Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Computación
IIC-2433 Minería de datos

"SISTEMA RECOMENDADOR DE ANIMES"

Profesor del Curso: Vicente Domínguez
Ayudante: Ricardo Schilling

Grupo: 05
Integrantes: Gonzalo Barros
Javiera Inostroza
Sebastián León
Samuel Zuñiga

Introducción

El proyecto del equipo consistirá en hacer un sistema recomendador de anime. En primer lugar se deberá obtener una base de datos que entregue la información necesaria para poder desarrollar los algoritmos y obtener valor de los datos. Para ello, se buscará en una comunidad reconocida como lo es myanimelist.net, desde donde se extraerá la información de cada anime. Junto con esto se obtendrán los datos de los diferentes usuarios de la comunidad para utilizarse en la búsqueda de *itemsets* frecuentes. Estos serán utilizados para verificar si es que en la base de datos de animes existen coincidencias entre los diferentes atributos de cada anime perteneciente al itemset y de este modo observar si existe algún parámetro que sea considerado por los usuarios para ver diferentes animes. De esta forma, se obtendrán reglas de asociación entre los diferentes animes y sus atributos, pudiendo obtenerse recomendaciones para cada usuario según lo que le ha calificado positivamente (*rating* mayor a cierto valor) y los atributos que se repiten dentro de los animes que le gustan.

Para esta primera entrega se realizó *webscrapping* de la información de los animes y de la información de cada usuario. Luego se limpiaron los datos para poder utilizarse y se realizó un breve análisis de la información obtenida. Existen datos que se encontraron en la página web que no contienen mayor información más que el nombre, estos datos fueron quitados de la base de datos ya que no son útiles para lo que se busca. Por otro lado también hay algunos que no contienen su información pero si tienen un rating asignado, en este caso, se conservará el dato a pesar de no tener todos los atributos, esto ya que de todas formas puede realizarse una asociación con los datos existentes.

Descripción de los datos

La primera base de datos entregada corresponde a los animes. Contiene una lista de más de 17.000 animes, cada uno con su información específica. La información fue obtenida desde la página web myanimelist.net, desde la cual se obtuvieron los siguientes datos:

- Name: Nombre del anime en cuestión
- Rating: Calificación promedio colocada por los usuarios de esta comunidad
- Type: Tipo de animación, entre los cuales se encuentra TV, OVA, Movie, ONA entre otros.
- Aired: Período en el que se encontró en emisión, contiene tanto la fecha de inicio como la de término.
- Episodios: Cantidad de episodios del anime.
- Genre: Géneros a los cuales pertenece el anime, entre los cuales destacan shonen, seinen, misterio, terror, acción, etc.
- Studios: Estudio o estudios que llevaron a cabo la animación
- Producers: Productores que participan del proceso de creación trayendo actores de voz, obteniendo los horarios de emisión, la banda sonora, y más.
- Licensors: empresas encargadas de las licencias de los animes para ser vendidos o transmitidos en diferentes partes del mundo
- Voice_Actors: los actores de voz de los personajes principales que participan en la versión en japonés
- Link: el link de la página de myanimelist.net que contiene toda la información y más descripción

Esta base de datos inicial servirá para poder observar si es que algunos datos son relevantes a la hora de que los usuarios de esta comunidad prefieran un anime por sobre otro. Esto se evaluará con ayuda de otra base de datos de usuarios, la cual contiene los animes vistos por cada usuario y su respectiva calificación. De este modo, puede obtenerse a los usuarios y los diferentes animes que les han gustado.

Para generar el dataframe de usuarios y sus calificaciones, se utilizaron los archivos csv de usuarios ubicados en la carpeta users_csv del repositorio. Estos archivos contienen el nombre de usuario y su género, el cual puede ser *male*, *female* o *non-binary*. También se utilizaron los archivos csv de *ratings* ubicados en la carpeta ratings_csv del repositorio, los cuales contienen el nombre de usuario, el anime calificado y el *rating* dado por el usuario. Estos datos fueron obtenidos al hacer *scrapping* de la página https://myanimelist.net/users.php. A continuación se presenta la estructura de los datos:

- User: Nombre de usuario
- Gender: Género de usuario. Puede tomar los valores *Male*, *Female* o *Non-Binary*.
- Anime: Nombre del anime que el usuario calificó.
- Rating: Calificación dada por el usuario al anime, la cual puede ir de 0 a 10 puntos.

Cabe destacar que los ratings fueron obtenidos, específicamente de la lista de animes completamente vistos de cada usuario (*completed*). El código de los *scrappings* se encuentra en la carpeta scrapping_anime_list_by_user. El procesamiento de estos datos se realizó en users_preprocessing.ipynb, en donde los archivos .csv fueron importados como

DataFrames. El dataframe de usuarios y sus calificaciones consiste en la unión de todos los archivos generados por los *scrappings* mencionados anteriormente.

