UNIVERSIDADE PRESBITERIANA MACKENZIE
FACULDADE DE COMPUTAÇÃO E INFORMÁTICA
TECNOLOGIA EM CIÊNCIAS DE DADOS

PROJETO APLICADO III

HistFlix: Um sistema de recomendação personalizado de filmes históricos e documentários

PROFESSORA: CAROLINA TOLEDO FERRAZ

GRUPO:

BRUNO BALTUILHE – 10424822 – 10424822@mackenzista.com.br
ISAQUE PIMENTEL – 10415608 – 10415608@mackenzista.com.br
KELLY GRAZIELY PENA – 110416108 – 10416108@mackenzista.com.br

São Paulo
2025

RESUMO

Redigir ao final do trabalho.

SUMÁRIO

1. INTRODUÇÃO

O cinema tem um papel fundamental na sociedade, não apenas como entretenimento, mas também como meio de transmitir conhecimento de maneira acessível e interessante. Documentários, séries e filmes históricos proporcionam uma visão única sobre eventos do passado, permitindo a os espectadores absorver informações valiosas enquanto são imersos em uma narrativa envolvente.

O consumo de filmes, séries e documentários tem se tornado cada vez mais personalizado, impulsionado pelo avanço da tecnologia e pelos sistemas de recomendação. Essa personalização melhora a experiência do usuário e amplia o acesso a conteúdos relevantes que poderiam passar despercebidos.

O objetivo deste projeto [1]é **desenvolver um sistema de recomendação que auxilie os usuários a encontrarem filmes e documentários de alta relevância cultural e educativa, considerando não apenas preferências passadas, mas também fatores contextuais como estado emocional**, ampliando o acesso ao aprendizado histórico por meio da sétima arte e proporcionando recomendações mais adequadas e personalizadas.

1.1. CONTEXTO DO TRABALHO

Somos um grupo de alunos de Ciências de Dados desenvolvendo um projeto de Sistema de Recomendação para melhorar as técnicas aprendizagem de disciplinas escolares além da sala de aula. Propomos o **HistFlix**, um sistema de recomendação de filmes e documentários de qualidade e relevância histórica e educacional, para estender o aprendizado da História além da sala de aula. Após o desenvolvimento do nosso produto, o apresentaremos para avaliação da disciplina de Projeto Aplicado III da Universidade Mackenzie.

[1] Segue a URL do projeto no Github: `https://github.com/isaque-pimentel/projeto-aplicado-3`

1.2. MOTIVAÇÃO E JUSTIFICATIVAS

A escolha do tema é impulsionada pelo crescente interesse em métodos de ensino alternativos que possam complementar os modelos tradicionais de educação. Filmes, séries e documentários despertam interesse em diferentes temas, como cultura, história e ciência, através de uma abordagem lúdica e visual. Além disso, o avanço da inteligência artificial permite que sistemas de recomendação personalizem essas experiências, sugerindo conteúdos alinhados às preferências dos usuários e suas necessidades de aprendizado.

Este projeto também busca aprimorar os sistemas de recomendação existentes, tornando-os mais dinâmicos e sensíveis ao contexto do usuário. Ao incorporar inteligência artificial para interpretar emoções expressadas por meio de uma interação textual, o sistema poderá fornecer sugestões mais alinhadas às necessidades específicas de cada momento. Isso contribui para um consumo mais significativo e envolvente de conteúdos audiovisuais, reduzindo a frustração causada por recomendações irrelevantes.

A HistFlix busca suprir a necessidade de um sistema especializado que forneça recomendações precisas e relevantes para estudantes, pesquisadores e entusiastas da história. Além disso, o projeto alinha-se aos Objetivos de Desenvolvimento Sustentável (ODS) da ONU, promovendo educação de qualidade ao facilitar o acesso a conteúdos educativos.

1.3. OBJETIVOS

Objetivo Geral

Desenvolver um sistema de recomendação de filmes, séries e documentários que utilize um modelo híbrido de recomendação e inteligência artificial para interpretar as emoções do usuário e sugerir conteúdos alinhados ao seu estado emocional e preferências específicas, com o propósito de aumentar o interesse pela história e democratizar o acesso a conteúdos audiovisuais educativos.

Objetivos Específicos

- Coletar e processar dados sobre filmes, séries e documentários, utilizando a base de dados **MovieLens**, para criar um modelo de recomendação personalizado.
- Implementar um modelo híbrido de recomendação, combinando filtragem colaborativa (que analisa o comportamento e as avaliações de outros usuários com perfis semelhantes) e filtragem baseada em conteúdo (que considera características específicas das obras audiovisuais, como gênero, duração, elenco e temática).
- Desenvolver uma interface interativa na qual os usuários possam expressar suas emoções e preferências momentâneas.
- Integrar técnicas de Processamento de Linguagem Natural (PLN) para interpretar sentimentos e preferências expressas textualmente.

