Clustering de usuarios de MyAnimeList: Hallando patrones y tendencias por comunidad

Integrantes:

Benjamín Ojeda

Nicolás Benavides

Víctor Marques

Matías Valenzuela

GitHub del proyecto:

Proyecto Grupo 12

Índice

Contexto y motivación:

En cualquier afición o actividad de interés con cierto nivel de popularidad, tienden a generarse comunidades muy distintas que comparten el mismo pasatiempo en común. A fin de cuentas, los humanos somos diferentes a lo largo del mundo, y existen hobbies de tan nivel de amplitud y variedad que pueden coexistir múltiples razones para que sean atrayentes, pero... ¿cuáles son dichas razones? ¿Cuán heterogéneo resulta ser el universo interesado en cierto tema arbitrario? En este proyecto, principalmente se tratará de identificar las diversas comunidades presentes en aquellos fans de la animación japonesa.

La industria del anime moderno nació alrededor de la década de 1960, y desde entonces cada vez ha gozado de una mayor popularidad. Según ciertas estadísticas (Ferjan, 2023), alrededor de más de un tercio de la población mundial ve anime, lo que equivale a 2,88 mil millones de personas. Además, la base de datos de artes multimedia japonesas cuenta con más de 136.700 obras bajo la categoría de anime en sus registros. En los Estados Unidos, se estima que más del 50% de la generación Z ve anime (Ferjan, 2023), y se proyecta que conforme avance el tiempo el reconocimiento de esta industria ciertamente será mayor en nuestras vidas. En un ámbito más local, según un análisis de Google Trends, se vio que Chile es el quinto país en el cual el anime es más popular (Ferjan, 2023).

Por otro lado, se procederá a dar un poco de contexto respecto a lo que es MyAnimeList. Es un sitio web fundado en 2004 por el estadounidense Garrett Gyssler, que tiene como propósito primario el ser una herramienta para que sus usuarios puedan registrar el anime y manga que han visto, inscribiéndose en el concepto de **comunidad virtual de catalogación**; término que Urban Dictionary define de la siguiente manera:

> "*Un subconjunto de los medios sociales, un sitio web o aplicación web que permite a los usuarios catalogar digitalmente las cosas que les preocupan -música, libros, películas, arte, gastronomía, artesanía, bricolaje, etc-, mientras que al mismo tiempo pueden compartir su contenido en una comunidad en línea*". (Urban Dictionary, 2011)

Hoy en día, MyAnimeList es una compañía de bolsa, operada por el grupo Media Do desde 2019. La creación de una cuenta de usuario con cerca de todas las funcionalidades asociadas es gratis, mas también se ofrece un servicio mensual de 2,99 dólares para ser un "seguidor de MAL"; que ofrece los beneficios de eliminación de anuncios de la página junto a la posibilidad de una mayor personalización del perfil propio. Como grupo, creemos que estos beneficios que se obtienen por pagar mensualmente son escasos y no valen mucho la pena, por lo que, frente a esta problemática, planteamos una factible posible solución. Ella consiste en que se ofrezcan como funcionalidades pagadas el que sitio web le permita al usuario saber al respecto sobre en cuál comunidad probablemente se identifique más; qué serie, en base a lo anterior, le parezca mejor y sea recomendado que vea, entre un sinfín de otras estadísticas personalizadas que se podrían obtener trabajando correctamente el enorme banco de datos que sustenta MyAnimeList. Nótese que, muy recientemente, el sitio web ha implementado una página de estadísticas personalizada para cada usuario, que justamente recoge la idea de que si se desean saber todos los datos o gráficos sea necesario pagar. El inconveniente con ello es que aún esas

estadísticas que se ofrecen son muy simples a nuestro parecer; no siendo lo suficientemente interesantes o con un trabajo serio por detrás; y es debido a ello que nos interesó la posibilidad de ampliar masivamente la cantidad de información valiosa que podría estar ahí. Básicamente, el único tipo de estadísticas que brinda MyAnimeList son del carácter descriptivo elemental, como cantidad de animes visualizados (entre muchas otras cosas), promedios, distribuciones de calificaciones o desviaciones estándar; medidas que cualquier persona podría fácilmente calcular o graficar con el suficiente tiempo.

