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

En los últimos años, la industria del anime ha ido creciendo de forma sostenida tanto en Japón como en el resto del mundo. En 2017 por primera vez en la historia, el mercado del anime superó la marca de los 2 billones de yenes (Eiji, 2019). Esto, junto a la creciente popularidad del anime en el extranjero, la difusión de los servicios de streaming como Netflix o Amazon Prime, y el rápido crecimiento de los juegos para smartphones creados por los japoneses, han llevado a que cada día se sumen nuevos interesados en el mundo de la animación japonesa.

Desde el éxito del Viaje de Chihiro en los Oscars del año 2002, la industria del anime se hizo un hueco en occidente y el número de seguidores ha ido en aumento con el pasar del tiempo. Con la popularidad alcanzada de la película Your Name (Andaur, 2017) en años recientes, quedó demostrado que la animación japonesa se consolidó como una industria de entretenimiento que atrae a seguidores de todo el mundo. Además, las personas han tenido cada vez mayor acceso a los animes estrenados en cada temporada y se hace importante para los usuarios el saber qué ver y qué no ver. Debido a lo anterior, surge la importancia de un sistema recomendador para consumir esta cantidad creciente de contenido de forma que las personas vean aquellas obras que realmente van a disfrutar.

Temática o Problemática

La forma en la que se emiten los anime por año es a través de las estaciones. Es decir, cada verano, otoño, invierno y primavera son lanzadas al aire una gran cantidad de series de todo tipo de géneros. Debido a que la cantidad de animes que se transmiten cada temporada en los últimos dos años ha sido en promedio de 40 (Crunchyroll, 2020), los usuarios son incapaces de ver y analizar cada trailer o sinopsis de las series a estrenar, por lo que es difícil predecir qué series deberían ver cada temporada en relación a los gustos de estos. Dado este problema, es importante que exista una forma de poder tomar una decisión correcta para disfrutar animes que sean del gusto de cada persona.

Con lo anterior en mente, nace la interrogante de cómo recomendar algo del gusto del usuario para facilitar su elección. Para ello, se consideró utilizar información previa sobre las calificaciones de distintos animes vistos por el usuario. Con estos datos sería posible obtener relaciones entre lo que fue visto por cada persona y determinar que sería altamente probable que le gustara. Sin embargo, para poder saber esto, primero es necesario acceder a esta información para luego procesarla y entregar recomendaciones.

Descripción de los datasets ocupados

En primera instancia, se encontró una base de datos en la página kaggle, la cual contenía información que llevaba cuatro años sin actualizar (CopperUnion, 2016), por lo que se tomó la decisión de crear una base de datos propia para poder realizar las recomendaciones correspondientes. La información actualizada se extrajo desde la página myanimelist.net, donde se puede obtener los datos de tanto los animes como los distintos usuarios que pertenecen a esa comunidad. Esta página fue seleccionada ya que cuenta con más de 12 millones de visitantes mensuales (MyAnimeList, 2020) para más de 17 mil animes

distribuidos en diferentes categorías y géneros. Para obtener toda esta información, se utilizaron dos algoritmos de web scraping: una para los animes y otra para los usuarios. Una vez retirada esta información, fue almacenada en archivos .csv para ser posteriormente utilizada.

En lo que respecta a los datos rescatados de usuarios, la información contiene el género del usuario, el rating que entregó el usuario al anime, y los identificadores de anime y de usuario. Esta fue retirada a partir de la pestaña de comunidad, en el apartado de usuarios, donde se activó el filtro por género para comenzar a acceder a los diferentes perfiles de la comunidad. Luego, se retiró cada dato existente sobre las puntuaciones que colocaba el usuario para cada anime y esto se fue agregando al archivo de usuarios correspondiente. Gracias a este método, se obtuvieron alrededor de 17 mil datos entre hombres y mujeres y 1.085.758 calificaciones. Con esta información, los métodos a aplicarse más adelante usarán los datos divididos por género, teniendo recomendaciones distintas para cada uno.