1.4. BASE DE DADOS

Utilizaremos a base de dados **MovieLens 1M**, amplamente utilizada para pesquisas em sistemas de recomendação. Esta base contém aproximadamente 1 milhão de avaliações de filmes realizadas por usuários anônimos, permitindo a implementação e a validação de algoritmos de recomendação.

Detalhamentos da Base de Dados

- **Fonte:** MovieLens 1M Dataset do GroupLens Research.
- **Forma e período de coleta:** Os dados foram coletados a partir interações reais de usuários com a plataforma durante o período de 2003 e 2004, registrando avaliações numéricas em uma escala de 1 a 5.
- **Estrutura dos Dados:** Avaliações de mais de 6.000 usuários anônimos sobre aproximadamente 4.000 filmes, totalizando cerca de 1.000.000 de avaliações.
- **Limitações:** A diversidade de gêneros é limitada aos filmes listados na base, podendo afetar a variedade de recomendações. Além disso, a base de dados é antiga, e muitos filmes recentes não estão incluídos, reduzindo inicialmente a relevância de algumas recomendações.

1.5. PIPELINE IDEALIZADO

O pipeline a ser adotado é o seguinte:

1. **Coleta de dados**: Importação das avaliações e metadados dos filmes a partir de MovieLens 1M.
2. **Processamento e Armazenamento**: Organização dos dados em um banco estruturado para facilitar buscas e análises.
3. **Interação com o Usuário**: Implementação de um chatbot baseado em PLN para interpretar emoções e preferências momentâneas.
4. **Geração de Recomendações**: Uso do modelo híbrido para sugerir conteúdos com base no histórico e no estado emocional do usuário.
5. **Aprimoramento Contínuo**: Análises estatísticas para otimizar a precisão do sistema ao longo do tempo.

Este projeto busca oferecer recomendações eficientes e criar uma experiência personalizada, permitindo que os usuários explorarem conteúdos alinhados também ao seu estado emocional, enriquecendo a experiência de aprendizado sobre História.

1.6 CRONOGRAMA DO PROJETO

O cronograma proposto para o projeto do Sistema de Recomendação **HistFlix** é o seguinte:

Etapa	Atividade	Prazo	Descrição
Etapa 1	Concepção do Produto	Semana 1-2	Identificação de necessidades da comunidade e levantamento de filmes relevantes para educação histórica.
Etapa 2	Definição do Produto	Semana 3-4	Validar critérios de recomendação e ampliar a abrangência do sistema.
Etapa 3	Metodologia e implementação do Modelo e Testes	Semana 5-6	Desenvolvimento do sistema com feedback de usuários.
Etapa 4	Resultado e conclusão	Semana 7-8	Apresentação do projeto.

2. COLETA E PROCESSAMENTO DOS DADOS

2.1. ETAPA DE EXTRAÇÃO E PROCESSAMENTO DOS DADOS

A etapa de extração e processamento dos dados consiste em transformar os arquivos originais do **MovieLens 1M** em tabelas estruturadas dentro de um banco de dados relacional SQLite. Essa transformação é essencial para facilitar a manipulação, consulta e análise dos dados durante o desenvolvimento do sistema.

O processamento e o armazenamento dos dados em um banco de dados relacional (e.g. SQLite) foi realizado pelo script *data_extraction.py*. Abaixo temos os passos dessa tarefa:

- Leitura dos arquivos *.dat*:
  - Os arquivos *users.dat*, *ratings.dat* e *movies.dat* foram carregados usando a biblioteca **pandas**;
  - Cada arquivo foi carregado em um DataFrame, com as colunas devidamente nomeadas de acordo com a documentação do MovieLens 1M;
- Limpeza e normalização dos dados:
  - Remoção de valores ausentes ou inconsistentes;
  - Conversão de tipos de dados para formatos mais eficientes (e.g., int32, float32);
  - Normalização de colunas, como transformar o gênero em valores numéricos (e.g., 0 para "F" e 1 para "M");
  - Extração de informações adicionais, como o ano de lançamento dos filmes a partir do título;
- Criação do banco de dados relacional SQLite:
  - Um banco de dados SQLite foi criado usando a biblioteca **sqlite3**;
  - As tabelas *users*, *ratings* e *movies* foram criadas no banco de dados, e os dados limpos foram inseridos diretamente a partir dos pandas DataFrames;
  - Os dados foram armazenados no banco de dados SQLite, permitindo consultas SQL eficientes e integração com outras ferramentas de análise;

2.2. BENEFÍCIOS DO PROCESSAMENTO

Existem inúmeros benefícios do uso de uma base de dados relacional sobre arquivos originalmente em formato de texto bruto (no caso no formato *.dat*), dentre elas se destacam:

- Eficiência: Uma estrutura relacional como a do SQLite permite consultas rápidas e organizadas;
- Portabilidade: Esse banco de dados SQLite é mais leve e pode ser facilmente compartilhado (ver arquivo em dataset\sqlite\movielens_1m.db);
- Manipulação facilitada: O uso do SQL simplifica a extração de informações específicas para análises ou treinamento de modelos. É comum o conselho no ambiente de ciência de dados: “Domine as consultas SQL antes de avançar no seu conhecimento”.

