadas con este contexto se utilicen, aumenta la probabilidad en escala logarítmica de que surja un enlace entre dos animes.
- Los grupos de palabras como el clúster 0 (Lanzamiento de episodios), clúster 1 (Relaciones y experiencias personales), clúster 3 (Temporalidad), clúster 4 (Relaciones familiares), clúster 5 (Videos promocionales), clúster 7 (Viajes y misiones), clúster 14 (Ciencia ficción y contextos históricos) y clúster 17 (Bandas y música tradicional) mostraron tener una influencia significativa, aunque negativa. Esto sugiere que mientras más palabras relacionadas con este contexto se utilicen, reduce la probabilidad en escala logarítmica de que surja un enlace entre dos animes.
- La inclusión de descripciones detalladas y bien estructuradas en las plataformas de anime puede potenciar la formación de comunidades de usuarios, mejorando la experiencia general y facilitando recomendaciones más precisas.

Recomendaciones

Basado en los hallazgos del estudio, se sugieren las siguientes recomendaciones:

- **Optimización de Descripciones:** Las plataformas de anime deberían considerar la optimización de las descripciones de los animes para incluir términos y frases que han demostrado tener una influencia positiva en la creación de enlaces. Esto podría mejorar la cohesión de la comunidad y la interacción entre los usuarios.
- **Mejora de Sistemas de Recomendación:** Integrar el análisis de clústeres de palabras en los sistemas de recomendación puede aumentar la precisión de las recomendaciones, sugiriendo animes que no solo sean populares, sino también relevantes según las descripciones textuales.
- **Investigaciones Futuras:** Es recomendable realizar estudios adicionales que amplíen este análisis a una mayor cantidad de datos y otros géneros de entretenimiento, para validar la aplicabilidad de la metodología en diferentes contextos.
- **Establecimiento del Tema de los Clústeres:** Al momento de establecer un tema relacionado con los clústeres de palabras, resulta difícil dado que las palabras que hay en estos no parecen muy relacionadas.

Finalmente, este estudio destaca la importancia del contenido textual en la formación de redes de usuarios y sugiere que un enfoque combinado de análisis de texto y grafos puede proporcionar valiosas perspectivas para mejorar la experiencia del usuario en plataformas de anime y más allá.

A. Análisis detallado de tópicos por Clusters

A continuación se presenta el análisis de los temas principales para cada uno de los clusters obtenidos a partir del análisis de texto de las descripciones de animes.

- **Cluster 0: Lanzamientos de episodios y medios**
  - **Palabras clave**: episodes, episode, included, later, cap, aired, set, specials, tv, special, manga, bundled, unaired, dvd, released, powers, bluray, final, volume, box, bluraydvd, also, bonus, bddvd, show, birthday, anniversary, name, artist, limited, known, release, volumes, releases, supernatural, battle, edition
  - **Tema**: Este cluster se centra en la información sobre lanzamientos de episodios y medios relacionados con animes, incluyendo episodios especiales, ediciones en DVD y Blu-ray, y contenido adicional.
- **Cluster 1: Relaciones y experiencias personales**
  - **Palabras clave**: love, time, days, every, true, live, spends, together, falls, fall, come, space, spend, work, nature, identity, action, across, station, outer, hard, comes
  - **Tema**: Este cluster se enfoca en relaciones interpersonales y experiencias personales, incluyendo temas de amor, convivencia, trabajo y exploración del espacio.
- **Cluster 2: Música y medios digitales**
  - **Palabras clave**: animated, music, kouji, official, videos, nanke, youtube, posted, site, twitter, website, channel
  - **Tema**: Este cluster está relacionado con la música y las plataformas digitales, como YouTube y redes sociales, usadas para la promoción de animes.
- **Cluster 3: Temporalidad y distancias**
  - **Palabras clave**: years, girls, hundred, three, four, five, ago, passed, since, future, several, thousand, ten, thousands, away, long, academy, cinderella, kingdoms, ever, distant, near, run, taken, runs, far, minutes
  - **Tema**: Este cluster trata sobre la temporalidad y las distancias, abarcando grandes periodos de tiempo y lugares lejanos, a menudo en un contexto de fantasía o futurista.
- **Cluster 4: Personajes y relaciones familiares**
  - **Palabras clave**: young, girl, boy, people, man, many, magical, woman, men, age, mysterious, named, called, little, yearold, strange, whose, beautiful, organization, power, sister, know, secret, older, younger, doesnt, brother
  - **Tema**: Este cluster se enfoca en personajes de diferentes edades y sus relaciones, incluyendo temas de magia, misterio y relaciones familiares.
- **Cluster 5: Videos promocionales y juegos**
  - **Palabras clave**: video, safety, promotional, nhks, program, game, games, features, usic, directed, minna, affic, fire, traffic, uta, featured, mobile, card, suit, gundam
  - **Tema**: Este cluster está centrado en videos promocionales, programas educativos y juegos, incluyendo franquicias específicas como Gundam.
- **Cluster 6: Narrativas épicas y mundos fantásticos**
  - **Palabras clave**: story, world, follows, another, revolves, centers, begins, tells, around, real, save, war, human, domination, fate, digital, fantasy, outside, spirit, entire, thus, earth, humanity, grail, race, ii, civil, beings, federation, planet, holy, alien
  - **Tema**: Este cluster aborda historias épicas y narrativas que giran en torno a mundos fantásticos, guerras y destinos de la humanidad.
- **Cluster 7: Viajes y misiones**
  - **Palabras clave**: home, find, make, back, way, must, order, returns, return, journey, able, fight, matters, get, bring, protect, along, learn, face, embark, sets, evil, worse, trying, spirits