En síntesis, observamos una oportunidad actualmente malgastada en MyAnimeList, que podría significar una importante ganancia económica para Media Do si es implementada correctamente, y para ello es indispensable utilizar las herramientas y el conocimiento de Ciencia de Datos. A partir de lo anterior, resulta evidente señalar a los trabajadores de Media Do como el público objetivo de este proyecto, pues sin lugar a duda serían los mayores beneficiarios al ser ellos quienes lograrían mejorar la rentabilidad de su compañía tomando como base las ideas desarrolladas por este trabajo. Como comentario aparte, nótese que ello es coherente con las aspiraciones financieras de Media Do, dado que un antecedente existente en 2019 (Ressler, 2019) da cuenta que dicha compañía tiene intenciones de fortalecer MyAnimeList; mejorando las funcionalidades del sitio como una plataforma de *marketing* y distribuidora de libros electrónicos. De manera indirecta, débase decirse que este proyecto tiene como público a todas aquellas personas que les gustaría saber y estudiar las diversas comunidades y sus características, que existen, como mínimo, entre los actuales usuarios de MyAnimeList (cerca de 15.106.951 cuentas bajo un registro del 9 de octubre de 2023 (MyAnimeList, 2023)).

No siendo una cuestión menos relevante, débase decirse que la inspiración inicial para escoger este proyecto proviene de enterarse del proyecto hecho por el profesor Hernán Valdivieso, que vigentemente realiza clases de Programación Avanzada y Visualización de Información en la Pontificia Universidad Católica de Chile. Su proyecto consistió en, primeramente, armar un enorme *dataset* de MyAnimeList, que recopiló en el siguiente repositorio de GitHub, titulado: "MyAnimeList-Database". Este repositorio cuenta con la información de la lista de animes de 325.772 usuarios diferentes, y también con la información de 17.562 animes distintos. Nótese que dichos datos fueron extraídos en parte (las preferencias de los usuarios) con la Jikan API, entre el 26 de febrero y el 20 de marzo de 2020. Sin embargo, para nuestro proyecto usaremos la MyAnimeList API, dado que es la oficial del sitio web; siendo la Jikan una API no oficial. Ello tiene coherencia dado que la API oficial no estaba disponible para su utilización en el momento donde el profesor hizo su *dataset*, puesto que en 2020 recién se habilitó desde el 6 de julio, según lo mencionado en el siguiente post. A modo de abreviar, considérese que escribir MAL es equivalente a escribir MyAnimeList.

Otro comentario importante que vale la pena destacar es la cuestión de la procedencia de las etiquetas. Como contexto, cada entrada de anime en MAL cuenta con dos elementos cuya intención es resumir lo fundamental o estilo de cada serie: géneros y temas. El problema con estos apartados es que resultaría realmente difícil trabajar con ellos puesto que se requeriría realizar primero la labor previa de comprobar la verdadera importancia de cada uno; además de que complejizaría innecesariamente *a posteriori* un adecuado *clustering*. Para paliar dicha situación, el docente nos informó de que otro sitio web, similar en funcionalidad a MAL, llamado Anime-Planet, cuenta con un sistema mejor de

etiquetación de sus animes. En este, existe un único apartado llamado -tags-, que podría decirse que combina los géneros y temas que coloca MAL; y además suelen existir más (lo cual genera que se pueda reconocer con una mayor facilidad las temáticas y "esencia" de la serie; característica sumamente positiva para nuestra idea de proyecto). Se verá este aspecto inmediatamente a continuación con un ejemplo (*Welcome to the NHK!*) del cual se tiene el conocimiento para aseverar si es notoria la mejora de este apartado en Anime-Planet en comparación a MyAnimeList.

Géneros de *Welcome to the NHK!* en MAL: *Comedy, Drama, Romance*.

