

Aplicação de Análise de Dados na Adaptação de Obras Literárias para Animações Japonesas

Mesquita S. Arthur¹

¹8º Período de Engenharia de Computação, CEFET-MG Campus V

Divinópolis, MG, Brasil

arthur.santana.mesquita@gmail.com

Abstract. Este trabalho tem como objetivo aplicar conceitos de raspagem de dados e implementação de estruturas de dados de forma a definir mangás promissores para adaptações para animes. A atividade foi realizada ao longo do período 2025.2 para as disciplinas de Ciência de Dados e Algoritmos e Estrutura de Dados, no curso de Engenharia de Computação do CEFET-MG - Campus V tendo sua entrega final no dia 10 de Dezembro.

Keywords: Algoritmos, Anime, Dados

1 Introdução

Desde 2007, com a criação da plataforma online de acesso digital a filmes da Netflix, mais e mais serviços de streaming têm sido criados, e sua importância têm crescido cada vez mais em relação a TV tradicional. Como consequência, muitos dos shows previamente disponíveis na Netflix, hoje foram divididos entre diferentes plataformas, incentivando a Netflix à criação de novos estímulos e diferenciais para sua plataforma.

Um incentivo que fica evidente é o investimento no mercado de obras orientais, que ainda não possuem a atenção das grandes empresas de entretenimento, e portanto, possuem o potencial para crescimento comercial através da aquisição de direitos de distribuição e investimento nos estúdios em seus países de origem.

Dentro desse contexto, o conhecimento de quais combinações de gêneros possuem mais potencial para criação de obras é essencial e pode proporcionar uma vantagem na tomada de decisões financeiras como investimentos externos ou internos e na criação de novas obras de ficção.

Através da análise de dados de consumidores de obras de entretenimento e os gêneros atribuídos a tais obras, agrupou-se os usuários em públicos-alvo e, por conseguinte, encontrou-se arranjos de elementos com potencial de atrair membros destes grupos a assinar uma plataforma de streaming. Tal processo foi explorado neste trabalho com foco em animações japonesas (animes) e baseado na plataforma de crítica de animes *MyAnimeList*.

2 Descrição do Problema

O projeto desenvolvido teve o intuito de analisar os gêneros atribuídos a diferentes animes e encontrar padrões de uso entre usuários que costumam classificar positivamente tais gêneros, de forma a encontrar combinações de gênero com potencial para criação de novas obras de entretenimento no futuro. Para tal, foi feita a raspagem de perfis da plataforma [MyAnimeList](#) diretamente, usando BeautifulSoup para obtenção dos mangás e shows, seus gêneros, assim como a mídia original daquela obra (mangá, light novel, manhwa ou diretamente para anime). Depois, o conjunto de obras e notas de cada usuário foi utilizado para calcular sua similaridade com outros usuários, aglomerando aqueles que possuem interesses similares em um grafo gerado. Por fim, o grafo foi usado para calcular possíveis associações de gêneros e permitirão a criação de shows com potencial de atingir o maior número de usuários de um aglomerado.

3 Implementação

3.1 Estrutura Geral

Foram construídos **7 scripts em Python**:

- **extract_anime.py**

Contém as funções que interagem com o `animes_cache.csv`, criando ou lendo o arquivo e armazenando novas entradas no vetor de animes com base no ID do anime fornecido.

- **extract_manga.py**

Script separado que possui a implementação das funções de raspagem de mangás, para criação do `manga_cache.csv`. Uma base de dados de mangás ainda não adaptados, que foi usada na etapa final do projeto.

- **extract_users.py**

Script separado que possui a implementação das funções de raspagem de IDs de usuário, para criação do `usernames.csv`. Uma base de dados de usuários ativos desde uma data especificada, que são os usuários que são analisados por `profiler.py`.

- **normalizer.py**

Contém as funções de normalização, alguns parâmetros que definem como os dados serão tratados na criação do `profiles.csv` além da função de criação do mesmo.

- **profiler.py**

Usando das funções em `extract_anime.py`, do cachê de animes gerado, e dos IDs de usuário verifica as características dos animes de cada perfil para a criação da assinatura única daquele usuário.