En lo que respecta al dataset de animes, este fue extraído de la misma página, obteniéndose alrededor de 17 mil datos de animes. Estos contenían características de rating, productores, cantidad de episodios, fecha de transmisión, link de la página, licenciadores, estudio de animación, géneros, nombre, identificador del anime, tipo de anime y actores de voz de los personajes principales. Después de rescatar toda esta información, existían diversos datos faltantes, por lo tanto se procedió a limpiar los datos. Se eliminaron los datos que no contenían género alguno, datos sin rating y se eliminaron las columnas de cantidad de episodios, fecha de transmisión y link de la página. Con esto se conservaron 12126 animes.

Una vez limpiados todos los datos, se procedió a combinarlos de modo que cada fila de usuario-anime contuviera los datos respectivos de cada anime y eliminar datos según fuera la necesidad de cada algoritmo. Además para evitar tener sesgos en las predicciones, dado que en el universo de datos de la página MyAnimeList el 68.95% (Marketing Kit MyAnimeList, 2020) representa a los hombres, se decidió separar los datos por género, por lo que finalmente se obtienen 2 *dataframes* para ser procesados por los algoritmos, es decir, uno para hombres y otro para mujeres. Esta decisión se tomó al estudiar la distribución de los datos, ya que las visualizaciones de hombres y mujeres mostraban evidentes diferencias. En los algoritmos se utilizó la información del rating que entregó el usuario al anime, el id de este, el id del anime y su rating correspondiente, esto se describe a continuación:

- User: Nombre del usuario
- Anime_id: Identificador del anime calificado
- Anime: Nombre del anime calificado
- Rating: Rating colocado por el usuario al anime

Exploración de los datos

Una vez obtenidos y filtrado los datos, se realizó un análisis estadístico de las características de cada dato. En el caso de los datos de usuarios, se destacan los siguientes gráficos:

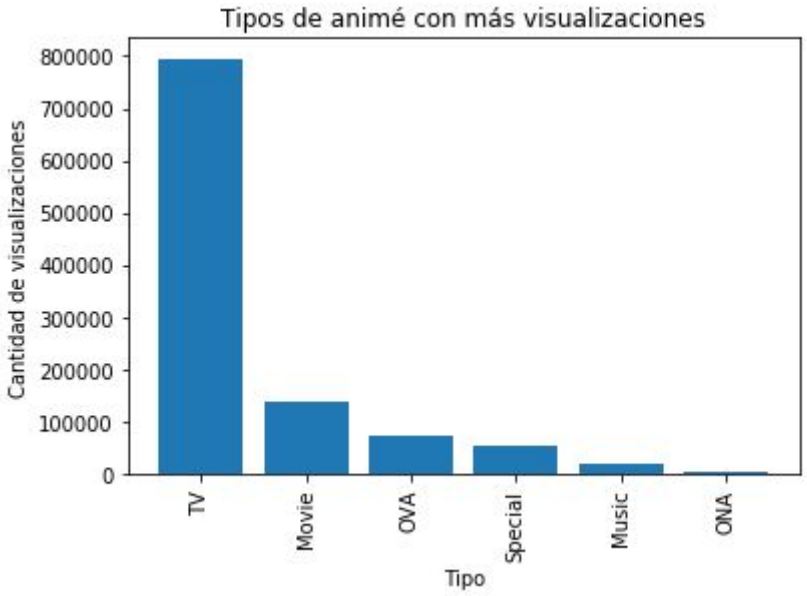

Gráfico 1: Tipos de anime con mas visualizaciones por los usuarios

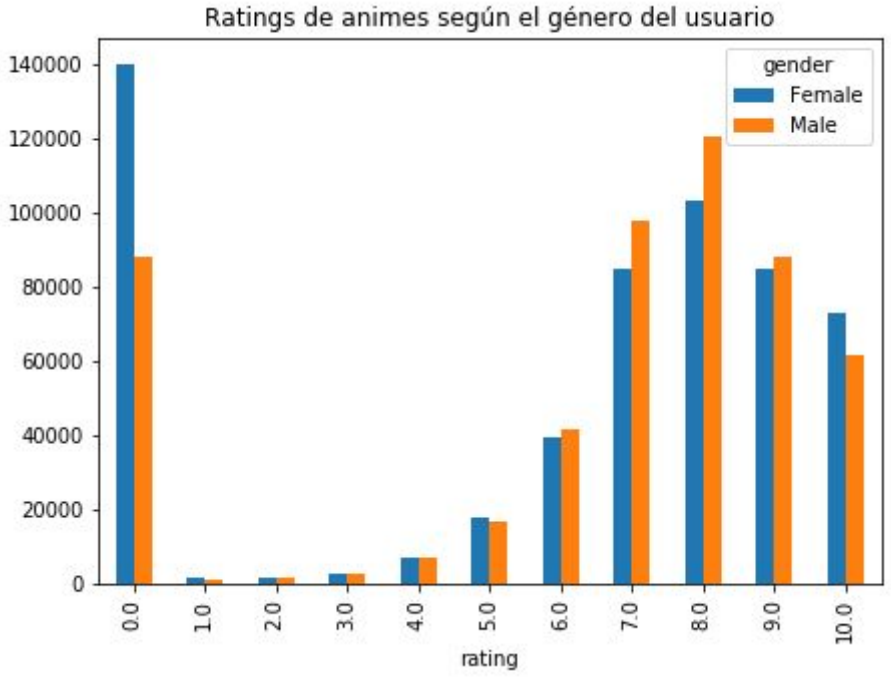

Gráfico 2: Ratings colocados por usuarios, según género

De lo que se puede observar de los gráficos es que, la mayoría de los usuarios consumen principalmente anime del tipo TV. En el gráfico 2 se observa una proporción similar entre hombres y mujeres de los ratings colocados. Es interesante destacar que existe un peak en los ratings 0. Ya que el objetivo principal es la recomendación según los gustos, se eliminarán aquellas puntuaciones iguales o menores a 5, dado que se consideran que al miembro de la comunidad no le gustó.

En lo que respecta a los datos de animes, se observan los siguientes gráficos:

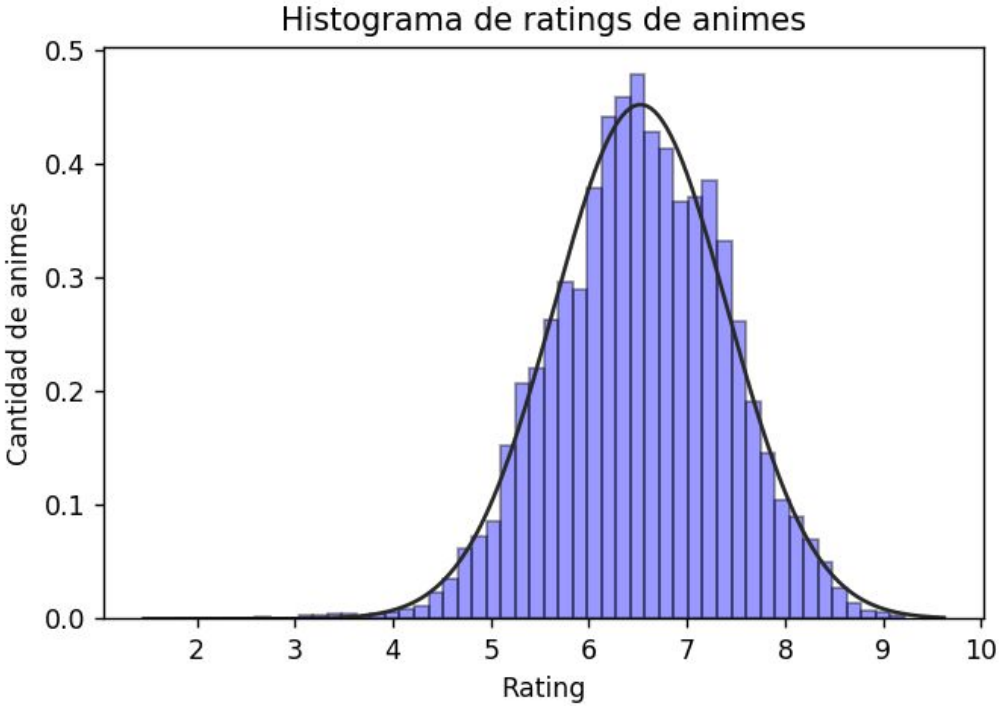

Gráfico 3: Histograma de ratings de animes

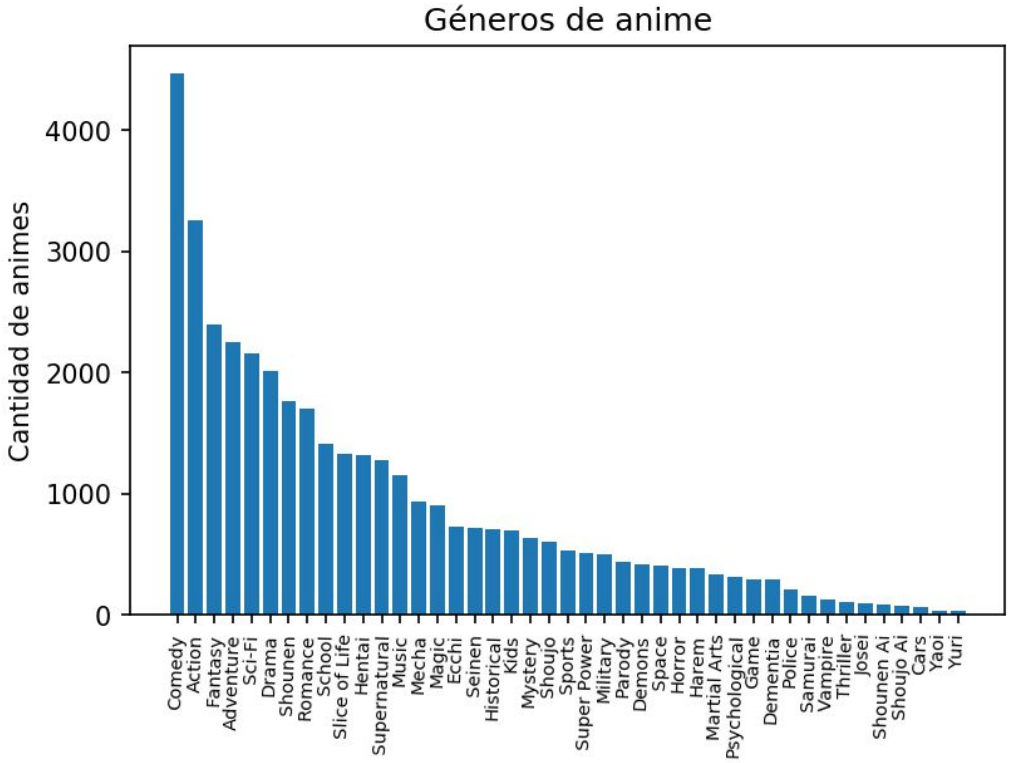

Gráfico 4: Cantidad de animes por género

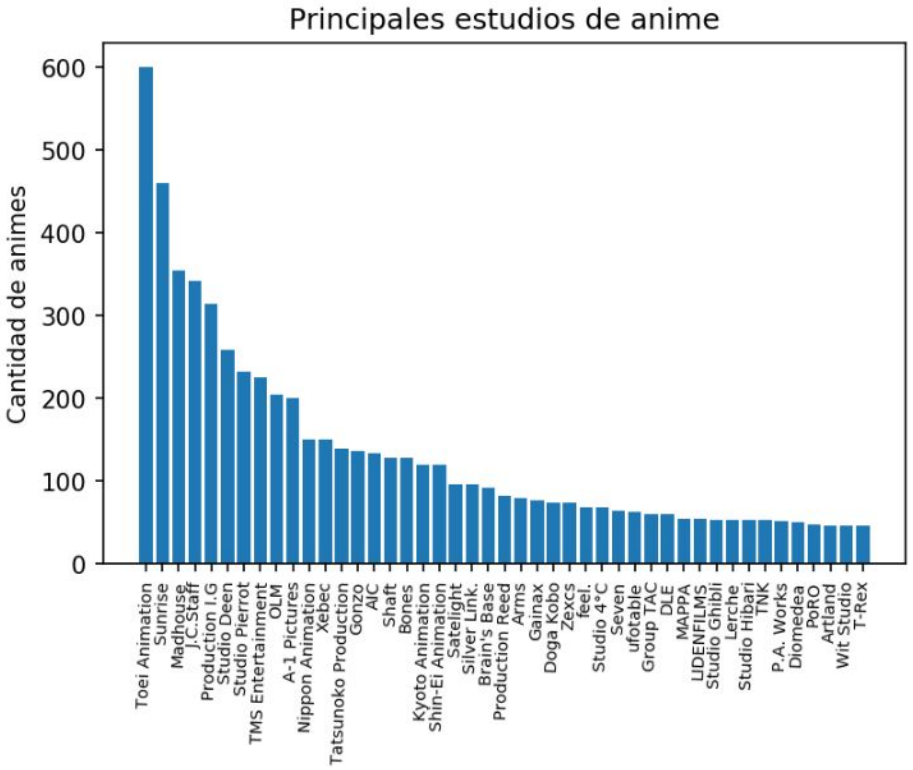

Gráfico 5: Cantidad de animes por estudio

De lo mostrado anteriormente se puede concluir que los ratings para los animes tienen una distribución normal centrada entre las puntuaciones 6 y 7. El principal género que se produce es el de comedia, seguido por acción y un poco más bajo el género de fantasía. Pasados estos géneros, hay una gran cantidad de géneros con cantidades menores a mil, además de notarse una clara diferencia para los animes de comedia y acción en términos de cantidades. En lo que respecta a estudios, el estudio Toei Animation es el que tiene una mayor cantidad de animes bajo su firma, seguido de Sunrise para finalizar el podio con Madhouse. Se puede observar también que existen diversos estudios con cantidades menores 100 animes que se consideran en los datos y que el estudio predominante, Toei, ha realizado casi 600 animes.

Descripción del proceso

Como se mencionó anteriormente, la obtención de los datos fue realizada mediante *web scraping* a la página MyAnimeList. Para realizar esto, se usaron las funciones de la librería Beautiful Soup para la extracción y posteriormente fueron procesados mediante la librería Pandas para condensar la información en *dataframes*, los cuales fueron guardados poco a poco en archivos .csv.

