Sistema de recomendação de Filmes utilizando Filtragem

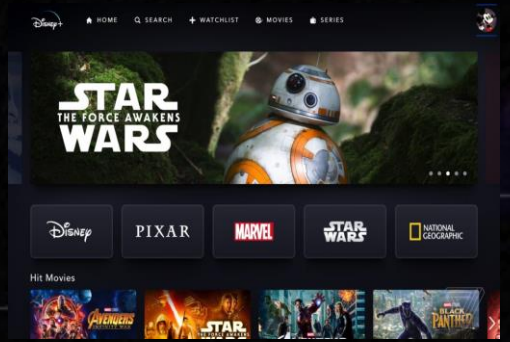
Nome: Guilherme Souza Mingroni
RA: 201027895

Sumário

01 Introdução

Introdução

- Importância dos sistemas de recomendação em plataformas de streaming
- Relevância no consumo de mídia atual.
- FilmMatch



02 Objetivos

Objetivos

Objetivo Geral: Desenvolver e avaliar um sistema de recomendação eficaz

Objetivos específicos:

- **Realizar uma revisão da literatura sobre técnicas e algoritmos de sistemas de recomendação de filmes, destacando as abordagens mais relevantes e eficazes.**
- **Coletar e pré-processar dados de avaliações de filmes e informações de usuários para construir um conjunto de dados representativo**
- **Desenvolver e implementar um sistema de recomendação de filmes utilizando técnicas de filtragem colaborativa e baseada em conteúdo, bem como técnicas de aprendizado de máquina.**
- **Avaliar o sistema utilizando métricas de similaridade como Similaridade de Cosseno, Correlação de Pearson , Índice de Jaccard, além de técnicas de clusterização**
- **Comparar o desempenho do sistema com outros sistemas de recomendação existentes no cenário.**

03 Justificativa

Justificativa

- Segundo dados da MPA, o número de assinantes de serviços de streaming de vídeo globalmente ultrapassou 1 bilhão em 2020

- Segundo dados da Streaming Global do FinderBrasil é o segundo país que mais assistem a streaming

- Segundo um estudo feito na Universidade de Washington, no aplicativo tik tok de 30 a 50% dos vídeos que os usuários veem são recomendados com base em seu engajamento anterior.

04 Pergunta de Pesquisa

Pergunta de Pesquisa

É possível oferecer sugestões mais diversificadas e otimizadas aos usuários através da utilização de diferentes métricas de similaridade no processo de filtragem colaborativa?”

- Ideia:

Avaliar se a utilização de diferentes métricas e técnicas de clusterização gera recomendações mais precisas e diversificadas em relação aos sistemas tradicionais

05 Trabalhos Relacionados

Trabalhos Relacionados

Nome do Autor	Ano de Publicação	Tipo de Filtragem (Exemplo de Plataforma)	Resumo da Obra
Uluyagmur	2018	Filtragem baseada em conteúdo (Rotten Tomatoes)	Sistema de recomendação utilizando características dos filmes (atores, diretores, gêneros) para personalizar recomendações ao perfil do usuário.
Sarwar et al.	2001	Filtragem colaborativa - cold start (Letterboxd)	Exploraram as limitações da filtragem colaborativa, especialmente o problema de 'cold start' para novos usuários ou itens sem histórico.
Salloum e Rajamanthri	2021	Filtragem colaborativa baseada em usuários	Abordagem de filtragem colaborativa baseada em usuários, identificando perfis semelhantes para recomendações mais precisas.
Dwivedi e Islam	2023	Filtragem colaborativa baseada em itens (Plataformas de recomendações sociais)	Abordagem de filtragem colaborativa baseada em itens, recomendando com base na similaridade entre itens previamente avaliados positivamente.

Trabalhos Relacionados

Nome do Autor	Ano de Publicação	Tipo de Filtragem (Exemplo de Plataforma)	Resumo da Obra
Koren	2006	Filtragem híbrida (Netflix, Amazon Prime Video)	Introdução de métodos híbridos avançados, como redes neurais e análise em larga escala para equilibrar personalização e descoberta.
Alsekait	2021	Filtragem híbrida com redes neurais profundas	Sistema híbrido com redes neurais profundas para capturar padrões complexos nas preferências dos usuários.
Setiawan e Arsytania	2024	Filtragem híbrida em cascata com redes neurais convolucionais	Sistema híbrido em cascata combinando redes neurais convolucionais e técnicas colaborativas para recomendações diversificadas.

