Um estudo demográfico sobre *animes*

Carvalho, Claudio
cco2@cin.ufpe.br

Gouveia, Guilherme
gglf@cin.ufpe.br

Tavares, Henrique
lhtc@cin.ufpe.br

Novembro, 2018

1 Introdução

Define-se como *anime*[1] (*animê*, em português brasileiro) qualquer animação produzida por estúdios japoneses. Apesar de variarem quanto ao gênero cinematográfico - comédia, drama, ação - animes tipicamente compartilham de um mesmo estilo artístico que atraiu diversos fãs ao redor do mundo. Culturalmente, portanto, possuem devida relevância.

Neste projeto, estudaremos mais a fundo padrões demográficos para cada estilo individual de anime, tomando como base atributos de espectadores como localização geográfica, idade e gênero. Além disso, queremos procurar entender como algumas características dos animes podem estar ligadas com seu público demográfico, como quantidade de episódios, produtora e data de lançamento.

2 Coleta dos dados

MyAnimeList é um *website* que fornece um amplo banco de dados sobre diversos títulos de *animes*. Seus usuários podem, portanto, criar listas dos episódios que assistiram ou que ainda desejam assistir, podendo, ainda, avaliá-los de acordo com sua qualidade. Portanto, existe um grande potencial quanto aos dados disponíveis.

No entanto, esses dados não são disponíveis diretamente pelo *website*. Portanto, uma busca mais confiável sobre onde podemos encontrá-los mostrou-se relevante. Confiavelmente, podemos realizar pesquisa no *Kaggle*, uma comunidade para cientistas e analistas de dados, muito útil para encontrar conjuntos.

Em nossa busca, encontramos o conjunto que queríamos em bom formato e atualizado constantemente. O autor dividiu os dados em três entidades - Anime, User e UserAnime - sendo a última caracterizando o relacionamento entre as duas primeiras.

Ainda para maior conveniência, o autor filtrou os dados que considerava falhos pelo *website*, como usuários com data de nascimento datadas antes do

ano 1900 ou aqueles que reportaram números absurdos de episódios assistidos. Essas escolhas do autor facilitaram nosso trabalho de pré-processamento.

3 Filtragem dos dados

Como mencionado no item anterior, os dados já possuem algum tipo de pré-processamento. No entanto, a quantidade de linhas na entidade UserAnime mostrou-se bastante extensa - o arquivo que a descreve contém 5GB. Essa extensão inviabiliza nossas pesquisas pois, mesmo que possuíssemos poder computacional para processá-los em tempo hábil, não é possível carregar todos esses dados diretamente na memória. Chegamos à conclusão de que algum método de redução deveria ser aplicado.

Como de costume para esse tipo de problema, o desafio será manter a mesma congruência dos dados, mesmo após reduzidos. Portanto, qualquer que fosse a medida utilizada para redução do conjunto de dados deveria levar isso em consideração. Após avaliação, as duas melhores medidas encontradas foram:

1. Eliminar as linhas do UserAnime que possuíssem baixa quantidade de episódios assistidos: nesse caso, estabeleceríamos um limiar de proporção mínima de episódios assistidos por quantidade de episódios do anime. Isto mostraria-se eficaz, já que nosso objetivo está interessado apenas nos usuários que demonstram interesse ao invés dos que apresentam falta dele. No entanto, para realizar essa filtragem, seria necessário analisar cada linha da entidade, o que se mostrou bastante lento e;
2. Selecionar aleatoriamente linhas do banco de dados: após estabelecer o número desejado de elementos, basta gerar um conjunto aleatório de índices e selecioná-los no conjunto de dados. Apesar de simples, mostramo-nos satisfeitos com este método. Calculamos que, dado a extensão do banco de dados original, a probabilidade de um filtro randômico enviesado é muito baixa.

Os dois métodos foram implementados em nossa pesquisa. Porém, por possuir tempo de processamento significativamente mais curto, optamos utilizar o segundo em nossos testes.

4 Pré-processamento dos dados

Após filtrar os dados, processá-los tornou-se uma tarefa menos árdua, uma vez que podemos acessar todos os dados diretamente da memória. No entanto, para nossos estudos, queremos utilizar algumas informações que, apesar de não explícitas diretamente no banco de dados, podem ser inferidas.

Primeiramente, temos que relacionar as entidades entre si. Para tal, a função *merge* da biblioteca Pandas mostrou-se útil, permitindo relacionar dois conjuntos de dados a partir de uma coluna chave. Relacionar animes e usuários foram realizados através das colunas *animeId* e *username* respectivamente.

Consideramos que idade e região dos usuários é de suma importância para descobrir a demografia de um *anime*. Esses dados podem ser inferidos a partir

das colunas *birthdate* e *location*. Calcular a idade através da data de nascimento é um tanto quanto trivial, sendo apenas necessário filtrar dados indesejados como datas inválidas. Região, no entanto, mostrou-se um pouco mais complexo.

O *website MyAnimeList* permite aos usuários fornecerem suas regiões livremente, apenas com um campo de *string*. Isso levou a várias respostas sem valor agregado - como "Minha casa" - que não explicitava a localização geográfica do usuário em questão. Respostas nesse formato precisaram ser filtradas.

Outras respostas se mostraram mais úteis, apesar de não bem formadas. Notamos que é bastante comum para usuários estadunidenses não fornecerem seu país de origem, mas uma dupla cidade/estado. Portanto, para termos acesso ao seu país, precisamos relacionar com um dicionário de estados americanos. Caso houvesse um casamento, poderíamos inferir o país desse usuário com bastante certeza.

