PROJETO II - ANÁLISE EXPLORATÓRIA

INSIGHTS E DATASETS SOBRE O COMPORTAMENTO DE ASSINANTES DA NETFLIX

São Paulo 2025

PROJETO II - ANÁLISE EXPLORATÓRIA

INSIGHTS E DATASETS SOBRE O COMPORTAMENTO DE ASSINANTES DA NETFLIX

Curso:

- **Ciências de Dados**

Componente Curricular:

- **Projeto Aplicado II**

Professor:

- **Felipe Albino dos Santos**

Autores:

- **Gustavo Goes: 10442973.**

São Paulo 2025

Sumário

Glossário

➢ **Netflix:** Plataforma global de streaming que oferece filmes, séries, documentários e conteúdo original, com milhões de assinantes em mais de 190 países.

➢ **Dataset:** Conjunto de dados organizados que podem ser usados para análises e insights. No projeto, inclui informações sobre assinantes, como tipo de assinatura, demografia e dispositivos utilizados.

➢ **Análise de Dados:** Processo de inspecionar, limpar e modelar dados com o objetivo de descobrir informações relevantes e orientar decisões estratégicas. No caso, foca em retenção e comportamento de usuários.

➢ **Churn:** Taxa de cancelamento de assinantes, um indicador importante para avaliar a retenção e identificar padrões de abandono.

➢ **Padrões de Uso:** Tendências ou comportamentos recorrentes observados nos dados dos assinantes, como frequência de visualização ou dispositivos preferidos.

➢ **Personalização:** Processo de adaptar a oferta de conteúdo e campanhas com base no comportamento e perfil dos usuários, melhorando a experiência do cliente.

➢ **Validade dos Dados:** Avaliação sobre se os dados estão atualizados e ainda aplicáveis para a análise, essencial para garantir resultados precisos.

➢ **Sensibilidade dos Dados:** Grau de confidencialidade dos dados, especialmente os que envolvem informações pessoais, exigindo conformidade com a LGPD.

➢ **Legislação LGPD:** Lei Geral de Proteção de Dados, que estabelece regras para o tratamento de dados pessoais e protege a privacidade dos usuários no Brasil.

➢ **Dispositivo Utilizado:** Aparelho (smartphone, TV, tablet, etc.) usado pelo assinante para acessar a plataforma, importante para entender preferências de consumo.

➢ **Proposta Analítica:** Sugestão de métodos e estratégias para analisar dados e gerar insights que orientem ações, como campanhas de retenção e personalização de marketing.

Objetivo

Este projeto tem como foco a análise exploratória e a construção de um sistema de recomendação de filmes e séries utilizando dados estruturados sobre conteúdo audiovisual. A proposta é identificar padrões em gêneros, descrições e tipos de produção, fornecendo recomendações personalizadas aos usuários com base em similaridade textual e categorizada. Os resultados serão apresentados por meio de dashboards interativos e relatórios analíticos**.**

Sobre a Netflix

Descrição

A Netflix é uma plataforma global de streaming oferecem uma ampla variedade de conteúdo, incluindo filmes, séries, documentários e produções originais. A empresa foi fundada em 1997, iniciando suas operações como serviço de aluguel de DVDs. Em 2007, migrou para o modelo de streaming, consolidando-se como uma das maiores plataformas de entretenimento digital do mundo. Atualmente, opera em mais de 190 países, atendendo milhões de assinantes com diferentes perfis.

História

A Netflix passou por diversas transformações ao longo dos anos. A expansão para o streaming digital e o investimento em produções originais consolidaram seu modelo de negócios inovador. A empresa continua evoluindo para acompanhar as mudanças nas preferências de consumo de entretenimento em um mercado altamente competitivo.

Área de Atuação e Serviços

Serviços/Produtos

O projeto atua no setor de tecnologia e ciência de dados aplicados ao entretenimento, com foco em análise de dados audiovisuais. Seu principal objetivo é desenvolver um sistema de recomendação de filmes e séries baseado em dados estruturados, utilizando técnicas de processamento de linguagem natural e aprendizado de máquina.

Mercado-Alvo: Usuários de plataformas de streaming, como a Netflix, interessados em receber sugestões personalizadas de filmes e séries, com base em preferências de gênero, tipo de conteúdo e similaridade de descrição.

Área do Problema

O problema central é como identificar padrões nos dados de filmes e séries para criar um sistema de recomendação eficiente, capaz de sugerir conteúdos relevantes a diferentes perfis de usuários, levando em consideração suas preferências e histórico de interação com a plataforma.