Essa etapa garante que os dados estejam prontos para serem explorados e utilizados nas etapas subsequentes do projeto, como a análise exploratória e a construção do sistema de recomendação.

3. ANÁLISE EXPLORATÓRIA DE DADOS (EDA)

A Análise Exploratória de Dados (EDA) foi conduzida com o objetivo de compreender as principais características do conjunto de dados da plataforma **HistFlix**, utilizando informações extraídas de um banco de dados SQLite. O processo envolveu a inspeção de estrutura dos dados, verificação de valores ausentes, distribuição de avaliações, análise de usuários e gêneros cinematográficos, bem como a evolução das avaliações ao longo do tempo.

Estrutura e Qualidade dos Dados

Os dados foram extraídos de um banco SQLite, contendo três principais tabelas: **users**, **ratings** e **movies**. A Tabela 1 apresenta a dimensão de cada conjunto de dados:

Tabela	Linhas	Colunas
users	6.040	4
ratings	1.000.209	4
movies	3.883	22

Tabela 1 Dimensão das tabelas do banco SQLite

A verificação de valores ausentes indicou que não há dados faltantes em nenhuma das tabelas, garantindo a integridade da base de dados.

Distribuição das Avaliações

A distribuição das avaliações reflete uma tendência dos usuários a atribuir notas mais altas, com 34,89% das avaliações correspondendo a 4 estrelas, seguidas por 26,11% para 3 estrelas e 22,63% para 5 estrelas. A Figura 1 ilustra essa distribuição, destacando a predominância de notas positivas no base MovieLens 1M.

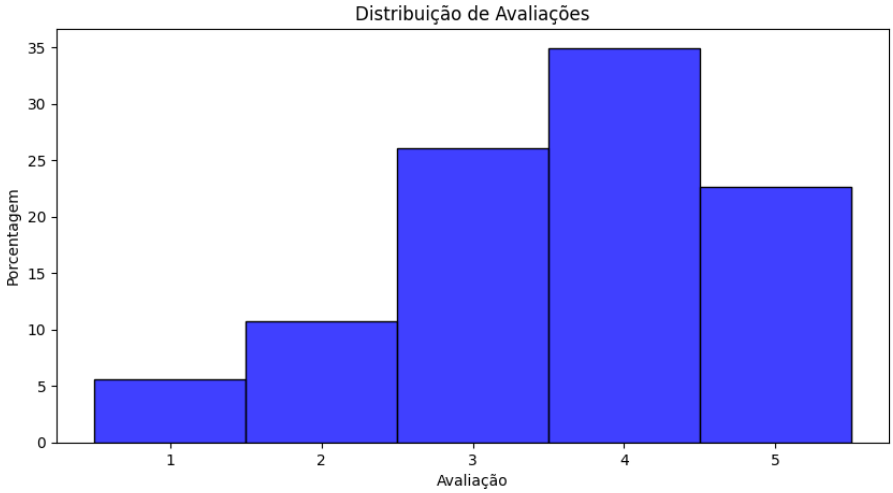

Figura 1 Distribuição de Avalições

Perfil dos Usuários

A distribuição de gênero entre os usuários da plataforma indica um predomínio do gênero masculino, representando 71,71% da base, enquanto o gênero feminino corresponde a 28,29%. A distribuição etária apresenta concentração de usuários na faixa de 25 anos (34,70%), seguida por 35 anos (19,75%) e 18 anos (18,26%), conforme representado na Figura 2. Essas características impactam diretamente o sistema de recomendação, pois diferentes grupos podem ter preferências distintas de filmes, exigindo abordagens personalizadas para melhorar a experiência do usuário.

