

**UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSOS DE CIÊNCIA E ENGENHARIA DA COMPUTAÇÃO**

**DISCIPLINA DE CLASSIFICAÇÃO E PESQUISA DE DADOS
PROF. LEANDRO KRUG WIVES**

RELATÓRIO TRABALHO FINAL

Equipe: “The Walking Data”

Integrantes: Brenda Schussler e Pedro Rigon

DESCRIÇÃO DO PROBLEMA

O surgimento e desenvolvimento das plataformas de streaming resultou também em uma revolução no mercado das produções audiovisuais. Nesse ramo, a empresa norte americana Netflix se tornou uma potência dominante e com isso filmes e séries têm influenciado, cada vez mais, o imaginário coletivo da população. Dessa forma, justificado pela sua relevância socioeconômica, o conjunto de dados que selecionamos para trabalhar contém as séries e filmes disponíveis na Netflix (atualizado em junho de 2021), o qual foi retirado do site do IMDB. Com isso, nossa aplicação trata-se de um catálogo com os títulos da Netflix, sendo possível fazer classificação e pesquisas nestes dados de acordo com filtros pré-estabelecidos.

PROJETO DE ARQUIVOS:

Inicialmente, foi feita a organização (divisão) dos dados em arquivos, separando-os de acordo com suas categorias. No arquivo principal (NetflixVideosDataCPD), o id_imdb de cada título está associado aos ids das suas respectivas informações (language, startYear, type, country, rankpop).

Nos demais arquivos, os ids de cada uma das categorias estão associados as suas respectivas informações, para que, quando feita a busca (filtragem) dos dados, as informações coletadas a respeito de cada título sejam corretas.

ORGANIZAÇÃO DOS ARQUIVOS:

A organização de arquivos foi baseada no modelo estrela, no qual há um arquivo principal que contém chaves de índice de acesso ao conteúdo que se encontra em arquivos auxiliares. Tal modelo foi aplicado devido a intensa repetição de dados que o DataBase carregava consigo, afim de facilitar a normalização e compactação dos dados através da remoção de dados repetidos ou inúteis e da sua padronização. Tais aspectos facilitam o trabalho de filtragem de dados e aceleram a busca dos dados através de arquivos armazenados em disco.

Para transformar o DataBase em um modelo estrela, utilizamos o programa Power BI do Windows que possui funções específicas para o tratamento de dados, como por exemplo: remoção de termos redundantes e inválidos, ligação entre colunas e inserção de índices nas colunas. Utilizando tais funções, moldamos nosso arquivo csv de entrada, obtendo como resultado os dados no seguinte modelo estrela:

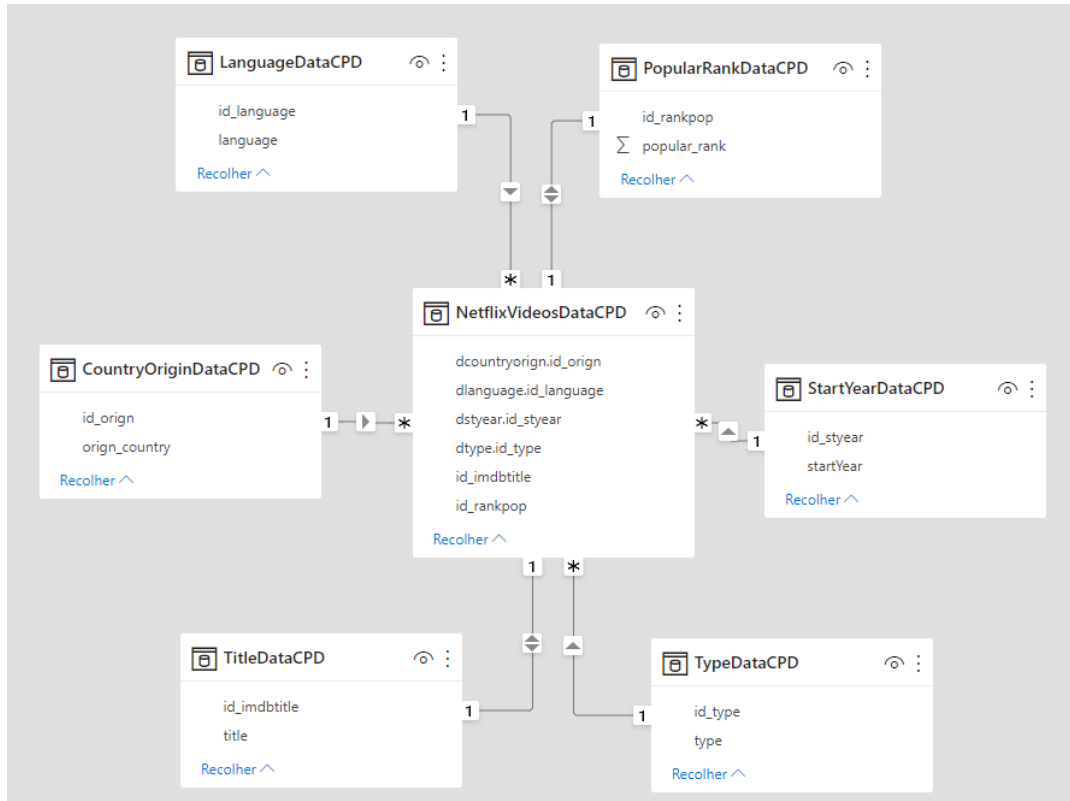


Figura 1: Modelo Estrela feito utilizando o Power BI.

Desta forma, obtivemos 7 arquivos csv, sendo 1 deles o arquivo central (NetflixVideosDataCPD), o qual possui apenas os índices de acesso às informações dos demais arquivos, e os outros 6, contendo suas respectivas informações específicas referenciadas pelo índice de acesso contido no arquivo csv principal.

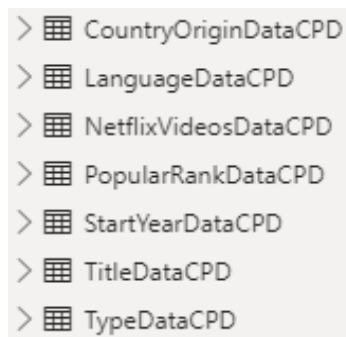


Figura 2: Arquivos resultantes da transformação dos dados

Além disso, foram tratados alguns problemas de informação contidos nos dados inicialmente extraídos, como:

- Espaço antes dos títulos;
- Letras maiúsculas;
- Letras desconfiguradas;
- Caracteres inválidos.

Para isso, utilizamos as próprias funções presentes no Excel, dentre elas ARRUMAR1(coluna) e MAIÚSCULA(coluna), além disso, implementamos através do modo desenvolvedor, uma função que remove, se existir na linha, caractere de <space> no início do texto.

	A	B	C	D	E	F
1	id_imdbtitle	id_language	id_year	id_type	id_country	id_rankpop
2	tt1064899	1	1	1	1	48
3	tt1734135	1	1	1	1	816
4	tt2788432	1	1	1	1	190
5	tt3322314	1	1	1	1	467
6	tt3787402	1	1	1	1	4387
7	tt3986586	1	1	1	1	549
8	tt4047038	1	1	1	1	470
9	tt4052886	1	1	1	1	1
10	tt4061080	1	1	1	1	405
11	tt4145054	1	1	1	1	313
12	tt4181172	1	1	1	1	458
13	tt4209256	1	1	1	1	248
14	tt4270492	1	1	1	1	89
15	tt4532368	1	1	1	1	80
16	tt4574334	1	1	1	1	23
17	tt4588068	1	1	1	1	4767
18	tt4592410	1	1	1	1	1437
19	tt4635282	1	1	1	1	348
20	tt4789300	1	1	1	1	1784
21	tt4955642	1	1	1	1	122
22	tt4971144	1	1	1	1	879
23	tt4973548	1	1	1	1	883
24	tt4998212	1	1	1	1	315
25	tt5028002	1	1	1	1	6
26	tt5075942	1	1	1	1	4746
27	tt5151816	1	1	1	1	151
28	tt5179408	1	1	1	1	1037
29	tt5228026	1	1	1	1	2430
30	tt5235950	1	1	1	1	3131

Figura 3: Arquivo Csv principal

A figura 3 exemplifica como ficou estruturado o arquivo csv principal, sendo que nele há apenas os índices de acesso às informações que se encontram nos outros arquivos, de modo, que cada linha apresenta a informação relacionada com o id_title. Portanto, cada linha carrega consigo índices de acesso à informação de um determinado título presente no streaming Netflix.