Finalmente se unen las dos bases de datos mencionadas anteriormente, formando un dataframe con la siguiente estructura:

- User: Nombre de usuario.
- Gender: Género de usuario. Puede tomar los valores Male, Female o Non-Binary.
- Name: Nombre del anime que el usuario calificó.
- Rating_By_User: Calificación dada por el usuario al anime, la cual puede ir de 0 a 10 puntos.
- Rating: Calificación promedio colocada por los usuarios de esta comunidad al anime.
- Type: Tipo de animación, entre los cuales se encuentra el anime: TV, OVA, Movie, ONA, entre otros.
- Genre: Géneros a los cuales pertenece el anime, entre los cuales destacan shonen, seinen, misterio, terror, acción, etc.
- Studios: Estudio o estudios que llevaron a cabo la animación
- Producers: productores que participan del proceso de creación del anime, trayendo actores de voz, obteniendo los horarios de emisión, la banda sonora, y más.
- Licensors: empresas encargadas de las licencias de los animes para ser vendidos o transmitidos en diferentes partes del mundo.
- Voice_Actors: los actores de voz de los personajes principales que participan en la versión en japonés del anime.

Tras el manejo de datos y la creación de visualizaciones, se observa que la calificación promedio de los animes tiene una distribución normal, concentrándose la mayoría de los animes entre calificaciones 6 y 7 (Figura 1). Además la mayoría de las calificaciones de los usuarios corresponden a 8, seguidas de 0 y 7 debido a que los datos presentan una distribución bimodal (Figura 2).

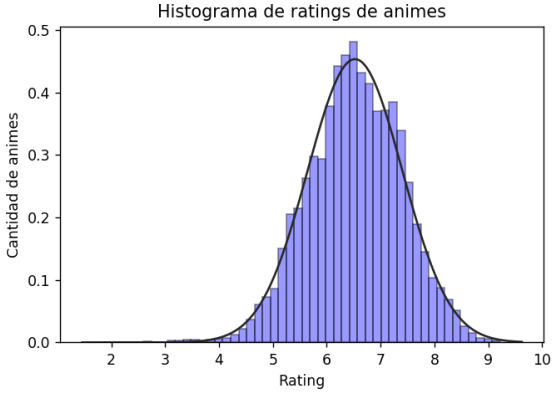

Figura 1: Histograma de ratings de animes
Fuente: Elaboración propia

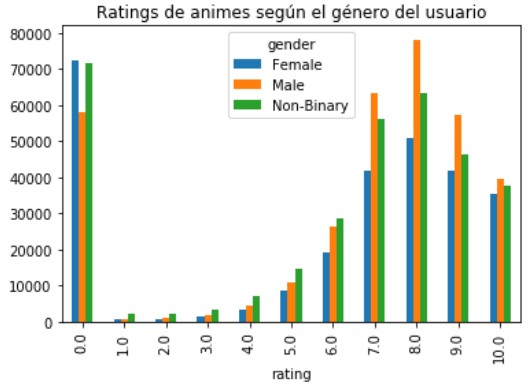

Figura 2: Gráfico de barra de ratings de animes
Fuente: Elaboración propia

Los estudios con mayor producción de animes son *ToeiAnimation*, seguido de *Sunrise*, sin embargo, no son los que tienen mayor cantidad de visualizaciones. *ToeiAnimation* se encuentra en cuarto lugar y *Sunrise* no se encuentra entre los veinte más vistos. A pesar de esto, Sunrise es el que tiene más visualizaciones dentro de los usuarios hombres y se encuentra en cuarto lugar de visualizaciones entre los usuarios no binarios.

Finalmente, la mayor cantidad de animes que se producen son transmitidos mediante el formato TV (Figura 3). También, son los que cuentan con mayor cantidad de visualizaciones superando por una gran cantidad a las de los animes que se transmiten en otros formatos (Figura 4). De las gráficas se pudo inferir que los usuarios no binarios presentan una preferencia por los animes transmitidos en formato OVA, dado que estas visualizaciones corresponden a 120 mil aproximadamente, seguido del formato Special y TV, con aproximadamente 25 mil cada una (Figura 5).