06 Fundamentação Teórica

Fundamentação Teórica

Evolução dos sistemas de recomendação

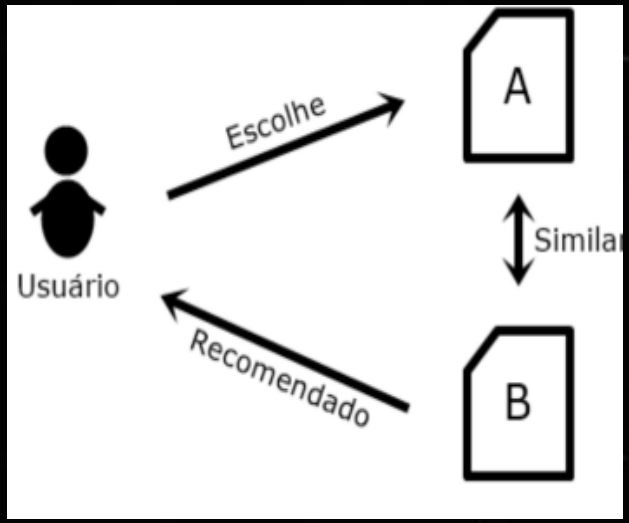
1. Filtragem Baseada em conteúdo -> Analisa características dos filmes para recomendações.
2. Filtragem Colaborativa -> Utiliza padrões de comportamento dos usuários.
3. Sistemas híbridos -> Combina abordagens para recomendações mais precisas.

Fundamentação Teórica

Filtragem Baseada em Conteúdo

1. Examina características específicas dos filmes.
2. Compara com preferências do usuário
3. Recomenda outro filme baseado nessa comparação e exame

- Limitação: Pode criar "bolha de filtragem"



Fonte: Adaptada de (Rolim et al. 2017)

Fundamentação Teórica

Filtragem Colaborativa

Filtragem que analisa padrões de comportamento e preferências de vários usuários para recomendar conteúdos populares entre usuários com gostos semelhantes.

Tipos:

1. Baseada em Usuários -> similaridade entre usuários
2. Baseada em Itens -> similaridade entre filmes

- Limitação: "cold start”

Fundamentação Teórica

Filtragem Hibrída

Combinação dos dois métodos anteriores (baseada em contéudo e colaborativa)

- Recomendações mais diversificadas e precisas.
- Desafios: Complexidade computacional e balanceamento cuidadoso.

Fundamentação Teórica

Métricas de Similaridade

Utilização de diferentes métricas para criação do projeto

- Similaridade cosseno -> mede o ângulo entre os vetores de características

Correlação de Pearson -> mede a relação linear entre duas variáveis.

- Indice de Jaccard -> mede intersecção entre conjuntos de características.

Clusterização também é utilizado nesse modelo (outro tipo de avaliação)

Fundamentação Teórica

Similaridade Cosseno

Mede o ângulo entre os vetores de características

$$\text{similarity}(A, B) = \cos(A, B) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

- A · B representa o produto escalar dos vetores A e B, que é a soma dos produtos dos valores correspondentes em cada vetor.
- ||A|| e ||B|| representam os comprimentos dos vetores A e B, calculadas como a raiz quadrada da soma dos quadrados de seus elementos.
- O resultado da similaridade cosseno varia entre -1 e 1

Fundamentação Teórica

Correlação de Pearson

Mede a relação linear entre duas variáveis.

$$P = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \cdot \sqrt{\sum(y_i - \bar{y})^2}}$$

- xi e yi representam os valores individuais das variáveis X e Y .
- x^{-} e y^{-} são as médias de X e Y , respectivamente.
- $\sum((xi - x^{-})(yi - y^{-}))$ calcula o somatório do produto das diferenças entre cada valor e a média das variáveis.
- $\sqrt{\sum}(xi - x^{-})$ 2 e $\sqrt{\sum}(yi - y^{-})$ 2 calculam as raízes quadradas do somatório das diferenças ao quadrado de cada valor e a média, correspondendo às variâncias das variáveis. •
- Essa fórmula resulta em um valor entre -1 e 1, onde 1 indica uma correlação positiva perfeita, -1 uma correlação negativa perfeita e 0 indica que não há correlação linear entre as variáveis.