Temas de *Welcome to the NHK!* en MAL: *Adult Cast, Otaku Culture, Psychological*

Tags de *Welcome to the NHK!* en Anime-Planet: *Comedy, Drama, Slice of Life, Dark Comedy, Loneliness, Mental Illness, NEET, Otaku Culture, Parody, Psychological, Satire, Based on a Light Novel*.

Se verifica exitosamente que son los tags de Anime-Planet los cuales captan de la forma más completa y fiel lo tratado en la obra audiovisual.

A partir de dicho trabajo realizado, nos dimos cuenta de que podría ser muy interesante el cumplimiento de ciertos objetivos; que en la posterior sección se explayarán.

Como documentación de los múltiples tags con los que cuenta Anime-Planet, en la siguiente página se describe cada uno.

Objetivos:

1. Catalogación de grupos de usuarios de MyAnimeList, fundamentalmente bajo los siguientes parámetros: Animes vistos, tags de los animes vistos, y las calificaciones que brindan los usuarios.
2. Indagar en la capacidad predictiva de los grupos de usuarios hallados, ya sea para toda métrica relevante del análisis. Con ello se busca responder la interrogante de la efectividad de un algoritmo de recomendación de anime que considere primariamente los animes observados y sus tags, y la respectiva calificación otorgada. En un ejemplo hipotético, si se observa que una persona pertenece a una comunidad "A" y no ha visto una serie que está muy bien evaluada dentro de su grupo, sería razonable concebir que es una buena idea que el sistema recomendara dicha serie a ese usuario.
3. Estudiar las características de los diversos grupos encontrados y determinar en base a ellos tendencias generalizadas que puedan extrapolarse a los fans de la animación nipona. De acuerdo con dichas tendencias, el sitio web podría determinar la viabilidad de la implementación de ciertas funcionalidades nuevas o la adaptación de las ya existentes para mejor.
4. Determinar si existe una correlación entre la cantidad de series vistas y la comunidad a la que pertenece un usuario en MyAnimeList. Naturalmente se pensaría que sí, pero resultaría interesante saber cuán fuerte es esta posible correlación.
5. Identificar los animes que generan la mayor polarización de calificaciones en diferentes comunidades. Un beneficio directo de este objetivo está orientado a una mejor administración del sitio web. ¿De qué manera se organizarían mejor los recursos? Bueno, si se sabe de antemano las series de animación que provocan la mayor disonancia en la gente, la administración de MyAnimeList podría destinar una mayor cantidad de moderadores de foros hacia aquellas series en particular. De esta forma, el debate sobre estas series podría mantenerse en un tono civilizado y razonable; propiciando así un mejor ambiente de charla en MyAnimeList que incentive a sus usuarios a interactuar en el sitio web, perfeccionando la imagen de la compañía.
6. Identificación de animes de culto, "clásicos"; series de animación que han adquirido un estatus renombrado (ya sea por sus altas calificaciones y fama o ambas características) y analizar qué características concretas los distinguen (en caso de existir) y de qué manera (si la hay) influyen en las comunidades de fans.

Audiencia objetivo: Sección de innovación o similar (que esté encargada esencialmente de implementar nuevas funcionalidades en MyAnimeList) de la empresa Media Do. La argumentación de ello ya fue desarrollada a cabalidad en la sección anterior de -Contexto y motivación-.

Datos:

1) Usuarios y su respectiva calificación a los animes (MyAnimeList API)

La extracción y conversión de estos datos a un *dataframe* ha quedado documentada en el GitHub del proyecto.

Consiste en un *dataframe* que contiene una cantidad arbitraria de columnas (depende de la cantidad de animes con los cuales se llegue a trabajar). Refleja a cada perfil de usuario en la primera columna, y también representa a cada anime en la primera fila. "anime_id" en este caso no representa a los usuarios, si no que representa a los identificadores de los animes dentro del sitio; por ejemplo, la id 30 corresponde a Neon Genesis Evangelion.