Para a ligação desses arquivos foram criadas árvores trie como índice de acesso ao arquivo principal. Há, árvores trie normais e invertidas, afim de facilitar o acesso, aumentar a velocidade dos filtros de busca e diminuir o tempo de execução das filtragens. Ao todo, foram criados os seguintes arquivos binários em forma de árvores trie:
















 CabecalhoPrincipal	5/9/2022 8:06 PM	Arquivo BIN
 country_trie	5/9/2022 8:06 PM	Arquivo BIN
 country_trieinvertido	5/9/2022 8:06 PM	Arquivo BIN
 language_trie	5/9/2022 8:06 PM	Arquivo BIN
 language_trieinvertido	5/9/2022 8:06 PM	Arquivo BIN
 netflix_trie	5/9/2022 8:06 PM	Arquivo BIN
 NetflixVideosDataCPD	5/9/2022 8:06 PM	Arquivo BIN
 rankpop_trie	5/9/2022 8:06 PM	Arquivo BIN
 rankpop_trieinvertido	5/9/2022 8:06 PM	Arquivo BIN
 styear_styear	5/9/2022 8:06 PM	Arquivo BIN
 styear_styearinvertido	5/9/2022 8:06 PM	Arquivo BIN
 title_trie	5/9/2022 8:06 PM	Arquivo BIN
 title_trieinvertido	5/9/2022 8:06 PM	Arquivo BIN
 type_trie	5/9/2022 8:06 PM	Arquivo BIN
 type_trieinvertido	5/9/2022 8:06 PM	Arquivo BIN

Figura 4: Arquivos binários criados para a execução do programa

FUNCIONALIDADES DA APLICAÇÃO:

A aplicação é organizada em menus, que permitem que o usuário realize o que desejar (dentro das funcionalidades disponíveis) com os dados trabalhados.

1. **Pesquisa de dados:** ao selecionar a opção 2) no menu principal, o usuário terá acesso ao menu de filtros, que permite fazer a pesquisa de dados de acordo com diversos filtros/categorias, conforme listados a seguir:
 1. Imdb_id: ao escolher essa opção, o usuário filtra a obra desejada buscando-a através do seu código de Imdb_id.
 2. Language: ao escolher essa opção, o usuário poderá buscar quais obras (dentro as contidas nos dados) tem como idioma original o digitado na busca do filtro. Por exemplo, ao filtrar por “french” a aplicação irá mostrar ao usuário quais os títulos possuem o idioma francês como idioma original.
 3. Popularity: essa opção permite ao usuário filtrar a obra de acordo com sua popularidade no ranking geral da Netflix. Por exemplo, ao buscar por “5” a aplicação irá apresentar ao usuário qual é o título que ocupa a quinta posição no ranking.

4. Start Year: ao selecionar essa opção, o usuário deve inserir um ano, e em seguida a aplicação mostrará quais títulos tiveram sua data de estreia no respectivo ano buscado.
5. Type: permite que o usuário busque quais as obras são do tipo buscado, sendo possível buscar as opções: tvseries, movie, tvspecial, tvminiseries, short, vídeo, tvmovie, tvshort, videogame, tvepisode.
6. Origin Country: essa opção permite que o usuário filtre as obras de acordo com seu país de origem. Por exemplo, ao buscar por “Brazil” a aplicação apresentará quais os títulos têm “Brazil” como país de origem.

