UNIVERSIDADE PRESBITERIANA MACKENZIE

Modelo de Sistema de Recomendação de Conteúdo para Títulos com Foco Educacional na Plataforma Netflix

FLÁVIO ESTEVAM NOGUEIRA ANDRADE	10441572
KAIQUE NASCIMENTO DE PAULA	10444268
MIGUEL SHIRAISHI	10431805
MOACYR SOUZA BARROS	10441179

São Paulo

SUMÁRIO

1 INTRODUÇÃO

A era do entretenimento digital é marcada por uma sobrecarga de conteúdo, com plataformas de streaming oferecendo vastos catálogos. Nesse cenário de abundância, é possível utilizar a ferramenta também para contribuir e atingir os objetivos das Nações Unidas em expandir o acesso e a melhora da educação, em conformidade com o item 4 dos Objetivos de Desenvolvimento Sustentável.

1.1 Contexto do trabalho

O setor de entretenimento digital, liderado por plataformas de streaming como a Netflix, vive uma era de conteúdo abundante. O principal desafio dessas empresas deixou de ser a oferta e passou a ser a curadoria eficiente. A solução para esse problema está nos sistemas de recomendação, modelos de predição baseados em algoritmos de Machine Learning. A disponibilidade de bases de dados públicas com metadados de títulos de plataformas como a Netflix tem impulsionado a pesquisa e o desenvolvimento de algoritmos de recomendação, fornecendo um campo de estudo prático para a comunidade científica.

1.2 Motivação

A motivação para este projeto é dupla: acadêmica e prática. Academicamente, a base utilizada é um benchmark ideal para o estudo de algoritmos de recomendação. Dominando a construção de um sistema de recomendação é uma habilidade altamente relevante nos dias de hoje e aplicável em diversas áreas.

1.3 Justificativa

Justifica-se este trabalho em três pilares:

- Técnico-Científico: Implementação e análise de um algoritmo de similaridade para a criação de um sistema de recomendação baseado em conteúdo.
- Extensionista: Alinhado com o Objetivo de Desenvolvimento Sustentável (ODS) 4, de "Educação de Qualidade", este projeto propõe que a mesma metodologia 1 utilizada para recomendar filmes de entretenimento pode ser aplicada a bases de dados com conteúdo educacionais. Assim sendo, um sistema de recomendação pode se tornar uma ferramenta valiosa para facilitar o acesso a documentários e filmes temáticos, contribuindo para a formação e o aprendizado contínuo.

- Prático: Simula o ambiente de trabalho de um cientista de dados, desde a compreensão do problema de negócio até a entrega de uma solução modelada.

1.4 Objetivos

Objetivo Geral: Desenvolver, implementar e avaliar o desempenho de um sistema de recomendação de títulos da Netflix, utilizando uma abordagem baseada em conteúdo.

Objetivos Específicos: Explorar, limpar e preparar a base de dados, compreendendo os desafios da curadoria de metadados.

- Implementar e analisar a eficiência de um modelo de recomendação baseado em conteúdo (TF-IDF e similaridade do cosseno);
- Avaliar o desempenho do modelo de forma qualitativa, por meio da análise da coerência das recomendações;
- Documentar todo o processo de forma clara e reproduzível.