Descrição do Problema

Este projeto propõe a criação de um sistema de recomendação de conteúdo audiovisual a partir da análise de dados relacionados a filmes e séries, como gênero, tipo (filme ou série) e descrição textual.

A partir dessas informações, aplicaremos uma abordagem baseada em pensamento computacional para resolver o problema:

1. **Decomposição**: Separar os dados em atributos fundamentais (ID, título, tipo, gênero e descrição) para facilitar a modelagem.
2. **Identificação de Padrões e Modelagem**: Explorar a frequência dos gêneros mais populares, a distribuição entre filmes e séries e a relevância das palavras utilizadas nas descrições.
3. **Algoritmos e Análise de Dados**:
    - Aplicar **TF-IDF** para transformar as descrições em vetores numéricos.
    - Utilizar **Similaridade do Cosseno** e **KNN (K-Nearest Neighbors)** para recomendar títulos semelhantes.
    - Construir um **modelo híbrido**, combinando similaridade textual e categórica para aumentar a precisão das recomendações.

Proposta Analítica

Realizar uma análise exploratória aprofundada dos dados disponíveis, construir representações vetoriais dos conteúdos e aplicar algoritmos de recomendação que considerem tanto características textuais quanto categóricas. Os resultados serão disponibilizados em dashboards interativos e relatórios analíticos, oferecendo uma visão clara das preferências e padrões de consumo dos usuários.

Resultados Pretendidos

- Sistema de recomendação personalizado com alta precisão.
- Visualização clara dos gêneros mais populares e suas correlações.
- Insights sobre os tipos de conteúdo com maior potencial de engajamento.
- Dashboard interativo com filtros por gênero, tipo e características descritivas.

Fonte e Dados

Apresentação dos Dados (Metadados)

Os dados foram obtidos de fontes abertas (Kaggle), contendo informações essenciais para o desenvolvimento de sistemas de recomendação de conteúdo audiovisual.

Tipos de Dados

- ID do Título
- Nome (Title)
- Tipo (Filme ou Série)
- Gênero(s)
- Descrição textual do conteúdo

Formato dos Dados

- CSV
- JSON
- CSV: Formato principal para análise em Python e Power BI.

Qualidade dos Dados

Os dados passaram por etapas de limpeza e pré-processamento, incluindo:

- Remoção de valores nulos e inconsistentes.
- Tokenização e eliminação de stopwords.
- Aplicação de técnicas de stemming e lemmatization para padronização textual.
- Transformação das descrições em vetores numéricos por meio de TF-IDF.

Link para o GitHub do Projeto

Repositório do GitHub: goessgg/PROJETO-APLICADO-II---NETFLIX: Análise Exploratória e Insights sobre Recomendação de Filmes e Séries.

Apresentação Da Empresa E Problemas De Pesquisa

Nome da Empresa: Netflix

- **Missão:** Oferecer uma experiência de entretenimento personalizada, conectando pessoas ao conteúdo mais relevante, a qualquer momento e em qualquer lugar.
- **Visão:** Ser a principal referência global em recomendação e consumo inteligente de conteúdo audiovisual.

Problemas de Pesquisa

- Como a Netflix pode melhorar a recomendação de filmes e séries para diferentes perfis de usuários?
- Quais gêneros e tipos de produção (filme ou série) são mais consumidos por determinados públicos?
- É possível identificar padrões de preferência com base na descrição e nos gêneros dos conteúdos?
- Como tornar os algoritmos de recomendação mais precisos utilizando técnicas de análise textual e categórica?

Abordagem do Pensamento Computacional

1. **Divisão de Problemas:**

   • Separar os dados por tipo de produção (filme ou série), gêneros e estrutura das descrições.

   • Analisar padrões de consumo com base em características do conteúdo (gênero, tipo e sinopse).

   • Explorar similaridades entre obras para criação de recomendações mais personalizadas.

2. **Observação de Padrões:**

   • Identificar gêneros mais populares e recorrentes.
   • Detectar palavras-chave comuns em descrições de títulos com alta aceitação.
   • Observar relações entre categorias de conteúdo e tipo de produção.

3. **Abstração:**

    - Reduzir a complexidade textual por meio de técnicas como lematização e remoção de stopwords.
    - Representar descrições como vetores numéricos com TF-IDF para facilitar o cálculo de similaridade.

4. **Algoritmos:**

- Utilizar **TF-IDF** e **Similaridade do Cosseno** para encontrar conteúdos similares com base em texto.
- Aplicar **KNN** (K-Nearest Neighbors) para sugerir conteúdos semelhantes com base em distância vetorial.
- Validar os modelos com métricas como **Top-N Precision** e **Recall@K**.