Também é possível fazer uma busca por prefixos no menu principal ao selecionar a opção 1). Ou seja, escolhendo a primeira opção no menu principal, o usuário pode pesquisar quais obras constam no catálogo tendo em seu título o prefixo pesquisado. Por exemplo, ao buscar “the” a aplicação irá listar todos as obras que tem como prefixo “the” em seus títulos. Ainda dentro desta opção, após a listagem dos títulos de acordo com a filtragem de prefixo inserida, o usuário pode escolher um dos títulos da lista para consultar todas as informações referentes a ele.



Figura 5: Menu Principal



Figura 6: Menu de Filtros buscando o título pelo seu imdb_id.

```
Menu de Filtros

Menu Interativo
-----
Buscar filme através do filtro:
[1] - Imdb_id
[2] - Language
[3] - Popularity
[4] - Start Year
[5] - Type
[6] - Origin Country
[7] - Sair do menu de filtros
Sua opcao: 2
Digite a linguagem: french
=====
Foram encontrados as seguintes obras no DataBase:
=====

17 filmes
8 rue de l'humanite
a tombeau ouvert
ad vitam
ad vitam
adn
adu
alors, heureux?
anelka: misunderstood
asterix
asterix
au service de la france
au service de la france
balle perdue
bigbug
blanche comme neige
blockbuster
boi
braqueurs
break
bronx
burn out
c'est quoi cette famille?!
caid
caid
climax
coco avant chanel
```

Figura 7: Menu de Filtros buscando os títulos pelo idioma.

```
Menu de Filtros

Menu Interativo
-----
Buscar filme através do filtro:
[1] - Imdb_id
[2] - Language
[3] - Popularity
[4] - Start Year
[5] - Type
[6] - Origin Country
[7] - Sair do menu de filtros
Sua opcao: 3
Digite a Popularidade: 4
=====
Foram encontrados as seguintes obras no DataBase:
=====

friends
```

Figura 8: Menu de Filtros buscando o título pela popularidade.

```
Menu de Filtros

Menu Interativo
-----
Buscar filme através do filtro:
[1] - Imdb_id
[2] - Language
[3] - Popularity
[4] - Start Year
[5] - Type
[6] - Origin Country
[7] - Sair do menu de filtros
Sua opcao: 4
Digite o Ano de Estreia: 2005
*****
Foram encontrados as seguintes obras no DataBase:
*****

avatar: the last airbender
bleach: burichi
courage & stupidity
doom
gamunui wigi: gamunui yeonggwang 2
get rich or die tryin'
grey's anatomy
hans liberg: tatatata
hell's kitchen
how i met your mother
james
johnny test
little einsteins
lord of war
meerkat manor
mörke
naboer
prison break
puppy
racing stripes
room
save the forest
supernatural
sweet moves
the adventures of sharkboy and lavagirl 3-d
the boondocks
```

Figura 9: Menu de Filtros buscando os títulos pelo ano de estreia.

```
Buscar filme através do filtro:
[1] - Imdb_id
[2] - Language
[3] - Popularity
[4] - Start Year
[5] - Type
[6] - Origin Country
[7] - Sair do menu de filtros
Sua opcao: 5

Tipos de Obras Disponíveis:
- tvseries
- movie
- tvspecial
- tvminiseries
- short
- video
- tvmovie
- tvshort
- videogame
- tvepisode

Digite o Tipo de Obra: short
*****
Foram encontrados as seguintes obras no DataBase:
*****

#happybirthdaysense8
13 reasons why: season 2 date announcement commercial
a 3 minute hug
a final cut for orson: 40 years in the making
a love song for latasha
a tale of two kitchens
after maria
after the raid
all in my family
american factory: a conversation with the obamas
angela's christmas
anima
antiracist baby
att doda ett barn
audible
bill hicks: reflections
birders
canvas
```