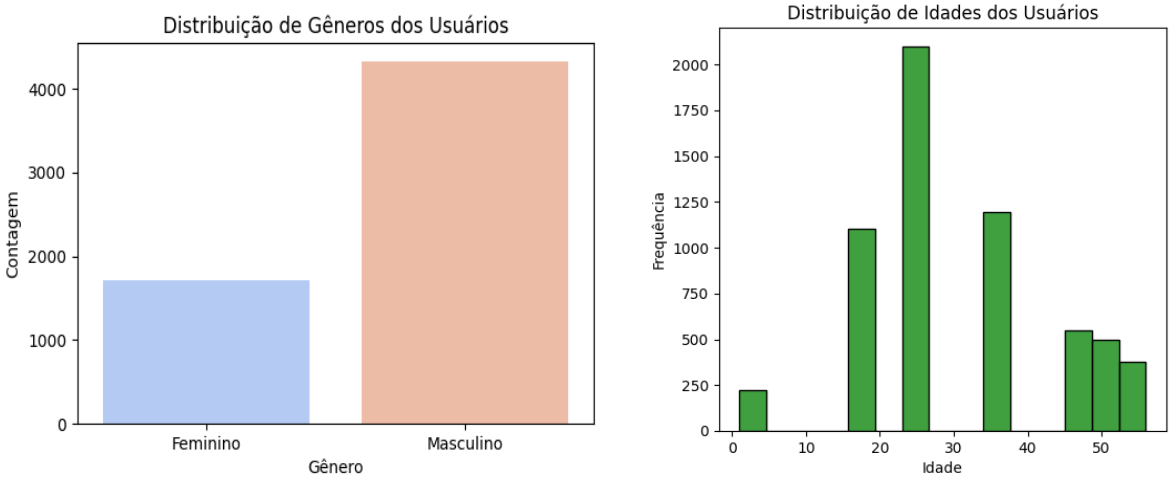

Figura 2 Distribuição dos Usuários por Gênero e Idade

Filmes Mais Bem Avaliados

Os filmes mais avaliados refletem títulos amplamente reconhecidos pelo público. O filme **"American Beauty" (1999)** lidera com 3.428 avaliações, seguido por clássicos da franquia *Star Wars* e outros sucessos como *Jurassic Park* (1993) e *Matrix* (1999). A Tabela 2 apresenta os dez filmes mais avaliados na plataforma. Esse aspecto é fundamental para o sistema de recomendação, pois filmes com muitas avaliações podem influenciar a personalização das sugestões, enquanto títulos menos populares podem precisar de técnicas avançadas para serem descobertos pelos usuários.

Posição	Filme	Número de Avaliações
1º	*American Beauty (1999)*	3.428
2º	*Star Wars: Episode IV - A New Hope (1977)*	2.991
3º	*Star Wars: Episode V - The Empire Strikes Back (1980)*	2.990
4º	*Star Wars: Episode VI - Return of the Jedi (1983)*	2.883
5º	*Jurassic Park (1993)*	2.672
6º	*Saving Private Ryan (1998)*	2.653
7º	*Terminator 2: Judgment Day (1991)*	2.649
8º	*The Matrix (1999)*	2.590
9º	*Back to the Future (1985)*	2.583
10º	*The Silence of the Lambs (1991)*	2.578

Tabela 2 Top 10 filmes mais avaliados

Média de Avaliação por Gênero Cinematográfico

A análise dos gêneros mostrou que os gêneros **"Film-Noir"**, **"Documentary"** e **"War"** possuem as médias de avaliação mais altas, superando 3,89 estrelas, enquanto **"Horror"** e **"Children's"** apresentam as menores notas médias. A Figura 3 exibe a distribuição da avaliação média por gênero e sua variação entre os diferentes gêneros.

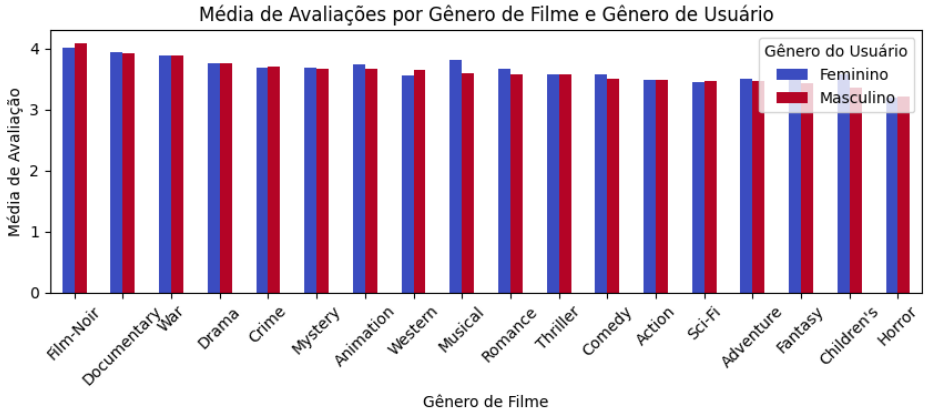

Figura 3 Distribuição da avaliação média por gênero de filme e de usuário

Evolução das Avaliações ao Longo do Tempo

A evolução do volume de avaliações ao longo do tempo indica um aumento expressivo entre os anos de 2000 e 2001, com um pico em novembro de 2000, seguido por um declínio gradual. Esse comportamento pode estar relacionado a eventos de promoção da plataforma ou lançamentos de filmes populares no período. A Figura 4 ilustra essa evolução, segmentada por gênero dos usuários. A variação do volume de avaliações ao longo do tempo pode indicar padrões sazonais ou mudanças de comportamento dos usuários, elementos que podem ser utilizados para aprimorar o sistema de recomendação, tornando-o mais adaptável a tendências e preferências dinâmicas.