- **Tema**: Este cluster se enfoca en viajes y misiones, donde los personajes deben regresar a casa, proteger algo o enfrentarse a desafíos.

- **Cluster 8: Vida escolar y cotidiana**
  - **Palabras clave**: first, year, school, day, one, life, second, however, daily, high, season, idol, students, elementary, boarding, middle, ordinary, normal, things, finds, present, next, everyday, knows, piece, night, thing, peaceful, peeping, half, soon, lives, schools, junior, schooler, nd, third, fourth, group, learns, becomes, realizes, discovers
  - **Tema**: Este cluster trata sobre la vida escolar y cotidiana de los personajes, abarcando varios niveles educativos y experiencias diarias.
- **Cluster 9: Adaptaciones y formatos de medios**
  - **Palabras clave**: short, Ponderada, based, series, anime, song, film, movie, stories, films, novel, ova, television, added, shorts, full, ducational, encyclopedia, adaptation, franchise, theme, feature, festival, light, synopsis, focuses
  - **Tema**: Este cluster se centra en las adaptaciones de historias en diferentes formatos de medios, como películas, series de televisión, novelas y OVAs.
- **Cluster 10: Amistad y relaciones escolares**
  - **Palabras clave**: new, two, friends, friend, transfer, old, york, threat, brand, boys, student, best, childhood, become, close, council, university, college, president
  - **Tema**: Este cluster está relacionado con la amistad y las relaciones escolares y universitarias, incluyendo transferencias y consejos estudiantiles.
- **Cluster 11: Diseño y desarrollo de personajes**
  - **Palabras clave**: main, characters, character, featuring, designs
  - **Tema**: Este cluster trata sobre los personajes principales y sus diseños en los animes.
- **Cluster 12: Participación y decisiones**
  - **Palabras clave**: take, takes, part, place, decides, care, go, took, taking, join
  - **Tema**: Este cluster se enfoca en la participación de los personajes en eventos y decisiones importantes.
- **Cluster 13: Información pendiente**
  - **Palabras clave**: yet, click, update, information
  - **Tema**: Este cluster parece ser una categoría de información que está pendiente de actualización o acceso adicional.
- **Cluster 14: Contextos históricos y de ciencia ficción**
  - **Palabras clave**: century, st, universal
  - **Tema**: Este cluster parece referirse a un marco temporal específico, posiblemente relacionado con un contexto histórico o de ciencia ficción.
- **Cluster 15: Técnicas de animación y estudios**
  - **Palabras clave**: animation, motion, puppet, stopmotion, toei, stop, nothing
  - **Tema**: Este cluster se centra en las técnicas de animación, incluyendo la animación stop-motion y el trabajo de estudios específicos como Toei.
- **Cluster 16: Libros infantiles y dramas**
  - **Palabras clave**: childrens, book, picture, drama
  - **Tema**: Este cluster está relacionado con libros infantiles, libros ilustrados y dramas.
- **Cluster 17: Bandas y música tradicional**
  - **Palabras clave**: irodorimidori, band, rock, fictional, japanese, traditional
  - **Tema**: Este cluster se enfoca en bandas, música rock y elementos tradicionales japoneses, posiblemente dentro de un contexto ficticio.