Los usuarios tienen el siguiente formato: "id_user:username", estos se encuentran en la primera fila, que también sirve de índice. Considere la siguiente imagen:

anime_id	6	20	30	32	47	227	269	550	777	1535	...	32801	32936	32961	33023
6618661:Strife9667	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	...	NaN	NaN	NaN	NaN
12960354:talesvl	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	10.0	...	NaN	NaN	NaN	NaN
15791607:kiritominz	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
9193223:Stari_Cross	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
16279944:RetroDoug	NaN	7.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
9328058:FrenzyX	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
12720513:Kremsev	0.0	NaN	NaN	NaN	0.0	0.0	0.0	NaN	0.0	0.0	...	NaN	NaN	NaN	NaN
6073632:SexyShadow	NaN	NaN	NaN	NaN	NaN	NaN	2.0	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
4785445:sunnnysideup	NaN	6.0	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0.0	...	NaN	NaN	NaN	NaN
12261756:R112	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	...	NaN	NaN	NaN	NaN
1112469:DanR7985	9.0	9.0	8.0	NaN	8.0	8.0	6.0	NaN	NaN	10.0	...	NaN	NaN	NaN	NaN
509046:Ty-kun	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
6151623:ShedimNuckelavee	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
17187137:Sharapatak21	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	10.0	...	NaN	NaN	NaN	NaN
17269271:zushizushi	NaN	NaN	8.0	8.0	NaN	10.0	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
12830514:Bryataro	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
7876237:Sam_kasper	NaN	8.0	NaN	8.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
14957485:Punoh47	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
15748558:Eminuel_	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
5499718:keelennis	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	0.0	0.0	0.0	0.0

Figura 1. Un vistazo al *dataframe* de 1)

Otra cosa que hay que tomar en cuenta es que este *dataframe* está traspuesto; si uno ve su forma normal, entonces se parecerá más a uno tradicional.

En un ejemplo sencillo, la imagen indica que el usuario "Strife9667" ha calificado con un 10.0 al anime que MyAnimeList le asigna el id de 6. Lo cual es real considerando esto y esto.

Las calificaciones que se pueden otorgar en el sitio van del 0.0 al 10.0.

anime_id	13249226:Daniel_Davi	14461699:Wizzmoo	10241534:Dinotronik	9125142:anime_battlebun	13331280:janiceyyey
30	10.0	NaN	NaN	7.0	7.0
31	10.0	NaN	NaN	NaN	NaN
32	10.0	NaN	NaN	NaN	NaN
2759	0.0	NaN	NaN	NaN	NaN
3784	0.0	NaN	NaN	NaN	NaN
...	...	...	...	...	...
33337	NaN	NaN	NaN	NaN	NaN
33513	NaN	NaN	NaN	NaN	NaN
37976	NaN	NaN	NaN	NaN	NaN
40620	NaN	NaN	NaN	NaN	NaN
42923	NaN	NaN	NaN	NaN	NaN

Figura 2. *Dataframe* de figura 1 en estado no traspuesto.

Debido a que los usuarios no califican todos los animes del sitio, (hay usuarios que solo ven o califican muy pocos animes) eso implica que la base de datos tendrá muchos NaN (tal como se ve en la imagen), dado que dicho usuario no tiene registrado dicho anime en su lista. En otro sentido, la calificación de 0.0 explica animes que la persona ha visto, ve, o planea ver, pero no ha calificado. En este caso, resulta indispensable decidir si eliminar estos datos o usarlos para otro fin más tarde.

2) Animes y sus tags (*Web scraping* de Anime-Planet)

La extracción y conversión de los datos de la página del anime particular a un *dataframe* quedará documentada en el GitHub del proyecto. La imagen que colocamos es un ejemplo de cómo debería quedar el producto final.

Consiste en un *dataframe* con dos columnas, "Anime" y "Tags". "Anime" tiene valores en formato *object*, que representan el nombre del anime en cuestión. "Tags" tiene valores en formato *object*, que representan los variados tags que coloca Anime-Planet.