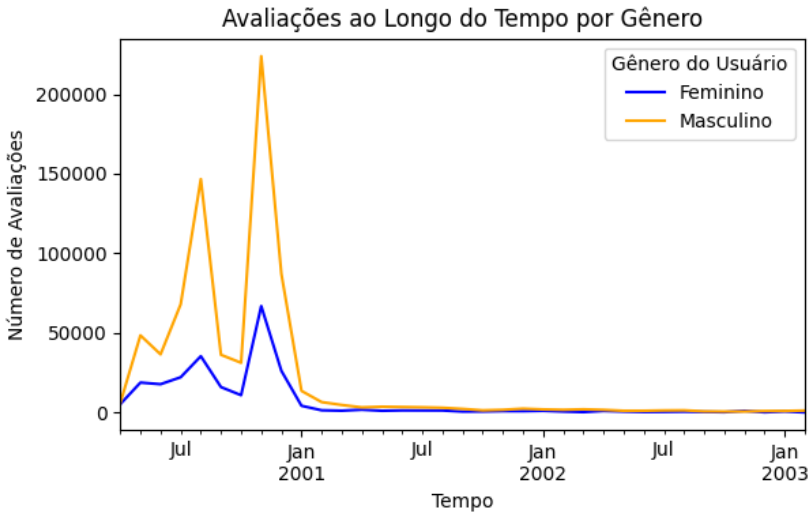

Figura 4 Evolução das avaliações ao longo do tempo

Considerações Finais

Os insights obtidos nesta EDA fornecem subsídios para a construção de um sistema de recomendação mais eficaz. O perfil dos usuários indica que diferentes

faixas etárias e gêneros podem exigir recomendações personalizadas, enquanto a popularidade dos filmes pode influenciar a forma como os títulos são sugeridos. Além disso, a evolução das avaliações ao longo do tempo sugere que um sistema adaptativo pode melhorar a relevância das sugestões conforme as tendências do público mudam.

Para um sistema de recomendação em produção, consideramos a utilização de uma base de dados maior, como o conjunto MovieLens 10M, que fornecerá uma amostragem mais ampla e representativa para otimizar os modelos de recomendação e garantir previsões mais precisas.

A análise exploratória permitiu uma compreensão detalhada da base de dados, identificando padrões relevantes para a modelagem preditiva. A distribuição de notas sugere uma tendência às avaliações positivas, enquanto a segmentação por idade e gênero dos usuários destaca diferenças na interação com o conteúdo. Além disso, a evolução das avaliações ao longo do tempo aponta para dinâmicas que podem influenciar estratégias de recomendação de conteúdo.

4. DEFINIÇÃO DAS TÉCNICAS PARA O TREINAMENTO DO MODELO DE RECOMENDAÇÃO

4.1. MODELO HÍBRIDO

4.1.1. JUSTIFICATIVA DA ABORDAGEM

O HistFlix adotará um modelo híbrido de recomendação, combinando duas abordagens principais:

- **Filtragem Colaborativa (FC)**: baseia-se no comportamento de usuários semelhantes, recomendando itens com base em avaliações anteriores.
- **Filtragem Baseada em Conteúdo (FBC)**: utiliza as características dos itens (ex.: gênero, duração, tema histórico) para gerar recomendações personalizadas.

A escolha por essa abordagem híbrida se justifica pelos seguintes fatores:

- A FC é altamente eficaz quando há um volume significativo de dados de usuários, mas sofre com o problema do cold start, dificultando recomendações para novos usuários ou itens pouco avaliados.
- A FBC mitiga essa limitação ao utilizar metadados dos filmes, sendo particularmente útil para perfis novos ou conteúdos pouco populares.
- O sistema precisa lidar com preferências explícitas (avaliações de filmes) e contextuais (emoções do usuário), tornando a combinação das duas técnicas essencial para recomendações mais relevantes.

4.1.2. IMPLEMENTAÇÃO DO MODELO

A implementação do modelo híbrido seguirá a seguinte abordagem técnica:

- **Filtragem Colaborativa**: utilizará **Singular Value Decomposition** (SVD), uma técnica de fatoração de matrizes amplamente aplicada em sistemas de recomendação.
- **Filtragem Baseada em Conteúdo**: aplicará **TF-IDF** combinado com **similaridade do cosseno** sobre descrições, gêneros e palavras-chave dos filmes.

- **Combinação das abordagens**: a fusão dos modelos será feita por uma média ponderada, ajustável conforme os experimentos e avaliações do sistema.

4.1.3. JUSTIFICATIVA DA COMBINAÇÃO DE SCORES

Para consolidar os resultados das duas abordagens, será empregada a estratégia de **fusão no nível de pontuação** (**score-level fusion**), conforme descrito na literatura sobre sistemas híbridos. A fórmula utilizada será:

$$score\ final = \alpha \cdot scoreFC + (1 - \alpha) \cdot scoreFBC$$

onde:

$scoreFC$: Nota prevista pelo modelo de Filtragem Colaborativa.