Para los datos extraídos de los usuarios, debido a que la muestra de la página puede sesgar por géneros, ya que existe una mayor proporción de hombres que mujeres en esa comunidad, se optó por realizar sistemas recomendadores para cada género por separado.

Para la limpieza de las calificaciones de los usuarios, primero se eliminan las columnas con datos faltantes, para posteriormente hacer un *merge* con la tabla de male_users y

female_users por el atributo anime_id. De esta forma las calificaciones quedan tanto con la información completa del usuario, como con la información del anime calificado.

El preprocesamiento posterior a esto fue realizado para limpieza de datos de los animes. En una primera etapa, existía una problemática debido a que los datos guardados en .csv contenían listas y estos, al leerse para poder utilizarse, se reincorporan al programa como strings, por lo que se vió en la necesidad de utilizar la librería Abstract Syntax Trees (abs). En esta librería se encuentra la función literal_eval, la cual sirvió para reconvertir los datos a listas nuevamente. Luego de estos percances, se notó que existían diversos datos con valores None, debido a que no se encontraban disponibles en la página. Este problema desencadenó en la necesidad de filtrar aquellos datos que eran necesarios para la clasificación. De esta manera, la limpieza de datos mediante Pandas, retiró las filas que no contenían datos de géneros de anime y las filas donde los animes no tenían un rating asignado. Además de esto, se eliminaron las columnas de 'Aired', 'Episodes' y 'Link' las cuales corresponden a fecha de emisión, cantidad de episodios y link a la página de MyAnimeList.