Figura 10: Menu de Filtros buscando os títulos pelo tipo.

```
Menu de Filtros

Menu Interativo
-----
Buscar filme através do filtro:
[1] - Imdb_id
[2] - Language
[3] - Popularity
[4] - Start Year
[5] - Type
[6] - Origin Country
[7] - Sair do menu de filtros
Sua opcao: 6
Digite o Pais de Origem: Brazil
=====
Foram encontrados as seguintes obras no DataBase:
=====
0.03
a garota invisivel
a primeira tentacao de cristo
a toca
afonso padilha: alma de pobre
alice junior
anitta: made in honorio
apenas o fim
boca a boca
bom dia, veronica
bruna surfistinha
bruno motta: melhor que os outros stand ups que eu ja fiz em 15 anos de carreira
bruno motta: o show do ano
cabras da peste
carnaval
cidade invisivel
cinderela pop
coisa mais linda
com a palavra, arnaldo antunes
crisalida
diario de um exorcista - zero
edmilson filho: notas, uma comedia de relacionamentos
emicida: amarelo - it's all for yesterday
encomenda
especial de ano todo com clarice falcao
especial de natal: se beber, nao ceie
faroeste caboclo
```

Figura 11: Menu de Filtros buscando os títulos pelo país de origem.


```
Menu Principal

Menu Interativo
-----
Opcoes:
[1] - Buscar NetflixTvShows
[2] - Filtrar Obra
[3] - Inserir Obra
[4] - Remover Obra
[5] - Top 10
[6] - Ordenar NetflixTvShows
[7] - Encerrar Programa
Sua opcao: 1
Digite o prefixo para Busca: Gre

Foram encontrados as seguintes obras no DataBase:

1) grease
2) great men academy
3) great news
4) great pretender
5) greatest events of wwii in colour
6) green beret's guide to surviving the apocalypse
7) green cross
8) green door
9) green eggs and ham
10) green is gold
11) green room
12) greenhouse academy
13) greenleaf
14) greg davis: you magnificent beast
15) grego rossello: disculpe las molestias
16) grenseland
17) grey's anatomy

Digite o numero da obra que deseja obter informacoes: 17
|||||
TITULO DA OBRA: GREY'S ANATOMY
LINGUAGEM ORIGINAL: ENGLISH
ANO DE ESTREIA: 2005
TIPO DE OBRA: TVSERIES
PAIS DE ORIGEM: UNITED STATES
RANKING DE POPULARIDADE ENTRE USUARIOS DA NETFLIX: 7
|||||
```

Figura 12: Menu Principal buscando por prefixo e em seguida exibindo as informações do título desejado dentre os listados com o respectivo prefixo.

2. **Ordenação dos dados:** ao selecionar a opção 6) do menu, o usuário entra no menu de ordenação, que lhe permite ordenar os títulos de acordo com os seguintes critérios: Ordem Alfabética Normal, Ordem Alfabética Inversa, Ranking Popularidade em ordem normal, Ranking Popularidade em ordem inversa.



Figura 13: Menu de Ordenação



Figura 14: parte da ordenação em ordem alfabética normal

```
bad samaritans
bad papa
bad investigate
bad day for the cut
bad boy billionaires: india
bad boy
bad blood
backstabbed
backlash
backfire
back with the ex
back to the outback
back street girls
babylon
baby reindeer
baby driver
baby ballroom
baby
babies behind bars
babies
baahubali: before the beginning
b: the beginning
azizler
aziz ansari: right now
aziz ansari: buried alive
aziz ansari live in madison square garden
azali
ayotzinapa, el paso de la tortuga
axone
away
awake: the million dollar game
awake
avlu
avicii: true stories
avengement
avatar: the last airbender
avatar: super deformed shorts
ava's possessions
autumn
autohead
autistic driving school
aurora
aunty donna's big ol' house of fun
auntie claus
audrie & daisy
audible
au service de la france
atypical
att doda ett barn
att angora en brygga
atone
atomic blonde
atlantique
athlete a
```

Figura 15: parte da ordenação em ordem alfabética inversa.

```
5 : ragnarok
6 : startup
7 : grey's anatomy
8 : sweet tooth
9 : the blacklist
10 : jupiter's legacy
11 : dirty john
12 : the walking dead
13 : peaky blinders
14 : shadow and bone
15 : breaking bad
16 : the woman in the window
17 : bo burnham: inside
18 : the mitchells vs the machines
19 : blue miracle
20 : love, death & robots
21 : ncis: naval criminal investigative service
22 : trouble
```

Figura 16: parte da ordenação pelo ranking em ordem normal.

```
100 : arrested development
99 : blue bloods
98 : sherlock
97 : the sinner
96 : jurassic park
95 : arrow
94 : lupin
93 : spider-man: far from home
92 : black space
91 : titans
90 : the umbrella academy
89 : billions
88 : agents of s.h.i.e.l.d.
87 : sex education
86 : star wars: the clone wars
85 : hawaii five-0
84 : riverdale
83 : family guy
82 : seinfeld
81 : the serpent
80 : legends of tomorrow
79 : black house
```

Figura 17: parte da ordenação pelo ranking em ordem inversa.