Fundamentação Teórica

Indice de Jaccard

Mede a intersecção entre dois conjuntos dividida pela união desses conjuntos

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

- $|A \cap B|$ representa o número de elementos em comum entre os conjuntos A e B (interseção).
- $|A \cup B|$ representa o número total de elementos na união dos conjuntos A e B.
- O índice de Jaccard varia entre 0 e 1, onde 1 indica que os conjuntos são idênticos (máxima similaridade), e 0 indica que não possuem elementos em comum.

Fundamentação Teórica

Métrica combinada

Utiliza a intersecção das 3 métricas anteriores para recomendação dos filmes

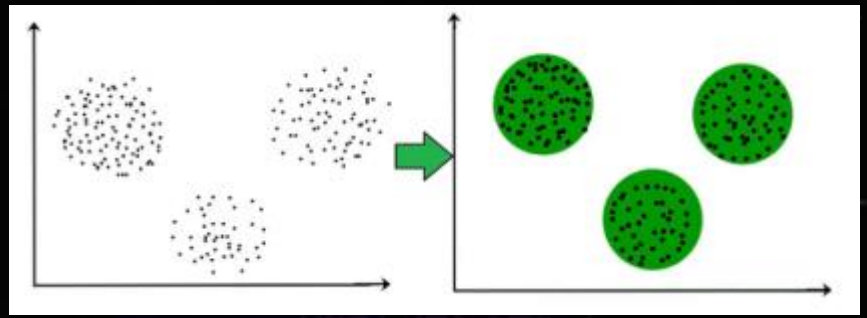
Utiliza uma média ponderada para recomendar os filmes, dando prioridade para gênero e atores, mas também utilizando outras características dos filmes e de usuários

Fundamentação Teórica

Clusterização

Técnica de agrupamento de itens por características semelhantes

- Tipo utilizado no sistema: DBSCAN
  - Agrupa pontos em regiões densamente povoadas, enquanto identifica pontos isolados como ruídos
  - Definição dos parâmetros de entrada, que incluem o raio de busca (ε) e o número mínimo de pontos (MinPts) necessários para formar um cluster.



Fundamentação Teórica

Métricas de avaliação

- Utiliza 2 métricas para avaliação:

1. Precisão -> mede a proporção de recomendações relevantes entre todas as recomendações feitas.

$$\text{Precisão} = \frac{TP}{TP + FP}$$

2. Recall -> mede a capacidade do sistema de recomendar filmes relevantes dentre todos os filmes que poderiam ser recomendados

$$\text{Recall} = \frac{TP}{TP + FN}$$

07 Modelo Proposto

Modelo Proposto

Utiliza os três tipos de filtragem para recomendação:

1. Filtragem baseada em conteúdo -> Indice de Jaccard e Clusterização
2. Filtragem Colaborativa -> Correlação de Pearson e Similaridade Cosseno
3. Filtragem Hibrida -> Similaridade Total

Similaridade Total

- Média ponderada:
  - **60% do índice de jaccard**
  - **20% de Pearson**
  - **20% de Similaridade Cosseno**

Objetivo: Dar ênfase no gênero e atores, diante dos demais itens que também são importantes para recomendação

Modelo Proposto

Funcionamento do sistema:

1 -> Usuário escolhe o filme

2 -> Usuário escolhe a metodologia

3 -> Usuário clica em calcular e abaixo do botão clicado, aparece os filmes recomendados

Selecione um filme da lista abaixo:

Filme: I Am Mother

Diretor: Grant Sputore

Ano de Lançamento: 2019

Duração: 114

Gênero: Drama, Mystery, Sci-Fi

Descrição:
Following humanity's mass extinction, a teen raised alone by a maternal droid finds her entire world shaken when she encounters another human.

Escolha a metodologia:

Método: Pearson

Calcular

08 Metodologia

Metodologia

Para criação do software foi utilizado 2 ferramentas

1. Google Colab
2. Python

Dentro de Python foram utilizadas algumas bibliotecas como:

- Pandas -> Juntar e transformar de datasets
- Ipywidgets -> Criar de menus interativos e botões
- Numpy -> Calcular Operações numéricas e cálculos vetorizados
- Scikit-Learn -> Implementar Métricas como Similaridade Cosseno e Pearson além de auxiliar no modelo de Clusterização em DBSCAN

Metodologia

Datasets utilizados:

- **Netflix Movies and TV Shows**

Fornece informações detalhadas sobre o catálogo da Netflix, incluindo título, diretor, ano de lançamento, duração, elenco e descrição. Essencial para análises de similaridade entre títulos.