Posterior al preprocesamiento de los datos de anime, se vió la necesidad de encontrar algoritmos que sirvan para generar recomendaciones a partir de todos los datos obtenidos. Con esto se llegaron a dos posibles alternativas, la primera consiste en la librería PyRecLab (Sepúlveda, Domínguez y Parra, 2017), la cual contiene diferentes métodos de recomendación, de los cuales se seleccionaron: User Average, Item Average y Slope One, y la segunda es la librería LensKit, en donde, de sus diferentes algoritmos de recomendación, se decidió utilizar el algoritmo de Item KNN, User KNN y ALS.

El uso de estos algoritmos se debió a la necesidad de evaluar diferentes tipos y enfoques que entregan los recomendadores frente a los datos obtenidos. Respecto a la librería PyRecLab: el primero de los algoritmos se debió a que entrega un enfoque específico a los usuarios al identificar perfiles similares y recomendar ítems a otros usuarios basado en el rating de personas parecidas. El segundo en cambio, se enfoca en los ítems y recomienda basado en la similitud de las evaluaciones al identificar animes similares en este caso. El tercero, es un algoritmo distinto a los anteriores ya que entra en la categoría de los sistemas de filtrado colaborativo que se encargan de combinar los ratings y usuarios para obtener recomendaciones personalizadas.

Por último, respecto a la librería LensKit se escogió Item KNN, User KNN y ALS, primeramente por su familiaridad con el grupo. El algoritmo Item KNN encuentra clusters basándose en qué tanto se parecen los animes vistos, mientras que el algoritmo User KNN encuentra clusters en base a qué tanto se parecen los usuarios. Finalmente, el algoritmo ALS corresponde a un tipo de algoritmo recomendador diferente a los anteriores, que se basa en la factorización de matrices la cual se encarga de generar valorizaciones de cada ítem para cada usuario.

Para poder evaluar los algoritmos de recomendación se deben usar métricas para comparar los resultados de cada uno de estos, por lo cual, se establece para las listas de predicción los siguientes indicadores:

MAE: (error absoluto medio): Medida que entrega el promedio del valor absoluto de la diferencia entre las predicciones y los datos observados. La fórmula viene dada por:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}.$$

RMSE (raíz del error cuadrático medio): Medida que entrega la raíz cuadrada de la diferencia promedio entre las predicciones y los datos observados.

$$\mathrm{RMSD}(\hat{\theta}) = \sqrt{\mathrm{MSE}(\hat{\theta})} = \sqrt{\mathrm{E}((\hat{\theta} - \theta)^2)}.$$

Y para las listas de recomendación los siguientes indicadores:

MAP (*Mean Average Precision*): Es la media del promedio de las precisiones, en específico, es la media respecto a los usuarios de cada precisión promedio de las listas de recomendación.

$$\mathrm{MAP} = \frac{\sum_{q=1}^{Q} \mathrm{AveP(q)}}{Q}$$

nDGC (*Normalized Discounted Cumulative Gain*): Medida que utiliza una escala de relevancia graduada de ítems del conjunto de resultados para evaluar la utilidad, o ganancia, de un ítem en función de su posición en la lista de resultados.

$$\mathrm{nDCG_p} = \frac{DCG_p}{IDCGp}.$$

Para los indicadores de las listas de predicción entre más bajos sean sus valores mejor será el algoritmo para predecir los ratings de los ítems, mientras que para las métricas de las listas de recomendación entre más altos sean sus valores mejor será su algoritmo para recomendar ítems a los usuarios

Evaluación de resultados

Mediante la librería LensKit se implementaron los algoritmos de Item KNN, User KNN y ALS. La evaluación de estos algoritmos se hizo para mujeres y para hombres por separado, utilizando las métricas error absoluto medio (MAE), raíz del error cuadrado medio (RMSE), precisión media promedio (MAP) y la ganancia acumulada descontada normalizada (nDCG), obteniendose los siguientes resultados:

Método LensKit	KNN - Item		KNN - User		ALS	
Género	Hombre	Mujer	Hombre	Mujer	Hombre	Mujer
MAE	0,66132	0,64024	0,58440	0,62916	0,18691	0,19196
RMSE	0,84719	0,81463	0,78540	0,80467	0,24853	0,25838
MAP	0,00187	0,00249	0,00059	0,00044	0,00122	0,00118
nDCG	0,01238	0,01011	0,00236	0,00111	0,02898	0,02919

Tabla 1: Resultados de LensKit

En el caso de PyRecLab, esta contiene diferentes métodos de predicciones y recomendaciones, además de encontrarse altamente documentada en su página de github, y al igual que LensKit, esta librería trae implementadas sus métricas.