- **graph_generator.py**

Contém as funções que geram o grafo e suas imagens, separando as assinaturas de usuário com base em sua similaridade e com as arestas do grafo sendo a similaridade de cosseno entre dois usuários.

- **recommender.py**

Contém as funções que, utilizando da base de dados de mangás e do grafo gerado, selecionam o mangá com maior similaridade à uma comunidade específica.

Sendo que na execução final, os scripts possuem a seguinte ordem de execução:

`extract_manga.py → extract_users.py → profiler.py →→ recommender.py`

3.2 Dependências

Para executar as diversas partes do projeto em *python* são necessárias algumas dependências:

- **Dependências para extração de dados:** *BeautifulSoup 4*, *LXML* e *Requests* através do seguinte comando do terminal:

```
pip install bs4 lxml requests
```

- **Dependências para criação de assinaturas e grafos:** *TQDM*, *NetworkX*, *Louvain* e *numpy* através do seguinte comando do terminal:

```
pip install tqdm networkX louvain
```

3.3 Extração

As funções de extração `extract_manga.py`, `extract_users.py`, `extract_anime.py`, são responsáveis por gerar a base de dados que foi analisada ao longo do projeto. Como todos os dados foram extraídos de uma plataforma online (MyAnimeList), existem algumas considerações que devem ser pontuadas sobre o método de extração.

- Os mangás analisados possuem nota mínima de aproximadamente 7, totalizando 9400 mangás, dos quais 4060 não possuem adaptação. Considerando que o algoritmo foi capaz de gerar as 5 recomendações consistentes para todas as comunidades sempre acima da nota 7.3, espera-se que não haja necessidade de considerar mangás abaixo deste limite no processo de recomendação.

- Na extração de usuários uma limitação do MyAnimeList, enquanto plataforma para extração, se apresenta: Para a criação de páginas de busca de usuários necessita-se de inserir algum dos seguintes parâmetros de filtro: *gênero, localização ou nome*.

Como nome é um parâmetro muito limitador, restariam apenas os outros dois, que são ambos opcionais durante a criação de uma conta no site. Assim, elimina-se da extração quaisquer usuários que não inseriram tal dado de filtragem no perfil. Inicialmente optou-se pelo gênero, porém após o início do processo de extração, percebeu-se a grande quantidade de contas vazias - possivelmente spam - com o gênero "Feminino".

Para evitar a criação de um viés de gênero na base de dados, optou-se então por utilizar o parâmetro de localização como "Brazil", Fazendo dessa pesquisa uma pesquisa de natureza regional no Brasil. Poderia-se incluir outros países como "United States", porém também verificou-se uma quantidade grande de contas inativas e vazias nesta região, totalizando algo próximo de 93% das contas encontradas - 22 de 24 usuários por página.

Por fim, para evitar contas inativas há muito tempo - possivelmente pessoas que não consomem mais animes - antes mesmo da verificação de conta vazia, verifica-se se o último acesso do usuário foi depois de 2017.

- Os animes foram analisados conforme apareciam com nota superior a 7 na lista de usuários extraídos, totalizando 5433 animes. Através desse método, espera-se que um pequeno número de obras com uma nota média maior que 7 não sejam adicionados à base de dados. Considera-se porém que por não se encontrarem na base de dados dos usuários (com nota alta o suficiente), estes não podem ser considerados **populares** dentro do conjunto de consumidores analisado, e são irrelevantes para a pesquisa. Ao aumentar-se o número de contas analisadas, espera-se que o cachê aumente também.

Após a execução de cada um dos scripts são gerados os arquivos csv contendo os dados extraídos de cada categoria. Como o script `extract_anime.py` normalmente é executado automaticamente dentro de `profiler.py`, para a criação do cache de animes com base nos animes disponíveis em cada lista, mas depende da base de usuários válida gerada por `extract_users.py`, este script gera o arquivo `usernames.csv`, que possui somente uma coluna, contendo todos os nomes de usuários extraídos.