$scoreFBC$: Nota prevista pela Filtragem Baseada em Conteúdo.

α: Peso atribuído a cada abordagem (por exemplo: 0.7 para FC e 0.3 para FBC).

Essa combinação permite balancear as vantagens de cada técnica, garantindo um modelo mais robusto e adaptável.

4.1.4. PERSONALIZAÇÃO CONTEXTUAL COM NLP

Além das técnicas tradicionais de recomendação, o **HistFlix** incluirá uma camada de **Processamento de Linguagem Natural (NLP)** para interpretar o estado emocional do usuário. Esse componente permitirá:

- Extração de emoções a partir do texto inserido pelo usuário (ex.: “curioso”, “reflexivo”, “entusiasmado”).
- Filtragem contextual das recomendações, ajustando o tom dos conteúdos sugeridos de acordo com o estado emocional identificado.

Essa adaptação dinâmica visa aumentar a personalização e o engajamento dos usuários no consumo de conteúdos históricos.

4.2. AVALIAÇÃO DE DESEMPENHO

A avaliação do desempenho do **HistFlix** será conduzida utilizando métricas amplamente empregadas (chamadas de **métricas offline**) na literatura de sistemas de recomendação. Para validação offline, serão utilizados dados da base **MovieLens 1M**, permitindo uma análise detalhada da eficácia do modelo híbrido.

A avaliação será realizada sob duas perspectivas:

4.2.1. AVALIAÇÃO POR ERRO DE PREDIÇÃO

Essa abordagem mensura a precisão das notas previstas pelo sistema em relação às avaliações reais dos usuários. A principal métrica utilizada será:

- **RMSE (Root Mean Square Error):** mede o desvio padrão dos erros de predição, sendo uma referência consolidada em competições como o Netflix Prize. Quanto menor o RMSE, maior a precisão das recomendações.

4.2.2. AVALIAÇÃO POR RANKING E RELEVÂNCIA

Como a qualidade da experiência do usuário também depende da ordem em que os itens são recomendados, será avaliada a eficácia do **ranking gerado pelo sistema**. Para isso, serão utilizadas as métricas:

- **Precision@K:** mede a proporção de itens relevantes dentro das K recomendações exibidas.
- **Recall@K:** mede a proporção de itens relevantes recuperados pelo sistema em relação ao total de itens relevantes disponíveis.

Essas métricas são apropriadas quando o sistema recomenda listas de filmes, e o objetivo é garantir que os filmes sugeridos sejam, de fato, relevantes para o usuário — tanto em termos de interesse quanto de emoção, no contexto do HistFlix.

5. REFERENCIAL TEÓRICO

Este referencial teórico fundamenta as escolhas metodológicas e técnicas adotadas no desenvolvimento do **HistFlix**, um sistema de recomendação inteligente voltado à sugestão personalizada de filmes e documentários históricos, sensível ao estado emocional dos usuários. Para isso, são abordadas teorias, modelos e algoritmos amplamente utilizados em sistemas de recomendação, com base em estudos consolidados e pesquisas recentes.

Além disso, esta seção contextualiza a aplicação de técnicas de **Processamento de Linguagem Natural (NLP)** na interpretação subjetiva das interações dos usuários e discute a sinergia entre essas abordagens. Ao reunir conceitos de filtragem colaborativa, filtragem baseada em conteúdo, modelos híbridos e análise de emoções, o referencial teórico embasa a proposta de um sistema robusto, contextual e orientado ao engajamento educacional.

5.1. SISTEMAS DE RECOMENDAÇÃO

Os **sistemas de recomendação** são soluções computacionais projetadas para sugerir itens potencialmente relevantes aos usuários, considerando seu comportamento anterior, preferências declaradas ou contexto atual (**Ricci et al., 2015**). Essas soluções são amplamente utilizadas em diversos domínios, como comércio eletrônico, educação, entretenimento e redes sociais.

De acordo com a literatura, os principais tipos de sistemas de recomendação são:

- **Filtragem Colaborativa (FC)**
- **Filtragem Baseada em Conteúdo (FBC)**
- **Modelos Híbridos**

Essas abordagens podem ser utilizadas isoladamente ou combinadas, dependendo dos objetivos do sistema e das características da base de dados.

5.1.1. FILTRAGEM COLABORATIVA (FC)

A **filtragem colaborativa (FC)** identifica padrões de comportamento entre usuários para prever itens de interesse. Entre os métodos mais eficazes, destaca-se o **Singular Value Decomposition (SVD)**, que decompõe a matriz de avaliações em fatores latentes representando similaridades entre usuários e itens (**Koren et al., 2009**).