2 REFERENCIAL TEÓRICO

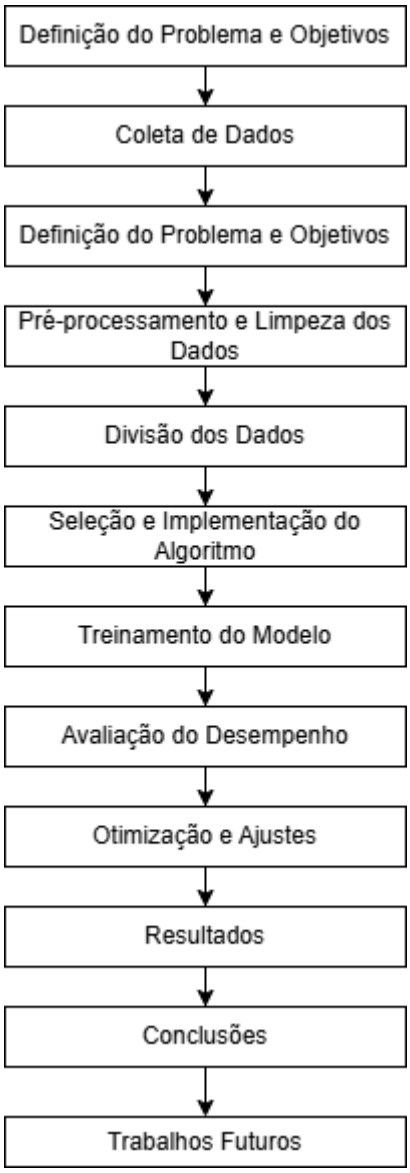
Os sistemas de recomendação têm se consolidado como ferramentas centrais para personalizar a experiência do usuário em ambientes digitais que oferecem grande volume de conteúdo, como plataformas de streaming, comércio eletrônico e redes sociais. Seu objetivo é reduzir a sobrecarga cognitiva e aumentar a relevância das sugestões apresentadas, filtrando informações de forma personalizada (RICCI; ROKACH; SHAPIRA, 2022). 2 As abordagens clássicas podem ser agrupadas em três grandes categorias: baseadas em conteúdo, filtragem colaborativa e modelos híbridos. Na recomendação baseada em conteúdo (Content-Based Filtering), a similaridade entre itens é calculada a partir de atributos explícitos (por exemplo, gênero, atores, diretores e sinopses no caso de filmes) e a preferência do usuário é inferida comparando itens consumidos previamente com novos candidatos (LOPS; DE GEMMIS; SEMERARO, 2011). Embora eficaz, essa técnica tende a gerar recomendações pouco diversificadas e restritas ao histórico já conhecido. A filtragem colaborativa (Collaborative Filtering), por sua vez, explora padrões de comportamento coletivo. Essa técnica pode ser user-based, quando usuários com histórico semelhante recebem recomendações mútuas, ou item-based, quando a similaridade entre itens é derivada da coocorrência de avaliações ou interações (SARWAR et al., 2001). O avanço desse método levou à utilização de fatoração de matrizes, como o Singular Value Decomposition (SVD) e o Alternating Least Squares (ALS), que se mostraram robustos para lidar com matrizes esparsas, típicas de ambientes de recomendação em larga escala (KOREN; BELL; VOLINSKY, 2009). Os modelos híbridos combinam múltiplas técnicas para superar limitações como o problema do cold start (situação em que há poucos dados sobre novos usuários ou itens) e a alta esparsidade das interações (BURKE, 2002). Essa estratégia oferece

maior cobertura e precisão em cenários complexos e heterogêneos. Com o avanço do aprendizado de máquina, surgiram soluções que incorporam redes neurais profundas e embeddings, capazes de capturar relações complexas entre usuários e itens. Modelos baseados em deep learning, como autoencoders e redes neurais convolucionais aplicadas a conteúdo multimídia, têm ampliado a capacidade de representar preferências e contextos, resultando em recomendações mais assertivas (ZHANG et al., 2019). A avaliação do desempenho é parte essencial na literatura de recomendação. Métricas como Root Mean Squared Error (RMSE) e Mean Absolute Error (MAE) são usadas em sistemas baseados em avaliações explícitas. Já em cenários de ranking, são comuns Precision, Recall, Mean Average Precision (MAP) e Normalized Discounted Cumulative Gain (NDCG), que medem a qualidade das listas 3 recomendadas (GUNAWARDANA; SHANI, 2015). A escolha da métrica deve refletir o objetivo do sistema e o tipo de interação disponível nos dados. No contexto específico de plataformas de streaming, estudos de caso mostram que a personalização é fator estratégico para engajar usuários e reduzir churn. A Netflix, por exemplo, combina fatoração de matrizes, enriquecimento semântico e aprendizagem contextual para sustentar recomendações altamente personalizadas (GÓMEZ-URIBE; HUNT, 2016). Esses fundamentos fornecem a base teórica para o presente projeto, que propõe desenvolver e testar um modelo de recomendação alinhado às melhores práticas acadêmicas e industriais.