3. **Top 10:** ao selecionar a opção 5) do menu principal, o usuário terá acesso ao menu Top 10, onde poderá listar os “Top 10” (primeiros 10 classificados de acordo com o ranking), de acordo com as categorias: Top 10 Filmes e Top 10 Series.

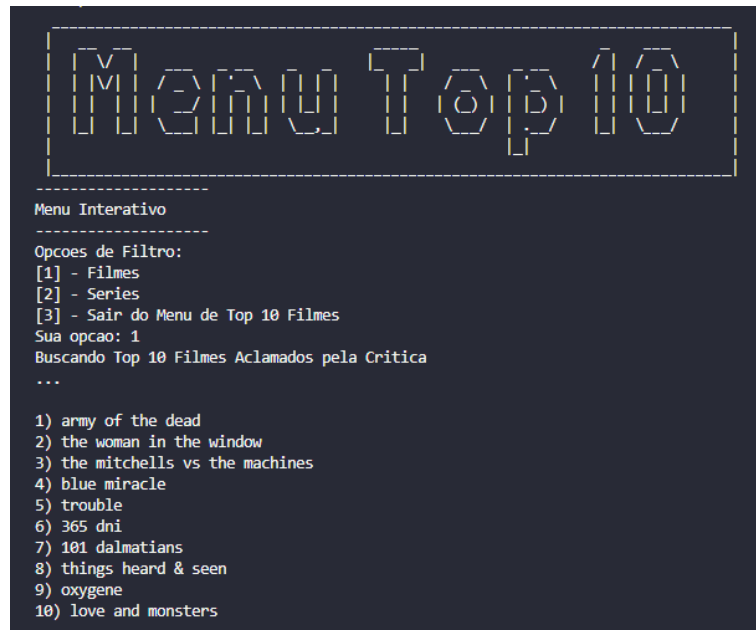


Figura 18: menu Top 10 exibindo Top 10 Filmes.

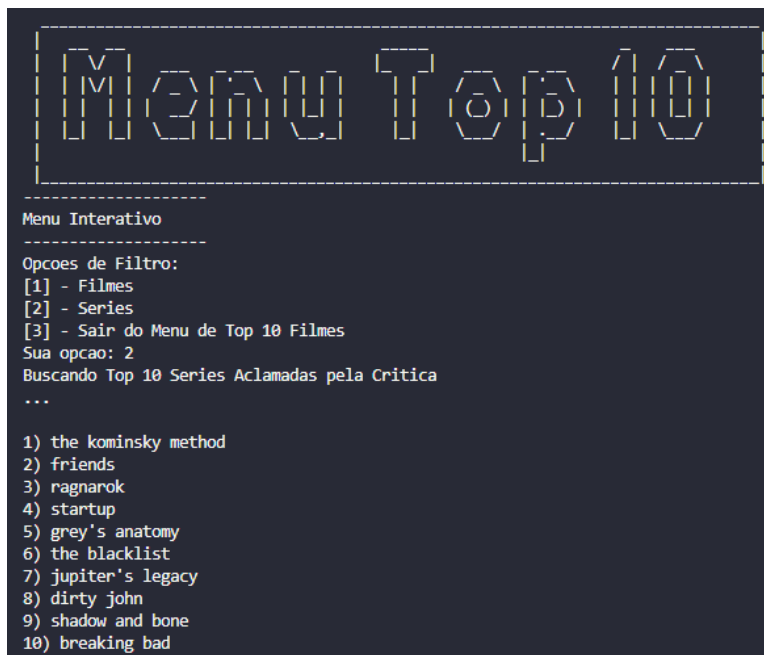


Figura 19: menu Top 10 exibindo Top 10 Series.