- **IMDB Movies Dataset**

Oferece dados como pôster do filme, título, gênero, classificação e número de votos.
Utilizado para enriquecer a interface visual e refinar a filtragem colaborativa.

- **TMDB Movies**

Base extensa com mais de 900 mil títulos, incluindo avaliações detalhadas dos usuários.
Fornece perspectiva adicional para a filtragem colaborativa e recomendações variadas

Metodologia

Processamento dos dados:

1. **Importação dos Bases**
   Utilização do Pandas para importar datasets, com parâmetros específicos para lidar com problemas de leitura de dados.

2. **Seleção e Renomeação**
   Seleção de variáveis relevantes e padronização dos nomes das colunas para consistência entre datasets.

3. **União dos Dados**
   Junção dos datasets com base na coluna de título, usando inner join para manter apenas filmes presentes em todas as fontes.

4. **Limpeza e Transformação**
   Remoção de valores nulos, transformação de variáveis categóricas em colunas binárias (One-Hot Encoding) e conversão de tipos de dados.

5. **Implementação de Similaridade**
   Desenvolvimento de funções para calcular similaridade entre filmes usando diversos métodos, como Índice de Jaccard e Similaridade Cosseno.

09 Implementações das funções de similaridade

Implementações das funções de similaridade

- Indice de Jaccard -> índice_jaccard
- Similaridade Cosseno -> cossine_similarity

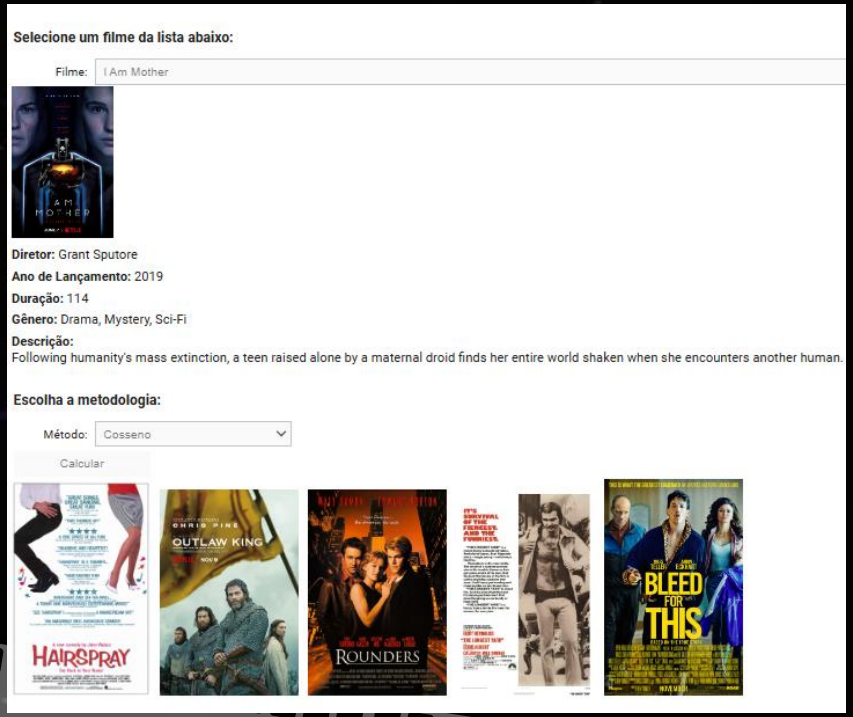
Essa função usa sklearn para padronizar colunas numéricas relevantes e depois calcula. Esse resultado é gerado e fica dentro de uma lista ordenada por conta da função calcular_cosseno