B. Resumen de Bondad y Ajuste del Modelo

Valores P para Grado, Distancia y Pareja Compartida

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Grado	0	0.64	1	0.79	1	1
Distancia	0	1	1	0.99	1	1
Pareja Compartida	0	0	0.08	0.29	0.56	1

Tabla 10: Resumen de los Valores P para Grado, Distancia y Pareja Compartida

Bondad del Ajuste para los Clusters

Cluster	Obs	Min	Mean	Max	MC p-value
edges	122481	121821	122179.0	122510	0.10
cluster 0	1705875	1695951	1702160.7	1707047	0.32
cluster 1	4095125	4074502	4088050.8	4098947	0.42
cluster 2	973720	968240	971946.7	974431	0.42
cluster 3	1241743	1234776	1239293.3	1242373	0.34
cluster 4	3286082	3268136	3279477.9	3288555	0.36
cluster 5	2229185	2216610	2224250.8	2229836	0.26
cluster 6	2412738	2400144	2408797.7	2415154	0.48
cluster 7	2386222	2374630	2382467.1	2388385	0.46
cluster 8	4144162	4122698	4136920.0	4148308	0.42
cluster 9	3493075	3476523	3487359.9	3496028	0.42
cluster 10	1501946	1493166	1498744.7	1502775	0.30
cluster 11	1575103	1566711	1571966.4	1576217	0.40
cluster 12	1659399	1650387	1656338.0	1660933	0.42
cluster 13	1451840	1444175	1449136.3	1452858	0.34
cluster 14	563071	559427	561712.6	563190	0.14
cluster 15	592388	589301	591573.8	593010	0.64
cluster 16	1895061	1884417	1891508.9	1896826	0.42
cluster 17	1246402	1239414	1244239.1	1247491	0.42

Tabla 11: Bondad del Ajuste para los Clusters

Referencias

[1] Lee, K., & Lee, K. (2015). Escaping your comfort zone: A graph-based recommender system for finding novel recommendations among relevant items. *Expert Systems with Applications*, 42(10), 4851-4858. Recuperado de https://doi.org/10.1016/j.eswa.2014.07.024.

[2] Chen, H., Gan, M., & Song, M. (2013). A Graph Model for Recommender Systems. In *Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013)* (pp. 878-881). Atlantis Press. Recuperado de https://doi.org/10.2991/iccsee.2013.221

[3] Abbe, E. (2018). Community Detection and Stochastic Block Models: Recent Developments. *Journal of Machine Learning Research*, 18(177), 1-86. Recuperado de http://jmlr.org/papers/v18/16-480.html

[4] Luque, C., Agudelo, I., Leal, K., & Sosa, J. (2024). Caracterización del discurso de posesión presidencial e identificación de comunidades políticas en Colombia: Aproximación empírica desde el análisis de redes sociales. Redes Revista hispana para el análisis de redes sociales, 35(1), 128–150. Recuperado de https://doi.org/10.5565/rev/redes.1021

[5] Bedi, P., Gautam, A., Bansal, S., & Bhatia, D. (2018). Weighted bipartite graph model for recommender system using entropy based similarity measure. En Advances in Intelligent Systems and Computing (pp. 163–173). Springer International Publishing.

[6] Shimomura, L. C., Oyamada, R. S., Vieira, M. R., & Kaster, D. S. (2021). A survey on graph-based methods for similarity searches in metric spaces. Information Systems, 95(101507), 101507. Recuperado de https://doi.org/10.1016/j.is.2020.101507

[7] IBM. (s.f.). ¿Qué es el procesamiento del lenguaje natural (PLN)? *IBM*. Recuperado de `https://www.ibm.com/es-es/topics/natural-language-processing`.

[8] Silge, J., & Robinson, D. (2017). *Text Mining with R*. O'Reilly Media, Inc. Capítulo 4: Relationships Between Words: N-grams and Correlations.

[9] Kolaczyk, E. D., & Csárdi, G. (2014). *Statistical Analysis of Network Data with R* (2nd ed.). Springer.

[10] Newman, M. (2018). *Networks* (2nd ed.). Oxford University Press.

[11] Luke, D. A. (2015). *A User's Guide to Network Analysis in R*. Springer.