SISTEMA RECOMENDADOR DE JUEGOS DE MESA BASADO EN LIGHTFM CON REPRESENTACIONES DE CONTENIDO DERIVADAS DE BERT Y TF-IDF

Sistemas recomendadores - Departamento de Ciencias de la Computación - Escuela de Ingeniería
Pontificia Universidad Católica de Chile
Profesor: Denis Parra
Integrantes: Matías Etcheverry - Martín González - Blanca Romero

Keywords: LightFM, Factorización Matricial, BERT, TF-IDF, Transformers, Sistemas Recomendadores, Basado en Contenido

Abstract

Actualmente, se presencia un auge en la industria de los juegos de mesa, con el lanzamiento de cientos de nuevos juegos. Ante esto, el presente trabajo busca introducir un sistema recomendador de juegos de mesa que ayude a los usuarios a encontrar aquellos que mejor se adaptan a sus preferencias.

Para generar el sistema con mejor rendimiento, se evalúa el desempeño de 7 métodos, los que varían entre filtrado colaborativo (iKNN), basado en contenido (TF-IDF y BERT), e híbrido (LightFM, con variaciones). Se hace uso de la base de datos Board Games Database de BGG en Kaggle y sus diversas tablas para entrenar de manera óptima cada método.

Los resultados obtenidos indican que los métodos híbridos de factorización matricial funcionan de mejor manera que los métodos puros, obteniendo métricas de precisión 10 veces mayores. También, se encuentra que dentro del grupo de métodos híbridos con LightFM, el que usa la matriz de *features* más sencilla de todos es el con mejores resultados. Se hace un análisis para comprender este comportamiento y se concluye que es dado el filtrado agresivo hecho sobre la base de datos para este trabajo.

1. Introducción

En los últimos años, la industria de los juegos de mesa ha visto un fuerte auge con el lanzamiento de cientos de títulos nuevos, y se espera que este crecimiento continúe. Esto genera una dificultad para los usuarios, ya que es cada vez más difícil encontrar aquellos juegos que se ajusten mejor a sus preferencias.

Como solución a esto, se propone la creación de un sistema recomendador de juegos de mesa, el cual recomienda títulos relevantes para un usuario a partir de sus preferencias previas. Para esto, se hace uso de diversos métodos de recomendación utilizados en problemas similares, con el fin de evaluar su capacidad de recomendación y utilidad para la tarea en cuestión.

2. Estado del arte y trabajos relacionados

El problema de la recomendación de juegos de mesa comparte grandes similitudes con otros problemas de recomendación ampliamente conocidos, como el de recomendación de series o películas para usuarios. Para dichas tareas, se suele utilizar tanto métodos de filtrado colaborativo como basados en contenido según los datos con los que se cuente, ya que ambos tienen sus ventajas y desventajas (Isinkaye et al., 2015).

También existe un trabajo que investiga el mismo problema de recomendación de juegos de mesa con la misma base de datos (Suazo, 2024), el que utiliza dichos métodos. En este, se comenta que el filtrado colaborativo sería aquel con mejores resultados, sin embargo, también comenta que se contradice a los resultados encontrados en otros papers. Esto podría deberse al filtro sobre la base de datos que el autor decidió aplicar, el que beneficia al primer método por sobre el segundo.

Ante esto, el equipo del presente documento desarrolló interés por saber cómo se comportan ambos métodos, así como un nuevo método híbrido de factorización matricial, ante un nuevo filtro de la base de datos. Con esto, se busca generar una contribución al área de sistemas de recomendación, a través de los resultados obtenidos y el análisis de estos.

3. Base de datos

Se utiliza la base de datos Board Games Database, de BoardGameGeek, en Kaggle, la que cuenta con más de 22.000 juegos, 411.000 usuarios y 19 millones de ratings. De esta base, se hace uso de las siguientes tablas:

- **user_ratings**: Muestra todos los ratings que cada usuario le ha dado a juegos.

- **games**: Presenta diversas características de cada juego, incluyendo una descripción textual.
- **themes**: Tabla binaria, indica el género de cada juego.