A partir de PyRecLab se utilizaron los algoritmos User average, Item average y Slope One. Cada uno de estos métodos fue evaluado tanto para hombres como para mujeres mediante la utilización de diversos métodos de medición como lo son el error absoluto medio (MAE), la raíz del error cuadrado medio (RMSE), la precisión media promedio (MAP) y la ganancia acumulada descontada normalizada (nDCG).

Método PyRecLab	User Average		Item Average		SlopeOne	
Género	Hombre	Mujer	Hombre	Mujer	Hombre	Mujer
MAE	0,92552	1,05655	0,97754	1,05655	2,96517	3,03601
RMSE	1,15702	1,29881	1,22161	1,29881	3,24532	3,33926
MAP	0,00023	0,00006	0,00025	0,00006	0,00003	0,00003
nDCG	0,00008	0,00003	0,00005	0,00003	0,00002	0,00001

Tabla 2: Resultados de PyRecLab

Finalmente, se procede juntar los resultados para poder realizar la evaluación conjunta de los algoritmos, obteniendo la siguiente tabla:

Librería	PyRecLab						LensKit					
Método	User Average		Item Average		SlopeOne		KNN - Item		KNN - User		ALS	
Género	Hombre	Mujer	Hombre	Mujer	Hombre	Mujer	Hombre	Mujer	Hombre	Mujer	Hombre	Mujer
MAE	0,92552	1,05655	0,97754	1,05655	2,96517	3,03601	0,66132	0,64024	0,58440	0,62916	0,18691	0,19196
RMSE	1,15702	1,29881	1,22161	1,29881	3,24532	3,33926	0,84719	0,81463	0,78540	0,80467	0,24853	0,25838
MAP	0,00023	0,00006	0,00025	0,00006	0,00003	0,00003	0,00187	0,00249	0,00059	0,00044	0,00122	0,00118
nDCG	0,00008	0,00003	0,00005	0,00003	0,00002	0,00001	0,01238	0,01011	0,00236	0,00111	0,02898	0,02919

Tabla 3: Evaluación de los Algoritmos

De lo observado en la Tabla 3, se puede apreciar que los mejores algoritmos, tanto para hombres como mujeres fueron Item KNN y ALS. Esto es en base a los resultados de las

diferentes métricas que se presentan ya que para las listas de predicción se obtienen los valores más pequeños entre los ratings de las predicciones y los observados, mientras que para las listas de recomendación se obtienen los valores más altos de las métricas de precisión y de ganancia.

Conclusiones

En base a la evaluación de resultados, se puede observar que el mejor algoritmo para la predicción fue ALS, seguido de User KNN. El resto de los algoritmos se descarta dado que sus métricas de evaluación se encuentran alejadas de estos dos primeros.

Respecto a los resultados obtenidos en la recomendación, el mejor algoritmo es KNN-Item para la métrica MAP. Para nDGC el mejor rendimiento se obtiene con ALS. El bajo rendimiento de los otros algoritmos puede deberse a la cantidad reducida de datos con las que se trabajó, debido a las limitaciones en los recursos de las máquinas virtuales, o a un posible sobre entrenamiento, lo cual quedará a investigar en trabajos futuros con una mayor cantidad de datos.