	Anime	Tags
0	Neon Genesis Evangelion	[Drama, Mecha, Sci Fi, Conspiracy, Kaijuu, Loneliness, Melancholy, Military, Monsters, Post-apocalyptic, Psychological, Super Robot, Original Work]
1	Fullmetal Alchemist: Brotherhood	[Action, Adventure Drama, Fantasy, Mystery, Shounen, Conspiracy, Death of a Loved One, Military, Siblings, Based on a Manga]
2	Legend of the Galactic Heroes	[Drama, Sci Fi, Military, Outer Space, Political, Rivalries, Space Opera, War, Based on a Novel]
3	Steins;Gate	[Sci Fi, Conspiracy, Psychological, Thriller, Time Loop, Time Travel, Based on a Visual Novel]

Figura 3. Un vistazo al *dataframe* de 2)

En un ejemplo, se tiene que los tags asociados al anime "Legend of the Galactic Heroes" son Drama, Sci Fi, Military, Outer Space, Political, Rivalries, Space Opera, War, y, Based on a Novel.

3) *Metadata* útil asociada a cada anime en MyAnimeList (MyAnimeList API)

La extracción y conversión de estos datos a un *dataframe* ha quedado documentada en el GitHub del proyecto.

Consiste en un *dataframe* que reúne la información que, a nuestro juicio, consideramos potencialmente valiosa de analizar para nuestro proyecto. Su cantidad de filas depende de la cantidad final de animes con los cuales trabajaremos. En cuanto a las columnas, dado que son tantas, se consideró mejor explicar cada una en detalle, lo que se hará a continuación.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1 entries, 0 to 0
Data columns (total 19 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   id                           1 non-null      int64
 1   title                        1 non-null      object
 2   start_date                   1 non-null      object
 3   end_date                     1 non-null      object
 4   synopsis                     1 non-null      object
 5   mean                         1 non-null      float64
 6   rank                         1 non-null      int64
 7   popularity                   1 non-null      int64
 8   num_list_users               1 non-null      int64
 9   num_scoring_users            1 non-null      int64
 10  media_type                   1 non-null      object
 11  status                       1 non-null      object
 12  num_episodes                 1 non-null      int64
 13  studios                      1 non-null      object
 14  alternative_titles.synonyms  1 non-null      object
 15  alternative_titles.en        1 non-null      object
 16  alternative_titles.ja        1 non-null      object
 17  start_season.year            1 non-null      int64
 18  start_season.season          1 non-null      object
dtypes: float64(1), int64(7), object(11)
memory usage: 280.0+ bytes
```

Figura 4. Detalle de las 19 columnas del *dataframe* de 3)

id: La identificación del anime dentro del sitio MyAnimeList, es un *int*.

title: El nombre principal del anime, es un texto.

start_date: Cuando empezó a emitirse originalmente, podría convertirse a *datetime*.

end_date: Fin de la emisión del anime, podría convertirse a *datetime*.

synopsis: Un breve resumen del anime, es un texto.

mean: La puntuación promedio del anime, es un *float*.

rank: Su lugar en el ranking de los animes con mejores puntuaciones en el sitio, es un *int*.

popularity: Su lugar en el ranking de los animes con más miembros; donde miembro es alguien que tenga el anime dentro de su lista. La variable es un *int*.

num_list_users: La cantidad de miembros del anime; donde miembro es alguien que tenga el anime dentro de su lista. Es un *int*.

num_scoring_users: La cantidad de usuarios que le han dado una puntuación al anime.

media_type: Como se estrenó originalmente este anime, generalmente se divide en anime de televisión, "tv", o una serie que se tiene que visualizar comprando discos de la serie, un "OVA". Similarmente, existe la categoría "ONA", que se utiliza para referirse a animes que primero salieron para internet. En este caso la variable podría ser una categoría.

status: Representa si sigue en emisión, o si ya terminó de transmitirse. Podría ser una categoría.

num_episodes: La cantidad de episodios que tiene el anime.