Dado el gran tamaño de esta base de datos, se opta por hacer un sampling de los datos de la siguiente forma:

- Primero se filtran los datos de **user_ratings** para mantener sólo aquellos *ratings* de usuarios que han dejado al menos 50 calificaciones. Luego se filtra para mantener sólo aquellos *ratings* de ítems que han recibido al menos 400 calificaciones. Este filtrado se repite hasta que no se eliminan más entradas de la tabla.
- A partir del filtrado anterior se reduce el dataset a aproximadamente 13 millones de datos, de la cual se obtiene una muestra aleatoria del 2%. De esta, se eliminan los datos nulos y usuarios con menos de 2 calificaciones (para que los conjuntos de entrenamiento y prueba puedan ser estratificados según los usuarios).
- Este muestreo de datos, en comparación al dataset original, resulta en una distribución razonable de temáticas entre los juegos, la que imita la realidad de los datos. (ver Fig 1 y 2 Anexo).

4. Metodología

Se realizan pruebas con diversos métodos, con el fin de encontrar el mejor sistema recomendador para la base de datos con la que se trabaja. Vale mencionar que para todos los métodos se utiliza una separación de 70-30 (%) para el set de entrenamiento (*train*) y prueba (*test*) respectivamente. Esta separación es estratificada por los usuarios, porque en el caso de algunos modelos no se pueden hacer predicciones para un usuario si no se tiene al menos un dato de este mismo en el *set* de entrenamiento.

Se decide evaluar los métodos de iKNN de filtrado colaborativo, TF-IDF (*Term Frequency - Inverse Document Frequency*) y BERT (*Bidirectional Encoder Representations from Transformers*) basados en contenido, y LightFM híbrido. Se hacen 4 modelos distintos de LightFM con distintas características, detalladas en el inciso de resultados. La razón que ayuda a decantar por estos modelos es el tipo de problema con el que se cuenta, las tablas existentes que cuentan con descripciones textuales de los juegos y la proporción de usuarios en comparación a juegos. El método iKNN usa los atributos de los juegos para recomendar otros similares, mientras que los métodos basados en contenido hacen uso de las descripciones textuales de los juegos, presentes en la tabla *games*: TF-IDF cuenta palabras y lo valiosas que son, mientras que BERT analiza simultáneamente las palabras y el contexto en el que se encuentra. Finalmente, la incorporación de LightFM se sustenta en que es un modelo híbrido utilizado con frecuencia, el que incorpora tanto filtrado colaborativo como basado en contenido.

En cuanto a las métricas, se utilizan 4 para la medición del desempeño de los modelos. Las principales son las de MAP (*mean average precision*) y nDCG (*normalized discounted cumulative gain*), las que se enfocan en la precisión de las recomendaciones en comparación a los datos reales. Las 2 métricas restantes son las de novedad y diversidad. La primera indica qué tan novedosa es la lista de juegos recomendados, basándose en la popularidad de los juegos (los que han sido calificados más veces son los más populares, por lo tanto menos novedosos). Por su parte, la diversidad indica qué tan similares son los elementos en la lista de juegos recomendados entre sí, la cual se calcula utilizando la matriz de similitud de ítems, calculada en base a la tabla **user_ratings** usando la distancia de coseno. Mientras mayor sea la similitud entre los ítems recomendados, menor es la diversidad y viceversa.

5. Resultados

5.1 iKNN - Filtrado colaborativo

Se entrena un modelo de ItemKNN, utilizando la distancia de coseno entre los ítems, y el parámetro k = 20 que indica que el algoritmo considera los 20 ítems más cercanos para la predicción.

Para la evaluación de este modelo se hace una muestra aleatoria del 1% de los usuarios del set de prueba (*test*) y una muestra aleatoria del 1% de los juegos vistos en el entrenamiento (*train*) para considerar en las predicciones. Este downsampling se hace con el fin de eficientizar el tiempo de ejecución.