Já os outros arquivos gerados, têm sua estrutura representada pelas tabelas de exemplo a seguir:

ID	Nome	Score	Gêneros	Tipo
656	Vagabond	9.27	"Action, Adventure, Award Winning"	Manga
147272	The Greatest Estate Developer	9.02	"Adventure, Comedy, Fantasy"	Manhwa
4632	Oyasumi Punpun (Goodnight Punpun)	8.99	"Drama, Slice of Life"	Manga
657	Real	8.96	"Award Winning, Drama, Sports"	Manga
3	20th Century Boys	9.89	"Award Winning, Drama, Mystery, Sci-Fi"	Manga

Table 1. Exemplo da representação de mangás não adaptados no arquivo gerado `manga-cache.csv`. As colunas principais utilizadas para a categorização são *Gênero* e *Tipo*

ID	Nome	Gêneros	Source
50265	Spy x Family	"Action, Award Winning, Comedy"	Manga
14289	"Suki tte Ii na yo. (Say ""I Love You."")"	"Drama, Romance"	Manga
47194	Summertime Render (Summer Time Rendering)	"Mystery, Supernatural, Suspense"	Manga
11757	Sword Art Online	"Action, Adventure, Fantasy, Romance"	Light novel
41353	The God of High School	"Action, Fantasy"	Manga

Table 2. Exemplo da representação de animes no arquivo gerado `anime-cache.csv`. As colunas principais utilizadas para a categorização são *Gênero* e *Source*

Repare que com relação ao arquivo de mangás, a coluna de *Score* foi removida, pois a nota relevante para a categorização é a do usuário, e não a nota média do anime, e a coluna *Tipo* foi substituída pela coluna *Source*,

pois todos os elementos possuem o mesmo tipo (Anime), mas podem ser **adaptações** de diferentes tipos de mídia, sendo sua fonte o elemento relevante.

3.4 Criação dos Perfis de Consumo

Para a criação dos perfis de consumo foi necessário acessar a lista de cada uma das contas extraídas e disponíveis em `usernames.csv`, extraíndo as características dos animes em questão. A partir disso foi criado a versão do `anime_cache.csv` que foi usada para os testes. Antes de qualquer coisa, os dados dos animes foram tratados para facilitar sua categorização: Alguns gêneros como "Award Winning" não foram considerados nos perfis, além de que várias origens de mídia não são literárias, e portanto também não devem ser consideradas diretamente. Estas foram todas incluídas na categoria "Other". Já as mídias Manhua e Manhwa também foram consideradas como o mesmo tipo - Para contexto, "Manhua" designa histórias em quadrinhos chinesas, enquanto "Manhwa" designa histórias em quadrinhos coreanas - o motivo disso é que ambas estão, em sua grande maioria, disponíveis na internet, de forma que seus públicos podem ser considerados como públicos de "Web comic". Já as Web Novels, foram categorizadas junto com Light Novels por se tratarem de mídias puramente escritas.

A estas fontes de obras foram atribuídos pesos que são utilizados para contrabalancear a grande predominância de alguns tipos de obra, mais especificamente, diminuir o peso de *mangás*, a mídia mais comum dentre adaptações, para que Manhwas e Light Novels possam aparecer mais frequentemente dentro das comunidades geradas ao final da execução.

Por fim, para cada elemento analisado, se não estiver presente em `anime_cache.csv`, extrai-se os seus dados, caso já esteja, inclui sua nota nos valores relativos de gênero e source de cada uma das colunas. Ao final da análise da lista do usuário, os valores são normalizados por proporção, se tornando valores de 0 a 1 indicando a porcentagem de afinidade de cada usuário com o elemento daquela coluna. Tais valores são então salvos no arquivo `profiles.csv`, conforme a seguinte estrutura:

Username	Manga	L-Novel	Anime	Manhwa	Other	Action	Adv	Comedy	Drama	Fantasy	SciFi	SoL	Rom	Thrill	Supnat	Sport
Usuário 1	0.4551	0.6049	0.1794	0.0587	0.0131	0.6525	0.3148	0.1762	0.3524	0.4290	0.1321	0.0000	0.2561	0.1974	0.2104	0.0000
Usuário 2	0.5993	0.2726	0.2311	0.0521	0.0104	0.3013	0.2166	0.2573	0.4528	0.2427	0.1482	0.0163	0.1954	0.1824	0.2638	0.0130
Usuário 3	0.6231	0.3059	0.1284	0.0162	0.0657	0.2474	0.2543	0.4913	0.2740	0.2468	0.1353	0.0173	0.3792	0.1382	0.0613	0.0145
Usuário 4	0.4252	0.4064	0.2224	0.0249	0.1075	0.2465	0.0873	0.2673	0.3476	0.2327	0.1108	0.0346	0.2271	0.1925	0.0803	0.0222
Usuário 5	0.4236	0.5099	0.2138	0.0152	0.0729	0.2945	0.2042	0.2458	0.3865	0.2730	0.1595	0.0487	0.3534	0.1926	0.0992	0.0076
Usuário 6	0.6445	0.2213	0.1206	0.1329	0.0452	0.8140	0.6013	0.1860	0.2027	0.4585	0.0963	0.0000	0.0233	0.1794	0.1296	0.0000

Table 3. Exemplo da representação de perfil de consumo de usuários no arquivo gerado `profiles.csv`. Os valores sofreram normalização por proporção e independem do número de animes consumidos por usuário.

3.5 Geração de comunidades e do grafo

Possuindo o arquivo final de perfis de consumo, agora pode-se agrupar tais perfis em comunidades. Neste passo porém, pode-se observar uma grande perda amostral, cuja causa principal é a grande quantidade de perfis com pouco número de animes, que são outliers no conjunto de dados - um exemplo é o **Usuário 36**, que possui 100% de afinidade com os gêneros e mídias em sua linha por ter assistido apenas 1 anime, cujo Score foi 10. A solução para tal foi apenas expandir o número de usuários extraído para mesmo com esta perda ainda possuir o número de usuários suficiente para a geração de comunidades. Assim, dos 1000 usernames no arquivo original, 534 possuíam um perfil não vazio, e destes, 281 fizeram parte do resultado final de comunidades e recomendações, com os outros 253 sendo considerados inválidos. Para facilitar o processo, imediatamente eliminou-se todas aquelas contas com menos de 5 animes, pois estas garantidamente seriam outliers.

Outra modificação importante resultante dos resultados iniciais foi a utilização do *Term Frequency – Inverse Document Frequency* nas colunas de afinidade com gêneros. Isso garante que gêneros que aparecem na maioria absoluta dos animes têm uma relevância menor nos resultados de agrupamento, pois em resultados iniciais quase todas as comunidades eram diferentes variações dos gêneros *Ação*, *Aventura* ou *Comédia*. Também foi aplicado um valor maior nas colunas de *Source* para garantir que estas sejam relevantes na criação dos grupos.

Por fim, aplicou-se similaridade de cosseno aos valores finais, gerou-se o grafo de comunidades utilizando-se o layout *spring* da biblioteca NetworkX para geração das comunidades. Sempre que uma comunidade fosse definida por um conjunto de gêneros anteriormente utilizado, um terceiro gênero seria especificado para melhor entendimento da variação dos padrões de consumo. Este mesmo grafo com gêneros e fontes atribuídos a comunidade foi então utilizado para gerar recomendações com base na lista de mangás `anime_cache.csv`, associando-se a

afinidade da coluna de *Tipo* com as *fontes* das comunidades, gerando-se o arquivo `output.dat` e a imagem do grafo `graph.comm.png`

4 Análise de Resultados

4.1 Grafo Gerado

Ao final da execução obtêve-se o seguinte grafo:

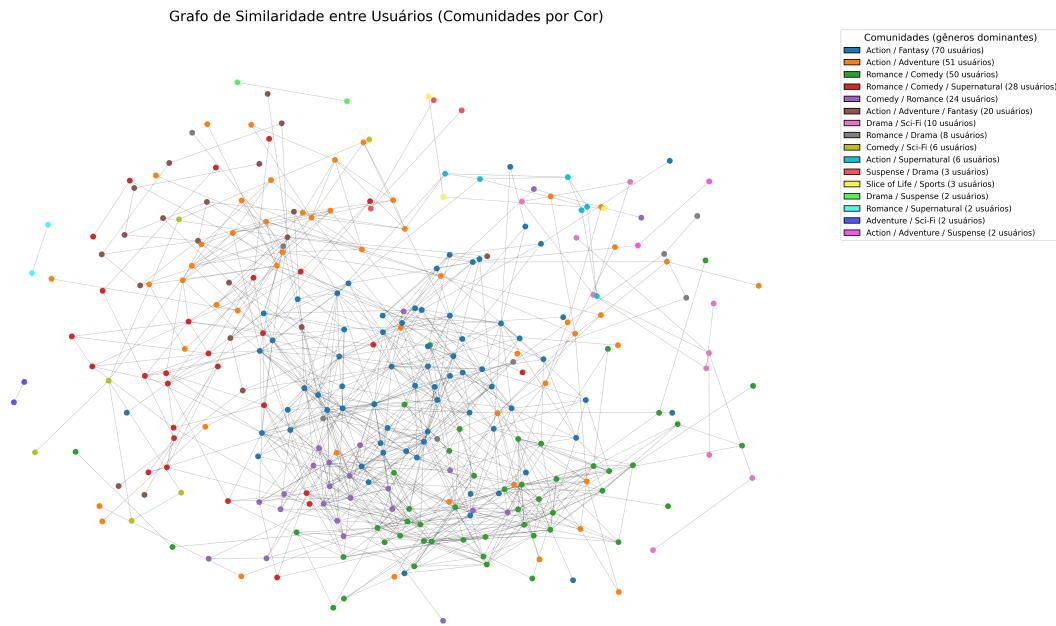


Figure 1. Grafo de comunidades de perfis de consumo dos usuários brasileiros do MyAnimeList com as cores dos vértices legendadas a partir dos gêneros dominantes da comunidade respectiva.

Já na imagem algumas conclusões interessantes podem ser tiradas, mas primeiramente é importante separar as comunidades *Romance/Comedy* (Verde), *Romance/Comedy/Supernatural* (Vermelho) e *Comedy/Romance* (Roxo). Apesar de parecerem se referir a mesma coisa, fica evidente a diferença entre os elementos que definem estas comunidades.

A comunidade vermelha, que é a mais distante das outras duas possui o gênero *Supernatural*, adicionando consequentemente elementos de *Fantasia* ao conjunto de interesses da comunidade. Por consequência pode-se observar que esta comunidade está mais próxima dos elementos de *Action/Adventure* (Laranja) e *Action/Adventure/Fantasy* (Marrom). Isso é um resultado esperado devido a grande proeminência da trope/subgênero *isekai*, que muito comumente adiciona elementos de comédia romântica em conjunto com os elementos já típicos de animes de Fantasia.

Em seguida é preciso colocar que *Comedy/Romance* (Roxo) e *Romance/Comedy* (Verde) apesar de próximos, não são a mesma coisa. Ambos podem ser descritos possivelmente como comédia romântica, porém é evidente que os usuários da comunidade roxa estão menos dispersos que os da comunidade verde. Isso se dá pelo fato de o foco maior da comunidade roxa ser nos elementos de comédia dos animes que assistem. Enquanto um anime pode facilmente inserir romance em sua narrativa e receber o gênero de romance sem isso entrar em conflito com outros elementos da sua estrutura, o gênero de comédia afeta diretamente a seriedade de outros elementos do anime, fazendo com que o gênero - e por conseguinte a comunidade interessada primariamente nos seus elementos - seja uma comunidade menos distribuída.

Action/Fantasy não só tem o maior número de usuários, como se localiza, em grande maioria no centro do grafo, mostrando que usuários que consomem este gênero estão muitas vezes próximos do perfil de consumo de outros gêneros. Algo que não pode ser dito de um gênero como *Comedy/Romance*, ou ainda melhor, *Suspense/Drama*, que claramente se localiza numa região isolada do grafo que pode ser atribuída a uma demografia diferente das outras regiões, tipicamente conhecida como *seinen*.

Para a melhor visualização destas regiões, uma versão do grafo por zonas de influência pode ser extremamente útil.