Esse método ganhou notoriedade em competições como o **Netflix Prize** e continua sendo amplamente adotado em bases de dados densas, como a **MovieLens** (**Harper & Konstan, 2015**).

5.1.2. FILTRAGEM BASEADA EM CONTEÚDO (FBC)

A **filtragem baseada em conteúdo (FBC)** recomenda itens com base nas características dos conteúdos previamente apreciados pelo usuário. Técnicas como **Term Frequency-Inverse Document Frequency (TF-IDF)** e **similaridade do cosseno** são amplamente utilizadas para vetorização textual e cálculo de similaridade semântica (**Lops et al., 2011; Aggarwal, 2016**).

Essa abordagem é especialmente útil em cenários de **cold start**, nos quais há poucos dados históricos sobre um novo usuário, permitindo a recomendação com base apenas nas características dos itens.

5.1.3. SISTEMAS HÍBRIDOS

Os **modelos híbridos** combinam múltiplas técnicas para superar limitações individuais, como a escassez de dados e a falta de diversidade nas recomendações. **Burke (2002)** propôs diversas arquiteturas híbridas, entre elas a **fusão de scores (score-level fusion)**, amplamente adotada por seu equilíbrio entre desempenho e flexibilidade.

Estudos mais recentes, como **Campos et al. (2021)**, destacam o potencial dos modelos híbridos em **contextos educacionais e culturais**, devido à sua capacidade de adaptação a múltiplos perfis de usuário. Essa abordagem permite que as recomendações considerem tanto o **comportamento coletivo dos usuários** quanto

as **características dos conteúdos históricos**, favorecendo uma experiência mais rica e personalizada.

5.2. PROCESSAMENTO DA LINGUAGEM NATURAL (NLP) E EMOÇÕES

O uso de **Processamento de Linguagem Natural (NLP)** permite que o sistema compreenda o contexto subjetivo do usuário a partir de suas interações textuais. **Bibliotecas como TextBlob e VADER** possibilitam a **análise de sentimentos** em textos informais, fornecendo insights sobre o estado emocional do usuário.

Estudos como **Chen et al. (2020)** demonstram que a integração de NLP com sistemas de recomendação potencializa a personalização emocional, criando experiências mais envolventes. Aplicações recentes, como o **Emotional Recommender System** (**Wang et al., 2023**), utilizam o estado emocional do usuário como parâmetro para ajustar sugestões em tempo real, tornando as recomendações mais empáticas e contextualizadas.

5.3. INTEGRAÇÃO PRÁTICA ENTRE FC, FBC E NLP NO HISTFLIX

O sistema **HistFlix** combina abordagens de recomendação para proporcionar sugestões personalizadas e emocionalmente adaptadas. O fluxo do sistema é estruturado da seguinte forma:

1. O usuário interage com o sistema digitando uma frase livre.
2. O sistema aplica técnicas de NLP para extrair:
    - **A emoção predominante** (ex.: relaxado, curioso, reflexivo).
    - **Temas e palavras-chave** de interesse (ex.: Segunda Guerra Mundial, civilizações antigas).
3. Em paralelo, são executados:
    - **Filtragem Colaborativa (SVD)** para prever quais filmes semelhantes outros usuários apreciaram.

   - **Filtragem Baseada em Conteúdo (TF-IDF + similaridade do cosseno)** para encontrar filmes próximos aos descritos pelo usuário.

4. Os resultados são combinados via **score híbrido**, ajustado por um fator de ponderação **α**.
5. Um **filtro emocional** reorganiza ou remove conteúdos da lista final, priorizando aqueles compatíveis com o estado emocional do usuário.

Essa abordagem possibilita recomendações mais **relevantes, humanizadas e educacionais**, alinhadas ao objetivo do projeto de promover o aprendizado histórico por meio do audiovisual.

5.4. TRABALHOS RELACIONADOS

Estudos recentes demonstram o potencial da integração entre sistemas de recomendação e análise emocional/contextual:

- **EduFlix (Silva et al., 2022):** sistema de recomendação para ensino, utilizando metadados e NLP para promover a aprendizagem por meio de vídeos educacionais.
- **RecEmo (Zhao et al., 2021):** recomendador emocional baseado em análise de sentimentos para modular sugestões conforme o humor do usuário.
- **DeepMovie (Kumar & Sharma, 2020):** recomendador de filmes baseado em aprendizado profundo e análise de reviews via NLP.

Esses trabalhos reforçam a viabilidade da proposta do **HistFlix**, destacando sua inovação ao unir recomendação, NLP e análise emocional no contexto educacional.

5.5. CONSIDERAÇÕES FINAIS

Com base nos estudos apresentados, fica evidente que a combinação de abordagens clássicas de recomendação com técnicas modernas de NLP e análise emocional é uma tendência crescente e eficaz para sistemas inteligentes.