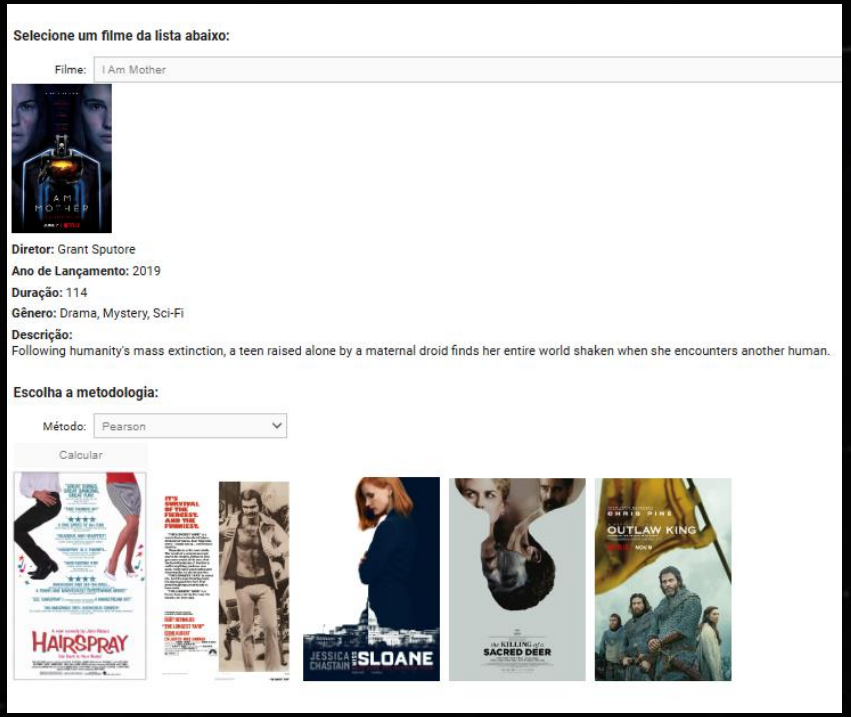
- Correlação de Pearson
- Clusterização com DBSCAN
- Similaridade Total

10 Resultados

Resultados

Parte gráfica -> Exemplo de comparação de recomendações usando Cosseno e Pearson





Resultados

Tempo de execução

Métrica	Complexidade Computacional	Tempo de Execução Estimado (para 1000 filmes)	Tempo de Execução Real	Observações
Índice de Jaccard	$O(n^2)$	2 a 5 segundos	2.3 segundos	Confirma a estimativa inicial de 2 a 5 segundos
Similaridade Cosseno	$O(n)$	0.5 a 1 segundo	0.7 segundos	Execução rápida e dentro da estimativa
Correlação de Pearson	$O(n)$	1 a 2 segundos	1 segundo	Aproxima-se da estimativa
Clusterização (DBSCAN)	$O(n \log n)$ a $O(n^2)$	1 a 3 segundos	2 segundos	Dentro do intervalo previsto de 1 a 3 segundos
Similaridade Total	$O(n^2)$	3 a 6 segundos	3 segundos	Confirmou o tempo estimado de 3 a 6 segundos

Resultados

Precisão

Sistema	Precisão
Nosso Sistema	0.81
ULUYAGMUR et al. (2012)	0.213
Dwivedi & Islam (2023)	0.85
Alsekait et al. (2024)	0.88
Setiawan & Arsytania (2024)	0.8695

- Indice que a maioria das recomendações feitas pelo sistema foi pertinente.
- É importante notar que sistemas híbridos de recomendação geralmente alcançam precisões que variam entre 0.7 e 0.85, como observado por (Jannach e Adomavicius 2016).

Resultados

Recall

Sistema	Recall
Nosso Sistema	0.07
ULUYAGMUR et al. (2012)	0.095
Alsekait et al. (2024)	0.85

- Indice que a maioria das recomendações que poderiam ser de gosto do usuário não foram recomendadas

11 Conclusão

Conclusão

Métricas de avaliação

- Resultado dentro do esperado na precisão (0.81 de precisão)
- Resultado negativo/ de menor expressão pensando no recall (0.07 de recall)

Há a necessidade de busca de equilíbrio entre precisão e recall

12 Perspectivas Futuras

Perspectivas futuras

Utilização de mais artifícios para melhoria do sistema:

- Utilização de novos algoritmos, interligação de mais métricas e varíaveis para aumento do recall
- Utilização de outros métodos de avaliação
- Utilização de plataformas que coloquem essa aplicação em uma rede (através de um site ou aplicativo)
- Utilização de API´s que detectem um novo adicionamento de filme e já coloque nesse banco de dados com padronização dos dados
- Detalhamento de atualizações (caso for um site ou aplicativo)

Referências:

- ALSEKAIT, D. M.; SHDEFAT, A. Y.; MOSTAFA, N. Next-generation movie recommenders: Leveraging hybrid deep learning for enhanced personalization. Natural Publishing, 2024. Disponível em: https://www.naturalspublishing.com/files/published/3nw1w641kciz46. pdf.