Para la muestra de usuarios del *test set* se predicen los 15 ítems más relevantes (dentro de la muestra de juegos establecida), y se extraen las métricas para el top 5, 10 y 15 de ítems, obteniendo los siguientes resultados:

	MAP@k	**nDCG@k**	**Novedad**	**Diversidad**
k = 5	0.0004	0.0008	10.7354	0.9858
k = 10	0.0004	0.0008		
k = 15	0.0005	0.0013		

Tabla 1: Resultados para iKNN

5.2 TF-IDF

Para este modelo se crea una matriz TF-IDF limitada a 5000 características en base a la columna "*Description"* de la tabla **games**. A partir de esta matriz se calculan recomendaciones en base a

los ítems que el usuario previamente había calificado positivamente (se considera positiva una calificación de 6 o mayor). Se recomiendan los ítems más similares según la matriz, excluyendo los ítems previamente vistos por el usuario. Se toman las top 15 recomendaciones con este método y se obtienen las siguientes métricas:

	MAP@k	**nDCG@k**	**Novedad**	**Diversidad**
k = 5	0.0021	0.0024	10.7297	0.9739
k = 10	0.0022	0.0027		
k = 15	0.0023	0.0032		

Tabla 2: Resultados para TF-IDF

5.3 BERT

Para este modelo se obtiene una matriz de características (factores latentes o *embeddings*) a partir de la columna *"Description"* de **games** usando BERT. Se utiliza el modelo MiniLM-L6 que entrega 384 factores latentes. A partir de esta matriz se calculan las recomendaciones en base a los ítems que el usuario había calificado positivamente, de la misma forma que para TF-IDF.

Se toman las top 15 recomendaciones con este método y se obtienen las siguientes métricas:

	MAP@k	**nDCG@k**	**Novedad**	**Diversidad**
k = 5	0.0011	0.0024	10.7798	0.9718
k = 10	0.0014	0.0027		
k = 15	0.0015	0.0032		

Tabla 3: Resultados para BERT

5.4 LightFM

Dado que LightFM es un método híbrido, se quiso probar distintas versiones de este, cada una basada en distintas representaciones (features) de los ítems. La primera versión utiliza una matriz de vectores binarios que representa los géneros de cada juego, extraídos de la tabla **themes**. La segunda versión emplea la matriz TF-IDF mencionada en el inciso 5.2, con la diferencia de que se usa TruncatedSVD para reducir la dimensionalidad de los vectores latentes a 100 dimensiones. La tercera versión utiliza la matriz de características generada por el modelo BERT descrito en el inciso 5.3, la que, al igual que en la versión anterior, recibe una reducción a 100 dimensiones para los vectores latentes utilizando TruncatedSVD. Finalmente, la cuarta versión

utiliza las características obtenidas en cada una de las versiones anteriores concatenadas en una sola matriz de features, y luego reducida con TruncatedSVD a 150 factores latentes.

Para todas las versiones de LightFM se utilizó BPR (*Bayessian Personalized Ranking*) como función de pérdida, una tasa de aprendizaje de 0.25, 40 factores latentes y 50 épocas de entrenamiento. Al igual que en los modelos anteriores se consideran las top 15 recomendaciones y se obtienen las siguientes métricas:

a. Para el modelo que utiliza los géneros de la tabla **themes**:

	MAP@k	**nDCG@k**	**Novedad**	**Diversidad**
k = 5	0.1847	0.2365	11.1434	0.5340
k = 10	0.1982	0.2715		
k = 15	0.2038	0.2900		

Tabla 4: Resultados para LightFM usando themes

b. Para el modelo que utiliza la reducción de las características TF-IDF obtenidas:

	MAP@k	**nDCG@k**	**Novedad**	**Diversidad**
k = 5	0.0155	0.0197	6.8150	0.6797
k = 10	0.0171	0.0248		
k = 15	0.0174	0.0260		

Tabla 5: Resultados para LightFM usando TF-IDF

c. Para el modelo que utiliza la reducción de los *embeddings* obtenidos con BERT:

	MAP@k	**nDCG@k**	**Novedad**	**Diversidad**
k = 5	0.0155	0.0197	6.8288	0.6761
k = 10	0.0171	0.0248		
k = 15	0.0178	0.0272		

Tabla 6: Resultados para LightFM usando BERT

d. Finalmente, para el modelo que incluye todos los *features* de los tres modelos anteriores, concatenados y reducidos a 150 dimensiones, se obtienen los siguientes resultados:

	MAP@k	nDCG@k	Novedad	Diversidad
k = 5	0.0080	0.0125	6.7853	0.7147
k = 10	0.0101	0.0183		
k = 15	0.0113	0.0233		

Tabla 7: Resultados para LightFM usando todas las features

Para facilitar la comparación entre los modelos, la siguiente tabla presenta las métricas de cada uno considerando solo el caso de k = 15:

Método (k = 15)	MAP@k	nDCG@k	Novedad	Diversidad
iKNN	0.0005	0.0013	10.7354	0.9858
TF-IDF	0.0023	0.0032	10.7297	0.9739
BERT	0.0015	0.0032	10.7798	0.9718
LFM + themes	0.2038	0.2900	11.1434	0.5340
LFM + TF-IDF	0.0174	0.0260	6.8150	0.6797
LFM + BERT	0.0178	0.0272	6.8288	0.6761
LFM + todos	0.0113	0.0233	6.7853	0.7147

Tabla 8: Resultados de cada método, para k = 15

Los resultados aquí presentados entregan valiosa información, y serán analizados en profundidad en el siguiente apartado.

6. Discusión

Los resultados experimentales permiten identificar patrones claros respecto al desempeño de los distintos enfoques de recomendación utilizados.

En primer lugar, LightFM con features de themes (LFM + themes) obtuvo los mejores resultados tanto en precisión (MAP@k = 0.2038) como en ranking (nDCG@k = 0.2900), superando ampliamente a los demás métodos. Este resultado sugiere que los géneros de los juegos aportan la información más relevante para capturar las preferencias de los usuarios, por el estilo y tipo de juego.

Además, todos los métodos basados en LightFM (LFM) superan significativamente en rendimiento a los modelos clásicos iKNN e incluso a los sistemas de contenido puro basados en TF-IDF y BERT, que obtuvieron valores de MAP@k y nDCG@k extremadamente bajos (del orden de 0.002–0.003). Esto indica que los métodos puramente basados en filtrado colaborativo o basados en contenido no se adaptan de buena manera al problema en cuestión. La diferencia que existiría entre estos resultados y los planteados en otros *papers* radicaría en el tamaño de las bases de datos utilizadas, lo que perjudica directamente el desempeño de estos métodos dada la falta de información total. En la misma línea, estos métodos no serían de uso práctico ya que requieren de un modelo de preferencias de los usuarios, además de recomendaciones pasadas de estos, por lo que no tienen buena capacidad para recomendar a potenciales usuarios nuevos (problema de *cold start*).

En contraste, LightFM combina efectivamente señales colaborativas con características de los juegos, lo que le permite generar recomendaciones personalizadas incluso con pocos datos por usuario. Este enfoque híbrido mitiga en parte el problema de *cold start* y aprovecha las correlaciones entre usuarios y juegos en el espacio latente. Ciertamente no entrega buenos resultados (MAP@k en torno a 0.015 y nDCG@k en torno a 0.026), pero la flexibilidad de este método híbrido sería suficiente para triplicar el rendimiento de los modelos puros.

También es muy interesante observar que al incluir múltiples tipos de features (TF-IDF, BERT, themes) simultáneamente (caso "LFM + todos"), el rendimiento disminuye respecto a todos los modelos que LightFM con un único tipo de matriz de features. Esto podría sugerir que combinar demasiadas representaciones puede introducir ruido o redundancia, limitando la capacidad del modelo para identificar patrones relevantes y hacer recomendaciones. Este descubrimiento es contrario a lo esperado por el equipo dado que se pensaba que más features resultan en mejores recomendaciones, sin embargo, se entiende que las matrices de features muestran información distinta en dimensiones latentes, lo que haría que no estén relacionados entre sí directamente, siendo esta la fuente de ruido.

Por último, al analizar las métricas de novedad y diversidad, se observa que los métodos no híbridos (como iKNN, TF-IDF, BERT) tienden a ofrecer recomendaciones más diversas y novedosas, lo cual es esperable al ser un potencial *trade-off* con respecto a la precisión de las recomendaciones. Dadas las distintas cantidades de ratings que hay para cada juego, es probable que se le recomiende un juego bueno pero común a un usuario (modelos con LightFM), mientras que una mala recomendación puede ser un juego poco común, muy enfocado en un nicho y con alta posibilidad de no ser acertado para el usuario, pero que tenga alta novedad (modelos puros).