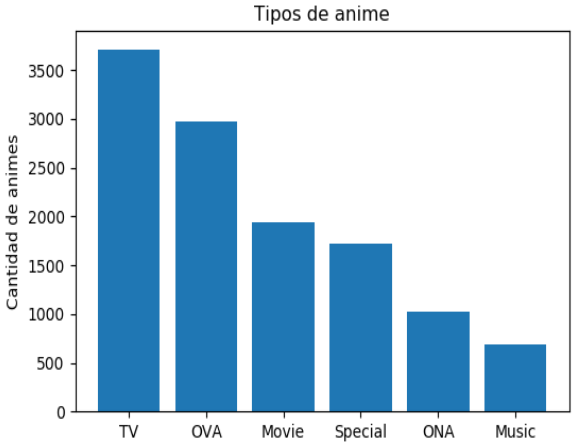

Figura 3: Gráficos de barra de tipos de anime
Fuente: Elaboración propia

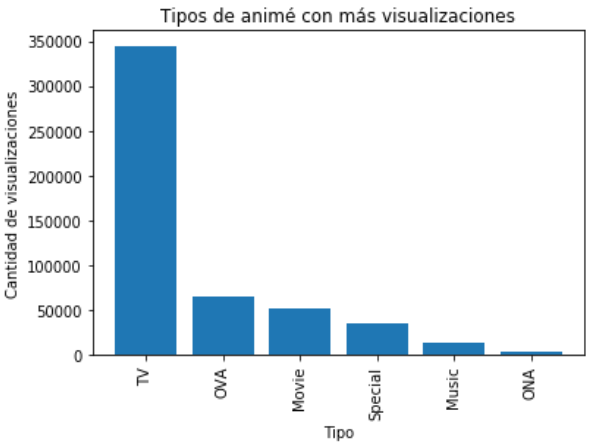

Figura 4: Gráfico de barras de visualizaciones por tipo de anime
Fuente: Elaboración propia

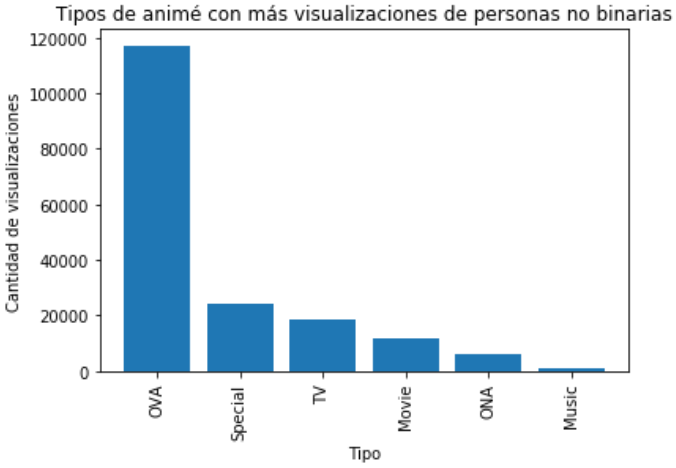

Figura 5: Gráficos de barras de tipos de anime con más visualizaciones de personas no binarias
Fuente: Elaboración propia

Temática o problemática central y describir cómo se abordará inicialmente

La temática central del proyecto consiste en crear un sistema recomendador de animes basado en los ya vistos por los usuarios. Para que un usuario obtenga una recomendación, este debe tener una lista con al menos un anime y cada uno de estos para ser recomendado debe tener al menos una valoración de un usuario. Estas recomendaciones se basarán en las calificaciones que tengan los diferentes animes, los gustos del usuario y las coincidencias entre los atributos de cada animación.

El proyecto se abordará inicialmente ejecutando KNN sobre los datos recolectados para obtener recomendaciones basadas en los animes que ya han sido vistos por un usuario. Además, se usará el algoritmo pyRecLab como segundo método de recomendación al usar los datos de las valoraciones de los usuarios. Posteriormente se evaluarán ambos algoritmos para obtener su confiabilidad y comparar los resultados de ambos. Finalmente se considerará la posibilidad de utilizar un tercer algoritmo para recomendar mediante *clustering.*

Trabajo pendiente para finalizar el proyecto

Revisando los datos obtenidos en esta entrega, se descubrió que al juntar ambos DataFrames disminuyó considerablemente la cantidad de datos. Esto se debió a las irregularidades en los nombres de los animes, por lo que se tomó la decisión de volver a realizar los scrapping tomando en cuenta los identificadores que myanimelist utiliza.

Respecto a los datos de los usuarios, la cantidad de datos de hombres, mujeres y no-binarios fue considerada como equitativa, lo cual no se asemeja a la realidad. Debido a esto, se volverá a realizar el scrapping considerando la proporcionalidad de géneros según los datos de la página (68.95% hombres, 30.95% mujeres, 0.1% no binarios) [Marketing Kit MyAnimeList, 2020].

Por último, queda pendiente la implementación de los algoritmos con las bases de datos obtenidas de los scrapping y su posterior análisis comparativo.

Referencias bibliográficas

MyAnimeList. Julio 2020. *Marketing Kit MyAnimeList.* (Julio de 2020). Disponible en: https://myanimelist.net/advertising.