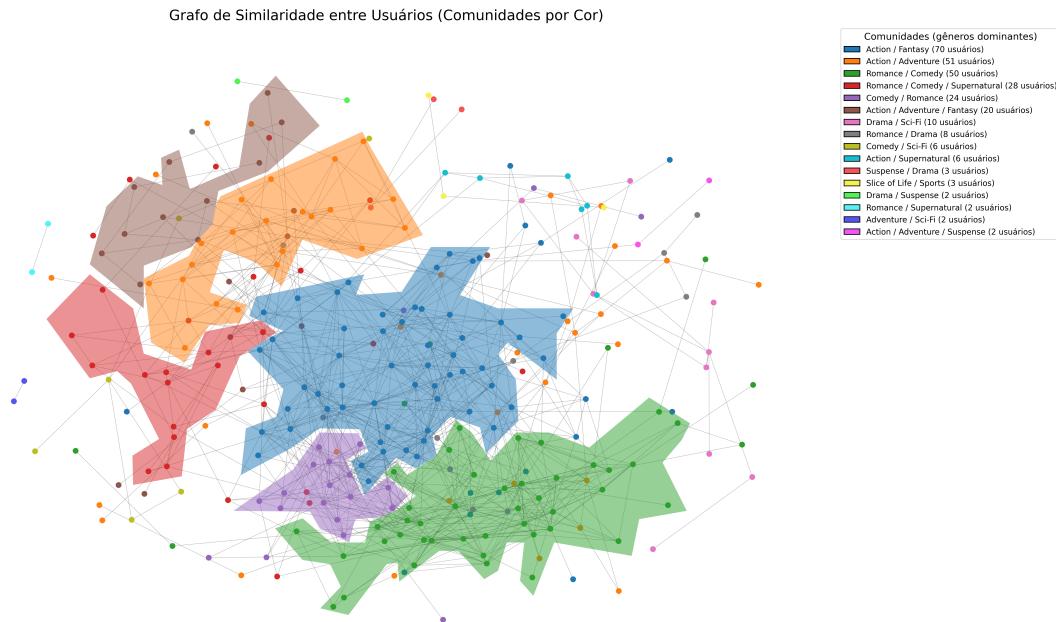


Figure 2. Grafo de comunidades de perfis de consumo dos usuários brasileiros do MyAnimeList com zonas de influência assinaladas a partir dos gêneros dominantes dos vértices da região respectiva.

No canto superior direito do grafo observa-se uma região com uma série de comunidades que não possuem vértices o suficiente de nenhuma delas para que seja considerada uma zona de influência. Verifica-se porém, que uma série de gêneros que descrevem estas comunidades podem, como dito anteriormente, designar animes de uma demografia diferente - e menor, quando comparada a massa de consumidores de animes - que consomem elementos mais adultos. Tais gêneros são *Drama*, *Sci-Fi*, *Suspense*, e *Slice of Life*. E pode-se também considerar o gênero *Sports* como parte deste conjunto por este tipicamente estar associado com *Slice of Life*, como mostrado na sua comunidade no grafo.

4.2 Obras recomendadas para adaptação

O seguinte output.dat gerado após a execução possui a seguinte estrutura.
 === RESULTADO DA CLUSTERIZAÇÃO E RECOMENDAÇÃO ===

Total de comunidades: 16

```
[COMUNIDADE 1] (70 usuários)
Foco Principal: Origem - Light_Novel. Gêneros - Action, Fantasy
--- Top 5 Mangás Não Adaptados Recomendados ---
[0.7878] Yu-Gi-Oh!
- Score MAL: 7.33 | Tipo: Light Novel | Gêneros: Action, Adventure, Comedy, Drama, F
[0.7782] Code Name wa Sailor V (Codename: Sailor V)
- Score MAL: 7.50 | Tipo: Manga | Gêneros: Action, Adventure, Comedy, Drama, Fantasy
[0.7782] Watashi no Messiah-sama
- Score MAL: 7.62 | Tipo: Manga | Gêneros: Action, Adventure, Comedy, Drama, Fantasy
[0.7727] One Piece Novel: A (One Piece Novel: Ace's Story)
- Score MAL: 7.93 | Tipo: Light Novel | Gêneros: Action, Adventure, Comedy, Fantasy
[0.7727] Fairy Tail
- Score MAL: 7.54 | Tipo: Light Novel | Gêneros: Action, Adventure, Comedy, Fantasy
```

[COMUNIDADE 2] (51 usuários)

[...]