3 METODOLOGIA

A metodologia adotada para o desenvolvimento deste Modelo de Sistema de Recomendação seguiu uma abordagem estruturada, alinhada ao ciclo de vida de um projeto de Ciência de Dados e focada na Filtragem Baseada em Conteúdo (FBC). Esta seção detalha a construção do modelo e os experimentos propostos, descrevendo a contribuição prática do grupo na área de sistemas de recomendação.

O processo foi dividido nas seguintes etapas, conforme implementado no código:



3.1 Definição do Problema e Objetivos

Nesta etapa inicial, definiu-se o foco do projeto: abordar a necessidade de curadoria eficiente em plataformas de streaming, com ênfase na recomendação de títulos com potencial educacional (documentários, filmes históricos e conteúdos culturais). O projeto

foi alinhado ao Objetivo de Desenvolvimento Sustentável (ODS) 4 – Educação de Qualidade.

O **Objetivo Geral** consistiu em desenvolver, implementar e avaliar o desempenho de um sistema de recomendação baseado em conteúdo para títulos presentes no catálogo da Netflix, utilizando metadados textuais para identificar relações de similaridade entre os títulos.

Entre os Objetivos Específicos, destacam-se:

- Selecionar e preparar uma base de dados adequada ao problema.
- Realizar pré-processamento textual e engenharia de features para melhorar a representação dos itens.
- Implementar uma abordagem de Filtragem Baseada em Conteúdo utilizando TF-IDF e Similaridade do Cosseno.
- Avaliar qualitativamente a coerência das recomendações geradas.
- Investigar possíveis caminhos de ajustes e otimização do modelo.

3.2 Coleta de Dados

A base utilizada foi o dataset público netflix_titles.csv, contendo 6.234 títulos da Netflix, com metadados como:

- título,
- diretor,
- elenco,
- país,
- data de inclusão na plataforma,
- classificação indicativa,
- duração,
- categorias,
- descrição.

A escolha da base atendeu aos requisitos do projeto, por fornecer dados textuais ricos e adequados para modelagem de similaridade entre itens.

3.3 Pré-processamento e Limpeza dos Dados

O pré-processamento envolveu limpeza, padronização e engenharia de atributos essenciais para a vetorização.

Limpeza de dados

As variáveis com dados faltantes (director, cast, country, date_added, rating, duration) foram tratadas da seguinte forma:

- preenchimento com string vazia '' para campos textuais,
- preenchimento com 'Missing' quando necessário para manter coerência semântica.

Engenharia de Features

A coluna tags foi criada pela concatenação das principais colunas textuais relevantes para identificar similaridade:

- director
- cast
- listed_in (categorias)
- description

Para melhorar a granularidade dos tokens:

- todo o texto foi convertido para minúsculas (str.lower()),
- espaços internos em nomes próprios foram removidos (ex.: *"peter cullen"* → *"petercullen"*), tornando cada nome um token único,
- essa abordagem aumentou a precisão da vetorização TF-IDF, pois evitou que nomes fossem divididos em tokens irrelevantes.

3.4 Divisão dos Dados

Como o modelo é baseado em cálculo de similaridade entre os itens (e não prevê uma variável alvo), não houve divisão tradicional em treino e teste.

O sistema de recomendação baseado em conteúdo gera uma matriz de similaridade calculada sobre o conjunto completo dos itens, razão pela qual a divisão de dados não se aplica a esta abordagem.