O projeto **HistFlix** fundamenta-se nesses princípios para desenvolver uma solução nova e educativa, capaz de compreender não apenas o contexto histórico dos conteúdos, mas também o estado emocional dos usuários, proporcionando uma

experiência mais envolvente e significativa no consumo de produções audiovisuais históricas.

6. CONCLUSÃO E TRABALHOS FUTUROS

Redigir a partir da aula 3.

7. REFERÊNCIA BIBLIOGRÁFICA

F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages. DOI=http://dx.doi.org/10.1145/2827872.

Chongchun Aggarwal. 2016. Recommender Systems: The Textbook. Springer. DOI=https://doi.org/10.1007/978-3-319-29659-3.

Robin Burke. 2002. Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction, 12, 331–370. DOI=https://doi.org/10.1023/A:1021240730564.

Pedro Campos, Fernando Díez, and Iván Cantador. 2021. Personalized and Adaptive Learning Systems Using Hybrid Recommender Models. International Journal of Artificial Intelligence in Education, 31, 845-867. DOI=https://doi.org/10.1007/s40593-021-00258-9.

Yong Chen, Jie Li, and Xiaojun Wang. 2020. Emotion-aware Recommendation: A Survey and Future Directions. ACM Transactions on Information Systems (TOIS) 38, 4, Article 34 (July 2020), 34 pages. DOI=https://doi.org/10.1145/3394621.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. Computer, 42, 8, 30-37. DOI=https://doi.org/10.1109/MC.2009.263.

Ravi Kumar and Anuj Sharma. 2020. DeepMovie: A Deep Learning-based Movie Recommender System Using NLP and Sentiment Analysis. International Journal of Advanced Computer Science and Applications, 11, 8, 317-325. DOI=https://doi.org/10.14569/IJACSA.2020.0110839.

Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. 2011. Content-Based Recommender Systems: State of the Art and Trends. In Recommender Systems Handbook, Springer, 73-105. DOI=https://doi.org/10.1007/978-0-387-85820-3_3.

Francesco Ricci, Lior Rokach, and Bracha Shapira. 2015. Recommender Systems Handbook (2nd ed.). Springer. DOI=https://doi.org/10.1007/978-1-4899-7637-6.

Marcelo Silva, Pedro Santos, and Rafael Almeida. 2022. EduFlix: A Recommender System for Educational Videos Based on Metadata and NLP Techniques. Journal of Educational Technology & Society, 25, 3, 117-132.

Lei Wang, Xiaojun Chen, and Yuhao Zhao. 2023. Emotional Recommender System: An Approach to Personalized Content Suggestions Based on User Mood. Journal of Artificial Intelligence Research, 76, 321-345. DOI=https://doi.org/10.1613/jair.2023.124.

Yunhao Zhao, Jian Huang, and Wei Lin. 2021. RecEmo: Emotion-Driven Content Recommendation Using Sentiment Analysis. IEEE Transactions on Affective Computing, 12, 2, 387-398. DOI=https://doi.org/10.1109/TAFFC.2021.3072674.

8. ANEXO

8.1. DEFINIÇÃO DAS BIBLIOTECAS PYTHON

Bibliotecas são fundamentais para a construção de qualquer software robusto e eficiente. No desenvolvimento do nosso sistema de recomendação, diversas bibliotecas Python foram empregadas para análise de dados, visualização, criação de recomendações e processamento de linguagem natural:

Manipulação e Análise de Dados

- **pandas**: manipula dados em formato de tabelas (chamados de DataFrames), organizando os dados da base MovieLens, como usuários, filmes e avaliações.
- **numpy**: realiza cálculos matemáticos eficientes com vetores e matrizes, auxiliando em operações numéricas e estatísticas, como normalização.

Visualização de Dados

- **matplotlib**: cria gráficos básicos, como histogramas e curvas de erro, úteis para exibição de desempenho do modelo.
- **seaborn**: complementa o matplotlib com gráficos mais intuitivos, como heatmaps e boxplots, facilitando a análise exploratória.

Sistema de Recomendação

- sklearn.metrics.pairwise.cosine_similarity: mede a similaridade entre vetores, essencial para calcular proximidade entre usuários ou filmes.
- **sklearn.feature_extraction.text.TfidfVectorizer**: converte descrições de filmes em vetores numéricos (TF-IDF), permitindo recomendação baseada em conteúdo.
- **surprise.SVD**: implementa filtragem colaborativa por meio de fatoração de matriz, prevendo avaliações de usuários.
- **surprise.Dataset** e **surprise.Reader**: estruturam dados no formato adequado para alimentar modelos de recomendação.

Processamento de Linguagem Natural

- **TextBlob**: analisa sentimentos e interpreta emoções dos textos inseridos, como "triste" ou "animado", personalizando recomendações.
- **nltk**: realiza tokenização e lematização, além de remover stopwords, facilitando o pré-processamento de textos antes da análise.