Ademais, por padronização, precisamos que todos os países possuam a mesma grafia. Precisaríamos tratar países ou regiões nos casos de sensibilidade ao idioma ("Brazil" e "Brasil"), capitalização ("Brazil" e "brazil") e nomes completos contra populares ("República Federativa do Brasil" e "Brazil"). Para tal, uma abordagem similar aos estados americanos foi aplicada - relacionar com um dicionário de países ao redor do mundo com diferentes grafias.

Finalmente, como uma medida extra de avaliação da afinidade de um usuário ao *anime*, julgamos necessário adicionar a coluna *myCompletion*, que calcula a proporção de episódios assistidos pelo usuário pelo total de episódios lançados de um *anime*.

5 Estória dos dados

Neste momento, vamos nos utilizar dos dados processados para analisá-los e chegarmos em conclusões. Primeiramente, queremos visar os *animes* populares e encontrar similaridades demográficas entre eles, com o objetivo de entender a comunidade de forma generalizada. Por último, vamos estudar detalhes mais intrínsecos como produtoras e gêneros, de forma a encontrar relações diversas.

5.1 Análise de animes populares

Após organizarmos animes de acordo com sua popularidade, testamos três diferentes parâmetros:

5.1.1 Origem

Animes podem possuir inspirações de outras formas de arte. Animações ocidentais, tradicionalmente, partem de ideias originais. No entanto, o mesmo caso não pode ser dito sobre as orientais que, em grande maioria, possuem *mangas* como material de origem.

Manga[2] (*mangá*, em português brasileiro) é o termo japonês utilizado para descrever histórias em quadrinhos. Apesar disso, globalmente, o termo possui

significado para um estilo artístico específico, marcado por traços semelhantes ao *anime*. Algumas surpresas emergiram, no entanto. *Animes* originais não são a segunda origem mais popular, este lugar fica com o *Light Novel* - um estilo de *manga*, porém, escrito em prosa. É interessante notar a separação entre *manga* e *web manga*, que também conseguiu um lugar na lista.

5.1.2 Gêneros dos espectadores

Assim como no item 5.1.1, também era esperado qual valor encontraria-se em primeiro lugar nessa lista. O gênero masculino predomina sobre os outros. Porém, surpreende a presença - apesar de mínima - o gênero não-binário. Uma boa demonstração de inclusão por parte do *website*.

5.1.3 Localização

Aqui, encontramos algumas surpresas. Intuitivamente, imagina-se forte presença de espectadores japoneses. No entanto, mal se encontram neste ranking. Supõe-se que o idioma do *website MyAnimeList*, por ser inglês, não seja popular em países orientais.

O primeiro lugar, como se esperado em pelo menos alto *ranking*, encontram-se os Estados Unidos, forte consumidor da cultura japonesa. O Brasil encontra-se na próxima colocação.

5.2 Análises específicas

Agora que realizamos um estudo generalizado, podemos partir para atributos mais específicos, visando melhor restrição demográfica. Assim como no item anterior, utilizamos diferentes parâmetros para encontrar diferentes conclusões. Nesta sessão, escolhemos gráficos mais interativos.

5.2.1 Ranqueamento x Popularidade

Queríamos entender se animes populares tendiam a ser melhores ranqueados. Ao contrário do que se é intuitivo, este não é o caso. Porém, esse estudo revela diferentes conclusões.

Primeiramente, apenas trocando a ordem dos animes a serem exibidos, pode-se notar semelhanças e diferenças interessante entre os gráficos. O *anime Fullmetal Alchemist: Brotherhood* mostrou ter correlação entre popularidade e boas avaliações, pois, em ambos os gráficos, apareceu em primeiro lugar.

Se ordenarmos por ranqueamento, notamos que alguns *animes* não são tão populares, mas possuem excelentes avaliações. Isso provavelmente indica uma preferência de nicho, como é o caso de *Gintama: Shirogane no Tamashii-hen*.

Quando organizamos por popularidade, podemos perceber *animes* que, apesar de alta popularidade, possuem baixas avaliações, como é o caso de *Sword Art Online II*.

5.2.2 Estação do ano de lançamento por gênero de *anime*

Através dessa análise, procuramos relacionar se a época de lançamento pode revelar conclusões interessantes. Notamos que a época do verão é a que possui menor quantidade de lançamentos. Interessantemente, *animes* lançados nessa estação possuem piores avaliações.

5.2.3 Relação com produtoras

Realizamos um estudo sobre as produtoras dos *animes*, procurando entender seu apelo demográfico. Encontramos as produtoras que mais lançam *animes*, e também que tipo de gênero mais lançam em dados absolutos e porcentagem. Assim, seria mais fácil de demonstrar quais as produtoras grandes e qual foco cada produtora tinha como principal. Isso seria interessante também para uma próxima análise onde vimos a quantidade de pessoas de cada gênero que assiste animes daquela produtora. Em geral, tudo segue o padrão de terem mais homens assistindo, mas é interessante observar desvios nesse padrão em certas produtoras, e em alguns casos, que produzem gêneros bem voltados para o público masculino. Também é relevante mencionar que em números absolutos, a representatividade de pessoas de gênero não-binário vendo algumas produtoras é expressivo, chegando a casa dos milhares em diversos casos.

References

[1] Lesley Aeschliman. Bellaonline. *What is Anime?*. 7 de novembro de 2007

[2] Prada, Edit *Manga ou Mangá?* 19 de abril de 2004