3.5 Seleção e implementação do algoritmo

A técnica escolhida foi a Filtragem Baseada em Conteúdo (Content-Based Filtering)**.**

Vetorização – TF-IDF

Utilizou-se o TfidfVectorizer, transformando o texto da coluna *tags* em uma matriz numérica de relevância de termos.

A teoria por trás da técnica assume que:

- quanto mais frequente e relevante for um termo em um item, maior seu peso;
- termos raros contribuem mais para diferenciar itens.

Foi utilizado o parâmetro:

- stop_words='english' para eliminar termos não informativos.

Métrica de Similaridade – Similaridade do Cosseno

A Similaridade do Cosseno foi empregada para medir a proximidade angular entre os vetores gerados pelo TF-IDF. O resultado foi uma Matriz de Similaridade, consolidando o relacionamento entre cada par de títulos da base.

3.6 Treinamento do Modelo

O processo de treinamento consistiu na construção da Matriz de Similaridade a partir do TF-IDF.

Etapas:

1. Aplicação do TF-IDF sobre a coluna *tags*.
2. Geração da matriz numérica TF-IDF.
3. Cálculo da Similaridade do Cosseno entre todos os vetores.
4. Armazenamento da matriz resultante para uso em consultas de recomendação.

Não houve parâmetros de treinamento supervisionado, já que o modelo apenas calcula similaridade entre textos.

3.7 Avaliação do Desempenho

A avaliação foi qualitativa, uma vez que não havia dados de interação de usuários (como ratings, histórico de visualização ou feedback). Assim, métricas quantitativas tradicionais (RMSE, NDCG, Precision@K, Recall@K) não se aplicam.

A validação consistiu na análise da coerência temática das recomendações geradas para títulos-chave.

3.8 Otimização e ajustes

Após a validação inicial, foram testadas alternativas de otimização:

Ajustes no TF-IDF

- Experimentação com max_features para reduzir dimensionalidade.
- Testes com *n-grams* (ex.: bigramas) para avaliar ganho de contexto.

Possíveis Melhorias Futuras

- Uso de vetorização semântica com Word2Vec**,** Doc2Vec ou Sentence-BERT para capturar contexto profundo.
- Combinação de abordagem baseada em conteúdo com filtragem colaborativa híbrida.
- Inclusão de dados de comportamento de usuário, caso disponíveis futuramente.

Essas melhorias representam caminhos para futuras versões do sistema.

4 RESULTADOS, CONCLUSÃO E TRABALHOS FUTUROS

4.1 Resultados Obtidos

A construção do sistema de recomendação baseado em conteúdo permitiu analisar, na prática, o comportamento do modelo diante de diferentes títulos do catálogo da Netflix. A matriz de similaridade gerada a partir do TF-IDF e calculada por Similaridade do Cosseno apresentou desempenho qualitativo satisfatório, produzindo recomendações coerentes com os temas, categorias e descrições dos itens.

Os testes realizados com títulos de naturezas distintas — séries históricas, dramas coreanos, animes, filmes de ação e conteúdos infantis — demonstraram que o modelo foi capaz de identificar relações semânticas relevantes entre obras. Isso confirma a adequação da abordagem escolhida para o propósito do projeto, especialmente considerando a ausência de dados de interação de usuários.

A tabela seguinte com os testes e seus devidos resultados, demonstra como o modelo treinado sugere títulos a partir de suas características como nacionalidade, gênero, diretor e afins.