- DWIVEDI, P.; ISLAM, B. An item-based collaborative filtering approach for movie recommendation system. In: Proceedings of the 10th International Conference on Computing, Communication and Automation (ICCCA). IEEE, 2023. Disponível em: https://ieeexplore.ieee.org/abstract/document/10112338/.

- GOMES, P. C. T. Clustering: O que é Cluster Analysis e quais suas Aplicações? 2024. Acessado em: 22 de Outubro de 2024. Disponível em: https://analisemacro.com.br/ data-science/clustering-o-que-e-cluster-analysis-e-quais-suas-aplicacoes/.

- JACCARD, P. Étude comparative de la distribution florale dans une portion des alpes et du jura. Bulletin de la Société Vaudoise des Sciences Naturelles, v. 37, p. 547–579, 1901. JANNACH, D.; ADOMAVICIUS, G. Recommender Systems: Past, Present, and Future. [S.l.]: Springer, 2016.

- KOREN, Y.; BELL, R.; VOLINSKY, C. Matrix factorization techniques for recommender systems. Computer, v. 42, n. 8, p. 30–37, 2009. Laboratório de Aplicações de Machine Learning em Finanças e Organizações. Sistemas de Recomendação usando Collaborative Filtering. 2018. Acessado em: 20 de Outubro de 2024. Disponível em: https://lamfo-unb.github.io/2018/09/29/ Sistemas-de-Recomenda%C3%A7%C3%A3o-usando-Collaborative-Filtering/.

- MONTEIRO, G.; CARL, H. Entendendo DBSCAN. 2020. Acessado em: 1 de Novembro de 2024. Disponível em: https://medium.com/@gabrielmonteiro/ entendendo-dbscan-por-gabriel-monteiro-e-hugo-carl-1234567890.

Referências:

• NGUYEN, T. T.; HUI, P. M.; HARPER, F. M.; TERVEEN, L.; KONSTAN, J. A. Exploring the filter bubble: The effect of using recommender systems on content diversity. In: Proceedings of the 23rd International Conference on World Wide Web. Seoul: ACM, 2014. p. 677–686. PEARSON, K. Note on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London, v. 58, p. 240–242, 1895.

- ROLIM, V. B.; FERREIRA, R.; COSTA, E.; CAVALCANTI, A. P.; FERREIRA, M. A. D. Um estudo sobre sistemas de recomendação de recursos educacionais. ResearchGate, October 2017. Disponível em: https://www.researchgate.net/publication/Um_Estudo_ Sobre_Sistemas_de_Recomendacao_de_Recursos_Educacionais.

- RUSSELL, S.; NORVIG, P. Artificial Intelligence: A Modern Approach. 3rd. ed. [S.l.]: Pearson, 2016. 43 SALLOUM, S.; RAJAMANTHRI, D. Implementation and evaluation of movie recommender systems using collaborative filtering. Journal of Advances in Information Technology, v. 12, n. 3, p. 189–196, 2021. Disponível em: https://www.jait.us/uploadfile/2021/0719/ 20210719052408995.pdf.

- SARWAR, B.; KARYPIS, G.; KONSTAN, J.; RIEDL, J. Item-based collaborative filtering recommendation algorithms. In: ACM. Proceedings of the 10th International Conference on World Wide Web. [S.l.], 2001. p. 285–295.

- SETIAWAN, E. B.; ARSYTANIA, I. H. Movie recommender system with cascade hybrid filtering using convolutional neural network. UAD, 2024. Disponível em: https://eprints. uad.ac.id/65078/1/1-Movie%20Recommender%20System%20with%20Cascade% 20Hybrid%20Filtering%20Using%20Convolutional%20Neural%20Network.pdf

Referências:

- THOMPSON, K.; BORDWELL, D. Film History: An Introduction. [S.l.]: McGraw-Hill Education, 1994.
- TURING, A. M. Computing machinery and intelligence. Mind, v. 59, n. 236, p. 433–460, 1950.
- ULUYAGMUR, M.; CATALTEPE, Z.; TAYFUR, E. Content-based movie recommendation using different feature sets. In: Proceedings of the World Congress on Engineering and Computer Science. [S.l.: s.n.], 2012. p. 17–24