studios: Los estudios, u organizaciones que trabajaron para el desarrollo del anime; sería una lista.

alternative_titles.synonyms: Una lista de otros nombres que tiene el anime.

alternative_titles.en: El nombre del anime en inglés, sería un *string*.

alternative_titles.ja: El nombre del anime en japones, sería un *string*.

start_season.year: El año en que comenzó sus emisiones. Sería un *int*.

start_season.season: La estación en que empieza su emisión: invierno, primavera, verano u otoño; podría ser una categoría.

Preguntas de investigación:

1. ¿Cuáles grupos de usuarios se pueden formar a partir de las *reviews* que ellos hacen respecto a los animes vistos, los diferentes animes visualizados y sus tags asociados? ¿Siquiera existen esos grupos? En caso afirmativo, ¿qué nombres específicos representarían la esencia de los grupos hallados?
2. O incluso si existen, ¿son suficientemente significativas las diferencias como para hacer buenas recomendaciones o predicciones de otras variables?
3. ¿Cómo se relacionan otras variables, como el tipo de media, cantidad de episodios, compañía de producción, popularidad, tags, año de lanzamiento, etc, con la calificación dada por los usuarios?
4. ¿Se pueden hallar patrones o tendencias regulares en las variables de los animes "clásicos" (entendiéndose que estos cuentan con una excepcional recepción positiva o fama o ambos)? ¿Cuáles son dichos patrones? ¿Tienen sentido con lo que normalmente se esperaría?
5. ¿Está correlacionada la cantidad de animes vistos con la "comunidad" que se pertenece? Dicha interrogante también es factible de plantear reemplazando la cantidad de animes vistos con otras variables del proyecto.
6. ¿Qué animes generan la mayor polarización (disonancia en cuanto a calificación: desviación estándar de las notas) en las comunidades halladas (dentro y fuera de una comunidad)?
7. Por otro lado, ¿qué animes generan la menor polarización (considerando los mismos supuestos de la pregunta anterior)?
8. ¿Existen patrones regulares en los nombres u otras variables recopiladas de los usuarios de un mismo grupo? Si es así, ¿cuáles son esos patrones y cuentan con una explicación coherente por detrás?

Diseño tentativo:

Para este proyecto, la herramienta estelar y basal que se usará es el *clustering*. Pero… ¿en qué consiste el *clustering*? Buscando en internet se encontró una muy buena y clara definición escrita por Joaquín Amat Rodrigo, un sénior *Data Scientist* español que cuenta con una web de divulgación de su área bastante completa e informativa.

> *"El término clustering hace referencia a un amplio abanico de técnicas cuya finalidad es encontrar patrones o grupos (clusters) dentro de un conjunto de observaciones. Las particiones se establecen de forma que, las observaciones que están dentro de un mismo grupo son similares entre ellas y distintas a las observaciones de otros grupos. Se trata de un método de aprendizaje no supervisado (unsupervised), ya que el proceso no tiene en cuenta a qué grupo pertenece realmente cada observación (si es que existe tal información). Esta característica es la que diferencia al clustering de los métodos de clasificación en el que sí emplea la verdadera clasificación durante su entrenamiento."* (Amat Rodrigo, 2020)

Como nuestra intención es precisamente encontrar conjuntos de usuarios similares de MyAnimeList, es sumamente razonable que utilicemos dicha herramienta, la cual se encuentra implementada en la librería scikit-learn de Python. En su página web, se describen los distintos algoritmos de *clustering* que se pueden aplicar, mas se reconoce que en esta etapa del proyecto se desconoce cuál finalmente se llegará a ocupar. Creemos que todavía nos falta el conocimiento necesario para responder dicha interrogante y, además, solamente luego de realizar un análisis exhaustivo a los datos del trabajo se podrá tomar una verdadera decisión más informada y consciente respecto a ello. Menciónese que dicha librería también podrá sernos útil en el caso de que veamos conveniente realizar regresiones.