Título	Recomendações
The Crown	Flowers | London Spy | Call the Midwife | Downton Abbey | The Frankenstein Chronicles | Hinterland | Black Earth Rising | The Real Football Factories | Traitors | Thomas and Friends
Squid Game	Prison Playbook | Love (ft. Marriage and Divorce) | Hospital Playlist | Chief of Staff | Man to Man | Color of Woman | The K2 | Hello, Me! | Run On | Persona
Air Force One	Tarzan 2 | What Lies Beneath | Magnolia | Boogie Nights | K-19: The Widowmaker | Paranoia | Behind Enemy Lines | Midnight Run | Krystal | Annabelle Hooper and the Ghosts of Nantucket
Control Z	You Cannot Hide | Ingobernable | Tijuana | El Dragón: Return of a Warrior | The House of Flowers | El desconocido | La Reina del Sur | 45 rpm | Who Killed Sara? | Miss Dynamite
My Little Pony: A New Generation	The Willoughbys | Hop | Ken Jeong: You Complete Me, Ho | Over the Moon | Walk of Shame | The Flintstones | Into the Grizzly Maze | Bureau of Magical Things | Dangerous Roads | Saving Zoë
The Garden of Words	Girls und Panzer der Film | EDENS ZERO | Words Bubble Up Like Soda Pop | Girls und Panzer | A Silent Voice | Nagi-Asu: A Lull in the Sea | Fireworks | The Could've-Gone-All-the-Way Committee | DRIFTING DRAGONS | BLAME!
Wheel of Fortune	Jeopardy! | Take Me | Taco Chronicles | Survivor | Ron White: If You Quit Listening, I'll Shut Up | The Circle Brazil | Cheer Squad | Sleepless Society: Insomnia | The Adventures of Sonic the Hedgehog | An American Tail: The Mystery of the Night Monster
RESIDENT EVIL: Infinite Darkness	Cyborg 009: Call of Justice | Transformers: War for Cybertron: Earthrise | Transformers: War for Cybertron: Kingdom | The End of Evangelion | Fullmetal Alchemist: Brotherhood | Neon Genesis Evangelion | Naruto | Ingress: The Animation | Naruto Shippuden: The Movie: The Lost Tower | EVANGELION: DEATH (TRUE)²
Good Morning Call	The Many Faces of Ito | Unlucky Ploy | Hormones | Roonpi Secret Love | Ex-Boyfriend | The Underclass | O-Negative, Love Can't Be Designed | How to Live Mortgage Free with Sarah Beeny | Miss in Kiss | A.I.C.O.
Ragnarok	Nobel | Kon-Tiki | Pee-wee's Playhouse | Fangbone | The Underwear | Sleepless Society: Insomnia | Bangkok Love Stories: Plead | Magnus | NSU German History X | Dracula

Nos casos das recomendações, percebe-se que:

- The Crown → recomendações de dramas históricos e séries britânicas;
- Squid Game → títulos relacionados a produções asiáticas e dramas intensos;
- Control Z → títulos hispano-americanos de investigação e suspense;

- The Garden of Words → recomendações dentro do gênero anime e temas semelhantes.

Entre os **pontos positivos**, destacam-se:
• coerência temática das recomendações;
• boa capacidade de distinção entre diferentes gêneros;
• execução leve e eficiente, adequada para prototipagem;
• reprodutibilidade do método.

Entre os **pontos negativos**, observou-se:
• tendência à baixa diversidade nas recomendações, típica de modelos puramente baseados em conteúdo;
• sensibilidade à qualidade dos metadados, especialmente descrições curtas ou genéricas;
• ausência de uma avaliação quantitativa formal, dado o cenário sem dados de usuários.

No geral, os resultados alcançados evidenciam que a técnica adotada é eficaz para exemplificação acadêmica e fundamenta possibilidades futuras de expansão e refinamento do sistema.

4.2 Conclusão

Este trabalho teve como objetivo desenvolver e avaliar um sistema de recomendação de títulos da Netflix utilizando uma abordagem baseada em conteúdo, com ênfase na construção de *features* textuais e no cálculo de similaridade entre itens. Além disso, alinhou-se ao ODS 4 – Educação de Qualidade, demonstrando como a metodologia pode ser aplicada na promoção de conteúdos educativos em plataformas de streaming.

Os objetivos propostos foram alcançados:

• Base de dados foi coletada, tratada e preparada adequadamente;

• Modelo de recomendação baseado em TF-IDF e Similaridade do Cosseno foi implementado;

• Foram realizados testes que demonstraram coerência temática das recomendações;

• Processo foi documentado de maneira clara e alinhada ao ciclo de vida de projetos de Ciência de Dados.