Es relevante también considerar las limitaciones y potenciales mejoras detectadas por el equipo, necesarias de tener en consideración para trabajos futuros. Estas reflexiones se presentan a continuación.

La mayor limitación que se enfrentó durante la implementación de los modelos de recomendación fue la capacidad de almacenamiento limitado que se tenía a disposición. Todo el proceso de entrenamiento se realizó en cuadernos de Google Colab, donde se tenían disponibles 12.7 GB de RAM compartidos en cuentas gratis de Google Drive con 1 GB de almacenamiento máximo. Esta limitación forzó a utilizar un 1.37% de los datos originalmente disponibles, lo que provocó una disminución en el desempeño de los modelos utilizados.

Las posibles mejoras que se pueden realizar es la utilización de más información para la creación de representaciones de ítems utilizados en el entrenamiento de los modelos de LightFM. Esto puede ser incluyendo tablas disponibles dentro del mismo dataset (como **subcategories**, **mechanics**, **publishers**, etc). También se puede contemplar el uso de información externa a la disponible en el dataset para enriquecer la representación de los ítems. Esto puede utilizar imágenes de los juegos, precios o tamaño de la caja por ejemplo. Cabe decir que no se implementó estas mejoras dado que implican un uso adicional de memoria que no se tenía a disposición.

7. Conclusiones

Este trabajo evaluó el rendimiento de distintos enfoques de recomendación de juegos de mesa, comparando modelos clásicos basados en filtrado colaborativo (iKNN), modelos basados en contenido (TF-IDF y BERT), y modelos híbridos utilizando LightFM con distintas representaciones de los ítems. Los resultados muestran de forma consistente que los enfoques híbridos superan ampliamente a los métodos puros en métricas de precisión (MAP@k y nDCG@k), destacando el modelo LightFM con géneros (*themes*) como el más efectivo. Esto evidencia el valor de incorporar tanto interacciones colaborativas como información estructurada del contenido en la construcción de sistemas recomendadores.

Asimismo, se observó que los métodos con más variedad de features no necesariamente presentan un mejor desempeño. De hecho, el modelo LightFM que combinó todas las matrices (TF-IDF, BERT y *themes*) tuvo peores resultados que aquellos con una sola fuente de features. Esto sugiere que la inclusión de múltiples representaciones puede introducir ruido y reducir la capacidad del modelo para identificar patrones claros, especialmente si las representaciones no están bien alineadas en el espacio latente.

Finalmente, si bien los modelos no híbridos presentan mayor novedad y diversidad en las recomendaciones, esto se da a costa de una pérdida considerable de precisión. En este contexto, LightFM se posiciona como una alternativa robusta y flexible para construir sistemas recomendadores eficientes incluso en escenarios con datos limitados, permitiendo un mejor balance entre personalización y relevancia.

8. Referencias

Agrawal, A. (2019, 18 febrero). Solving business usecases by recommender system using lightFM. *Medium*. https://medium.com/data-science/solving-business-usecases-by-recommender-system-using-lightfm-4ba7b3ac8e62

Board Game Database from BoardGameGeek. (2022, 17 enero). Kaggle. https://www.kaggle.com/datasets/threnjen/board-games-database-from-boardgamegeek/data?select=games.csv

Maciej Kula (2015). Metadata Embeddings for User and Item Cold-start Recommendations. In Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender Systems co-located with 9th ACM Conference on Recommender Systems (RecSys 2015), Vienna, Austria, September 16-20, 2015. (pp. 14–21). CEUR-WS.org.

Qaiser, Shahzad & Ali, Ramsha. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications. 181. 10.5120/ijca2018917395.

Isinkaye, F., Folajimi, Y., and Ojokoh, B. (2015). Recommendation systems: Principles, methods and evaluation. Egyptian Informatics Journal.