Primeramente, nuestro plan de proyecto parte con el paso de reunir los datos necesarios que necesitamos para el cumplimiento de nuestros objetivos. Veamos que la fuente original de nuestros datos se puede encontrar en sitios web, y, más específicamente, la data de MAL es factible de conseguir fundamentalmente utilizando su API oficial, mientras que la data de Anime-Planet se tendrá que conseguir bajo *web scraping*. Para estas labores, se utilizarán las librerías de pandas (para crear los *dataframes*), requests y BeautifulSoup (obtención de información). Sin embargo, plantéese que un inconveniente el cual tendremos que enfrentar es el hecho de que no es muy viable conseguir en un reducido tiempo la data de la lista de usuarios de los más de 15 millones de usuarios de MyAnimeList. Por lo tanto, se necesitará antes determinar una muestra aleatoria de los usuarios, y para ello una posibilidad es usar la librería matemática de Numpy. Ya con todos los datos recopilados, puede pasarse a la siguiente etapa.

Segundamente, lo elemental que debemos realizar para nuestro proyecto es una limpieza y exploración de los datos que hayamos recolectado. Esto, es especialmente relevante teniendo en cuenta el contexto de que los datos que trabajaremos serán obtenidos en base a la utilización de una API y *web scraping*, y puede que en ambos procesos existan fallas inherentes en los valores extraídos o una cantidad importante de información nula. Es nuestra responsabilidad revisar atenta y

minuciosamente cualquier tipo de desperfecto en dichos datos para corregirlos; así como igualmente descartar posible información que lleguemos a considerar innecesaria y podría sobrecargar la carga computacional asociada a los algoritmos que se usarán; disminuyendo y entorpeciendo nuestro avance en el proyecto.

Terceramente, se piensa realizar lo ya mencionado, el susodicho *clustering*. De acuerdo con sus resultados, se analizará cuál (si existe) es un nombre adecuado para cada grupo generado; viendo si reúne las características habituales que siguen las comunidades de fans de la animación (gusto o interés por una estética (temáticas) en concreto y/o juicio más o menos severo con respecto a lo pensado de la calidad de las obras visualizadas).

En cuarto lugar, se realizará una etapa en la cual se piensan tomar muchas medidas para tratar de responder a nuestros objetivos y preguntas planteadas. Para ello, en parte se usará la librería de SciPy, para obtener medidas estadísticas como desviaciones estándar, coeficientes de variación y coeficientes de correlación, entre otros estadísticos. Además, si se ve necesario, puede realizarse un análisis de componentes principales para simplificar los datos que tenemos. También podrían aplicarse técnicas de análisis factorial para identificar posibles correlaciones entre las variables de estudio. Para el apartado de identificar patrones de texto, pueden usarse las funciones implementadas de la librería spaCy, que ayudan a procesar lenguaje natural.

En quinto y último lugar, se tiene planeado armar gráficos y figuras que den cuenta bellamente de conclusiones de nuestro proyecto; comunicando en parte lo logrado. Para ello, se usarán las librerías de matplotlib, seaborn, bokeh o plotly.

Para resumir, se colocará una lista que da cuenta de las librerías que podrían utilizarse en este proyecto:

- pandas
- Numpy
- SciPy
- spaCy
- requests
- BeautifulSoup
- scikit-learn
- matplotlib
- seaborn
- bokeh
- plotly

Bibliografía

Amat Rodrigo, J. (Diciembre de 2020). Recuperado el 10 de Octubre de 2023, de CienciadeDatos: https://cienciadedatos.net/documentos/py20-clustering-con-python

Ferjan, M. (15 de Febrero de 2023). *HeadphonesAddict*. Obtenido de https://headphonesaddict.com/anime-statistics/#:~:text=Some%20estimates%20say%20more%20than,2.88%20billion%20people%20watch%20anime.

MyAnimeList. (9 de Octubre de 2023). Recuperado el 9 de Octubre de 2023, de MyAnimeList: https://myanimelist.net/forum/

Ressler, K. (15 de Enero de 2019). Obtenido de AnimeNewsNetwork: https://www.animenewsnetwork.com/news/2019-01-15/media-do-acquires-myanimelist-website/.142092

Urban Dictionary. (4 de Noviembre de 2011). *Urban Dictionary*. Obtenido de https://www.urbandictionary.com/define.php?term=social%20cataloging