4. **Inserção e Remoção de dados:** ao escolher a opção 3) do menu principal, o usuário poderá inserir uma nova obra no catálogo, assim como, ao escolher a opção 4) poderá excluir (remover) uma obra do catálogo.

BIBLIOTECAS UTILIZADAS:

- **pandas:** está biblioteca foi utilizada para fazer a limpeza inicial dos dados, e para auxiliar na implementação das funções de inserção e remoção de um título ao catálogo. Além disso, foi utilizada para a coleta dos dados do arquivo csv para um arquivo binário.
- **heapq:** foi utilizada para auxiliar na implementação da função de ordenação dos dados (heapsort).
- **pickle:** esta biblioteca foi usada para serializar os dados contidos em um arquivo binário. Portanto, os dados eram armazenados e carregados usando essa função que servia como uma interface entre arquivo binário e programa, facilitando a extração e simplificando processos que poderiam ser feitos manualmente, porém trariam um nível de complexidade e de tempo maior frente ao objetivo do trabalho.
- **csv:** foi utilizado para facilitar a identificação e a conversão dos dados csv para arquivos binários, cria uma interface arquivo-programa que facilita a aplicação das funções e dos módulos responsáveis pela filtragem de dados.
- **os:** usado para a exclusão dos arquivos binários, quando novos arquivos são criados ao executar as funções de adicionar ou remover obra ao catálogo.

COMO COMPILAR O CÓDIGO:

Após baixar o arquivo .zip ou .rar enviado, descompactar e abrir a pasta que se encontra dentro do arquivo como projeto de alguma IDE configurada para Python3. Após isso, basta executar o módulo main.py. Para percorrer as funcionalidades da aplicação basta ir escolhendo as opções desejadas dos menus, explicadas em cada um deles.

CONSIDERAÇÕES FINAIS:

Com a realização deste trabalho, o grupo além de desenvolver e praticar os conteúdos estudados em aula, aprimorou os conhecimentos em uma nova linguagem de programação (Python), visto que nenhuma disciplina anterior em nosso curso havia nos dado acesso a essa linguagem. A realização desse trabalho colaborou muito para concatenar os conhecimentos que pareciam soltos e dar razão e ordem a cada conteúdo apresentado, mostrando a importância de entender conceitos básicos e como as coisas funcionam “debaixo dos panos” de grandes aplicações usadas diariamente para controlar dados.

A parte que encontramos mais dificuldades foi a mais inicial do trabalho, em especial, a conversão de csv para binário e a construção das árvores trie. Inicialmente, levou demasiado tempo para evoluirmos pequenas coisas no trabalho, o que foi de certo

modo um pouco desgastante, mas com o passar dos dias, obtivemos prática e o entendimento do trabalho aumentou, tornando a execução e evolução do projeto mais simples e eficiente.

Se tivéssemos mais tempo, gostaríamos de aprimorar nossas funções de busca, incrementando o número de filtros, bem como, encadeando-os, possibilitando ao usuário buscar por mais de um filtro ao mesmo tempo. Também achamos que seria interessante implementar uma árvore B para pesquisas por alguns filtros específicos, visando facilitar o acesso ao arquivo principal, o que tornaria a filtragem um processo mais rápido e eficiente. Além disso, era de nosso interesse corrigir alguns bugs específicos, como o que obtivemos ao inserir um novo ano de estreia, sendo o único dado que não aceitou ser inserido com valor diferente dos já existentes no arquivo, dentre outras limitações e aprimoramentos de código que poderiam ter sido implementados.

Por fim, conclui-se que a realização do presente trabalho foi bastante proveitosa, pois nos possibilitou aplicar os conceitos aprendidos em aula, sendo que, após a execução do trabalho esses conceitos tornaram-se menos abstratos e o entendimento e motivos de usar cada método ficaram muito mais evidentes. Ademais, foi ótimo desenvolver técnicas novas e aprender uma nova linguagem de programação.