Suazo Pavez, D. E. (2024). *Aplicación de sistemas de recomendación a juegos de mesa modernos* [Memoria de título, Universidad de Concepción].

9. Anexos

Para replicar los experimentos mostrados en este paper dirigirse al repositorio https://github.com/blanca-romero/ProyectoRecomendacionJuegos y seguir las instrucciones descritas en el archivo README.md.

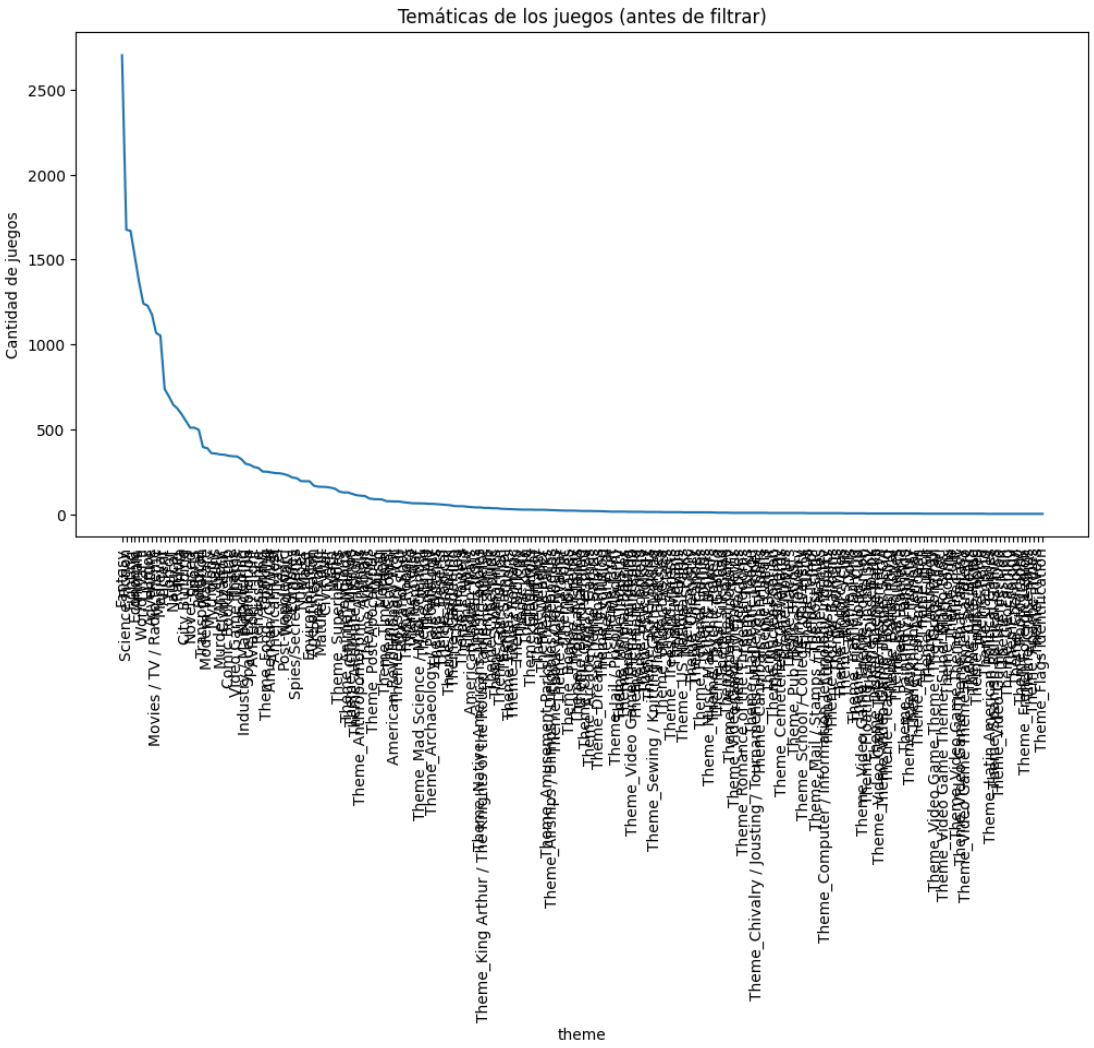

Fig 1: Distribución de frecuencias en las temáticas de los juegos del dataset antes del muestreo