Observou-se que em diversas comunidades houve a grande predominância de mangás, enquanto em outras houve uma maior distribuição entre mangás e light novels, algo que pode ser atribuído tanto ao maior consumo destas obras pelos membros da comunidade quanto ao maior número de light novels associadas aos gêneros definidos.

Outra coisa interessante de se observar é que recomendações como Yu-Gi-Oh, One Piece e Fairy Tail se referem a franquias que já foram adaptadas para anime em algum momento. O motivo que elas aparecem na recomendação final é porque estas obras em específico são histórias originais nos universos destas franquias, histórias estas que não foram adaptadas para anime, e portanto ainda fazem parte do conjunto de dados considerado para a recomendação. Muito pelo contrário, a presença destas obras já reconhecidas pode ser considerado um fator que comprova a efetividade do algoritmo de recomendação, já que pode-se assumir que novas obras destas franquias seriam bem sucedidas comercialmente com suas comunidades respectivas.

Outros dados importantes a pontuar são a maior afinidade encontrada na recomendação, com 97.21%, na comunidade *Slice Of Life/Sports*, que curiosamente gerou recomendações focadas em Terror e Mistério. Isso pode ser atribuído ao baixo número de usuários na comunidade, resultando em uma alta afinidade coincidental de todos os usuários com o gênero Terror. Já a menor afinidade encontrada é da comunidade Drama/Sci-Fi, com 72.13% de afinidade, provavelmente porque com o peso escolhido para a nota, as últimas recomendações não foram do gênero Sci-Fi. Por fim, a obra com maior nota que foi recomendada para receber uma adaptação para anime é o mangá **Spirit Circle**, com uma nota de 8.47 e os gêneros *Action, Adventure, Comedy e Romance*

5 Conclusão

Conclui-se portanto, que o algoritmo desenvolvido conseguiu extrair e tratar os dados da plataforma MyAnimeList, além de fazer o uso destes dados para chegar a uma representação fidedigna da divisão do conjunto de consumidores da mídia de animações japonesas.

As recomendações finais também podem ser consideradas válidas para o conjunto de fontes e gêneros identificado no grafo, apesar de que um trabalho futuro poderia melhorar o sistema de recomendação através de testes mais rigorosos dispondo de uma base de dados maior.

Uma possível forma de fazer isso seria a utilização de outro site de notas de animes, como o AniList, que possui uma API mais acessível, ou ainda a utilização de uma base de dados pronta como a do próprio Netflix, apesar de que seria necessário encontrar uma base de dados que estivesse atualizada depois da pandemia devido às grandes mudanças no mercado já apresentadas na introdução.

Assim, por fim, reitera-se a necessidade da criação de novos mecanismos que permitam a identificação de estratégias comerciais para a criação de entretenimento de qualquer natureza e a importância que os dados de consumo destas mídias podem ter para empresas que participam do processo de produção destas obras, sendo plataformas de notas uma fonte muito rica destes dados e uma forma de estudar o comportamento da massa consumidora.

6 Créditos

- **Arthur Santana de Mesquita:** Extração, Implementação, Análise de Resultados, Escrita do Artigo.
- **Higor Alexandre Duarte Mascarenhas:** Orientador e Professor de Ciência dos Dados.
- **Michel Pires da Silva:** Orientador e Professor de Algoritmos e Estrutura de Dados.

7 Referências

References

- [1] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, Cambridge, MA, 4 edition, 2022.
- [2] Leonard Richardson. Beautiful soup documentation, 2024. URL <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>. Accessed: 2025-12-12.
- [3] NetworkX Developers. Networkx documentation, 2025. URL <https://networkx.org/documentation/stable/>. Accessed: 2025-12-12.