Ainda que a abordagem apresente limitações inerentes à técnica — como dependência dos metadados e baixa diversidade — ela se mostrou eficaz para fins didáticos e para a construção de um protótipo funcional, servindo como ponto de partida para sistemas mais robustos.

4.3 Melhorias Propostas

O projeto pode ser aprimorado em diferentes frentes, entre elas:

Qualidade da Representação dos Itens

• utilização de modelos de linguagem mais avançados (Word2Vec, Doc2Vec, Sentence-BERT);
• inclusão de bigramas e *n-grams* mais ricos;
• refinamento da coluna *tags*, com técnicas de *stemming* ou *lemmatization*.

Métricas e Avaliação

• implementação de métricas de ranking (Precision@K, Recall@K, NDCG), caso dados de interação sejam incorporados;
• experimentação com diferentes parâmetros do TF-IDF para comparação estruturada.

Diversificação das Recomendações

• aplicação de técnicas de *re-ranking* para aumentar a diversidade;
• fusão com modelos colaborativos, criando uma solução híbrida.

4.4 Trabalhos Futuros

As técnicas exploradas neste projeto podem inspirar múltiplas extensões e novos produtos de pesquisa, tais como:

• **Sistemas híbridos de recomendação**, combinando filtragem colaborativa e conteúdo, para lidar com sparsidade e cold start;
• **Recomendação de conteúdos educacionais** para plataformas de aprendizagem, museus virtuais ou bases de conhecimento social;
• **Análise de preferências culturais**, estudando padrões de consumo por região ou faixa etária;
• **Recomendação multimodal**, integrando texto, imagens (pôsteres) e áudio (trilhas sonoras);
• **Implementação em ambiente de produção**, com API, pipeline MLOps e dashboards de monitoramento;
• **Treinamento de embeddings próprios** utilizando técnicas modernas de Deep

Learning, como Transformers e Large Language Models especializados em recomendação.

Essas direções representam oportunidades concretas de evolução do trabalho, ampliando tanto sua relevância acadêmica quanto sua aplicabilidade prática.

5 REFERÊNCIAS

BURKE, Robin. Hybrid recommender systems: Survey and experiments. User Modeling and User-Adapted Interaction, v. 12, n. 4, p. 331-370, 2002.

GÓMEZ-URIBE, Carlos A.; HUNT, Neil. The Netflix Recommender System: Algorithms, Business Value, and Innovation. ACM Transactions on Management Information Systems (TMIS), v. 6, n. 4, p. 1-19, 2016.

• GUNAWARDANA, Asela; SHANI, Guy. Evaluating recommender systems. In: RICCI, Francesco; ROKACH, Lior; SHAPIRA, Bracha (Org.). Recommender Systems Handbook. 2. ed. New York: Springer, 2015. p. 265-308.

KOREN, Yehuda; BELL, Robert; VOLINSKY, Chris. Matrix factorization techniques for recommender systems. IEEE Computer, v. 42, n. 8, p. 30-37, 2009.

LOPS, Pasquale; DE GEMMIS, Marco; SEMERARO, Giovanni. Content-based recommender systems: State of the art and trends. In: RICCI, Francesco; ROKACH, Lior; SHAPIRA, Bracha (Org.). Recommender Systems Handbook. New York: Springer, 2011. p. 73-105.

RICCI, Francesco; ROKACH, Lior; SHAPIRA, Bracha. Recommender Systems Handbook. 3. ed. New York: Springer, 2022.

SARWAR, Badrul; KARYPIS, George; KONSTAN, Joseph; RIEDL, John. Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International Conference on World Wide Web. New York: ACM, 2001. p. 285-295.

ZHANG, Shuai; YAO, Lina; SUN, Aixin; TAY, Yi. Deep learning based recommender system: A survey and new perspectives. ACM Computing Surveys (CSUR), v. 52, n. 1, p. 1-38, 2019.