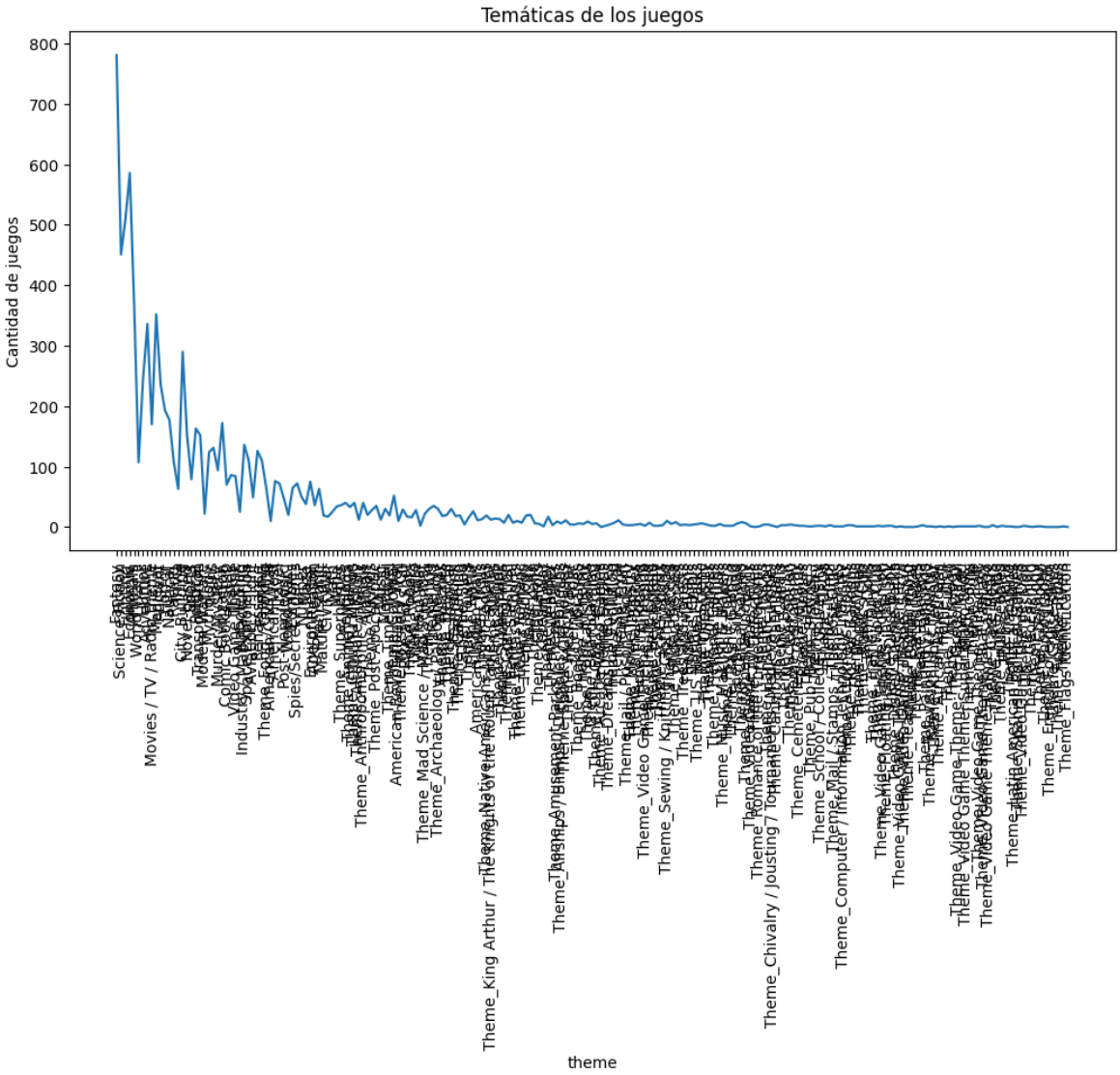

Fig 2: Distribución de frecuencias en las temáticas de los juegos de la muestra realizada