Dado el éxito que tiene hoy en día el mundo del anime y que la industria sigue creciendo, un algoritmo recomendador con KNN es, por ahora, una buena solución al problema de que los usuarios no saben qué ver dada la enorme oferta de cada temporada. Sin embargo, como la industria del anime no deja de crecer, se hace necesaria la búsqueda de algoritmos más eficientes, que permitan filtrar según las características de usuarios o los distintos atributos de los animes. Producto de esto surge la necesidad de realizar un trabajo futuro, que permita clasificar con otras features además del rating por usuario.

Trabajo futuro

Como trabajo futuro, para generar mejores recomendaciones se requiere de una mayor cantidad de datos de usuarios, es por esto, que si se quiere seguir adelante, es necesario realizar un *web scraping* más exhaustivo de estos. Una cantidad mayor de datos podría permitir recomendaciones más acertadas pero, requerirá de mayores cantidades de recursos computacionales para la resolución. Otra opción de proceder en un futuro para evitar este problema mencionado es buscar métodos que se ajusten mejor a la situación considerada para este trabajo, tanto para la cantidad de datos como para las métricas utilizadas.

Aparte de esto, la metadata disponible de cada anime no fue utilizada. Existe la posibilidad de utilizar estos datos para generar otro tipo de relaciones entre los animes y así, tener más posibilidades de predecir conexiones entre las animaciones. Entregando así recomendaciones donde se consideren estudios de animación o actores de voz, entre otras características, además de los ratings de cada anime.

Finalmente, evaluar otras métricas de precisión para definir qué algoritmo conviene al momento de realizar recomendaciones con KNN-Items y ALS, puesto que como se mencionó anteriormente, KNN-Items es mejor en la métrica MAP, mientras que ALS es mejor en nDGC.

Principales dificultades

Existieron diversas dificultades a través de cada etapa del proyecto. En el proceso de *web scraping*, la página no tenía una definición del lenguaje html para cada elemento que se quería obtener, por lo tanto, fue necesario hardcodear diferentes elementos para poder obtenerse. Otra complicación fue el manejo de listas de los *dataframes*, ya que las listas guardadas en los archivos .csv una vez que se querían utilizar, Pandas interpreta esos datos como *strings*, por lo que fue necesario buscar alguna librería para solucionar ese problema o programarlo desde cero. Respecto a la obtención de datos de usuarios, la página de MyAnimeList no contaba con una lista de todos los usuarios, de modo que se debió hacer el scrapping por género de usuarios, ya que en la página, al filtrar por este atributo, si era posible acceder a una lista.

Una vez solucionados los problemas con los datos, existieron dificultades con la utilización de cada librería y los algoritmos de recomendación que existían, ya que debía entenderse lo básico del funcionamiento y posteriormente calcular métricas para los desempeños, las cuales debían verificarse y asegurarse que los resultados fueran correctos para poder concluir algo. La utilización de cada librería no es trivial y se debe ser cuidadoso con los datos y parámetros que necesita cada algoritmo, debe leerse la documentación con atención para evitar cualquier error.

Otra dificultad que existió se debió a la utilización de la librería PyRecLab, debido a que esta no tiene soporte para sistemas operativos windows, por lo que se recurrió a otro entorno de programación como es Colab de Google, para poder utilizarla.

Referencias bibliográficas

Andaur, E. (2017). El filme de anime que cautivó al mundo: Your name. Disponible en: https://www.diarioconcepcion.cl/cultura-y-espectaculos/2017/10/05/el-filme-de-anime-que-cautivo-al-mundo-your-name.html

CooperUnion. (2016). Anime Recommendations Database. Disponible en:https://www.kaggle.com/CooperUnion/anime-recommendations-database

Crunchyroll. (2020). Invierno 2019 - Otoño 2020. https://www.crunchyroll.com/es/videos/anime/seasons#/es/videos/anime/seasons/fall-2020

Eiji, H.(2019). Anime Market Tops ¥2-trillion Mark for the First Time. Disponible en: https://japan-forward.com/anime-market-tops-¥2-trillion-mark-for-the-first-time/

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

Sepulveda, G., Dominguez, V., Parra, D. (2017). pyRecLab: A Software Library for Quick Prototyping of Recommender Systems. Disponible en: https://github.com/gasevi/pyreclab