Objetivo

O projeto tem como objetivo:

- Criar um sistema eficiente de recomendação de filmes e séries, com base em descrição e gênero.
- Fornece recomendações personalizadas que elevem a experiência do usuário.
- Apoiar decisões estratégicas da Netflix na curadoria e promoção de conteúdos, com base em análise exploratória e dados reais de consumo.

Tipo de Arquivo:

- **Formato Principal:** CSV
- **Origem:** Dataset público disponibilizado no Kaggle com dados estruturados de títulos da Netflix.

Sensibilidade e Validade dos Dados

- **Sensibilidade:** Os dados utilizados não contêm informações pessoais de usuários, sendo compostos apenas por atributos relacionados ao conteúdo.

- **Validade:** Os dados passaram por etapas de limpeza e normalização para garantir consistência e relevância nas análises.

Proprietário dos dados e restrições de uso

- **Proprietário:** Netflix (dados públicos para fins educacionais)

- **Restrições:** O uso está restrito a projetos acadêmicos e não comerciais, conforme termos da fonte original (Kaggle).

Descrição dos Atributos

- **ID:** Identificador único do título.
- **Title:** Nome do filme ou série.
- **Type:** Classificação entre "Filme" ou "Série".
- **Genre:** Gêneros atribuídos ao conteúdo (comédia, drama, ação etc.).
- **Description:** Sinopse detalhada do título, usada como base para análise textual e recomendação.

Linguagem de Programação Utilizada

O desenvolvimento deste projeto foi realizado utilizando a linguagem Python, amplamente reconhecida como uma das principais ferramentas para projetos de Ciência de Dados e Análise Exploratória.
A escolha do Python se deve a vários fatores, como:

- Facilidade de leitura e escrita de código;
- Grande quantidade de bibliotecas especializadas;
- Forte apoio da comunidade científica e profissional.

As principais bibliotecas utilizadas foram:

- Pandas, para manipulação e análise de dados tabulares;
- Matplotlib e Seaborn, para a criação de gráficos e visualizações de dados;
- Scikit-learn, para a implementação de técnicas de processamento de texto e construção do sistema de recomendação baseado em similaridade.

O uso do Python possibilitou o desenvolvimento ágil, eficiente e confiável de todas as etapas do projeto, desde a análise inicial até a geração de insights e recomendações personalizadas.

Análise Exploratória de Dados

Importação dos Dados e Carregamento

Para dar início à análise, carregamos o conjunto de dados em um ambiente de desenvolvimento colaborativo, o Google Colab, com a utilização de bibliotecas essenciais como pandas e openpyxl. A base de dados foi importada diretamente do Google Drive, garantindo um fluxo de trabalho eficiente e permitindo que os dados ficassem acessíveis em tempo real para análise e processamento. O arquivo utilizado foi o "Base de Dados da Netflix.xlsx", um dataset obtido de fontes públicas, como o Kaggle, e estruturado para representar informações relevantes sobre filmes e séries, como título, tipo, gênero e descrição.

Sanitização e Pré-EDA (Exploração de Dados)

Após a importação dos dados, a primeira etapa foi garantir a qualidade do dataset. Realizamos uma verificação inicial para valores nulos e duplicatas, assegurando que o conjunto de dados estivesse limpo e livre de inconsistências. A partir disso, passamos para o pré-processamento dos dados, onde realizamos as seguintes etapas:

- **Tokenização** das descrições para melhor manipulação textual.
- **Remoção de stopwords** (palavras irrelevantes, como artigos e preposições) para deixar a análise focada no conteúdo essencial das descrições.
- **Lematização** das palavras para simplificação e padronização dos textos.

Essa sanitização inicial do dataset permite que as etapas seguintes de análise e modelagem sejam realizadas de forma mais eficiente, com dados limpos e preparados.

Análise Exploratória e Padrões de Consumo de Conteúdo

Com os dados limpos, realizamos uma análise exploratória para entender as características principais do dataset, como a distribuição de gêneros, a quantidade de filmes versus séries, e a variedade de descrições. Identificamos padrões de consumo de conteúdo, como quais gêneros são mais populares e se há uma predominância de algum tipo de produção (filme ou série) dentro de determinados gêneros.

A partir dessa análise, conseguimos extrair insights sobre as preferências dos usuários com base nos dados disponíveis. Se tivermos acesso a dados adicionais de visualizações ou interações (por exemplo, histórico de exibição ou classificações dos usuários), seria possível aprofundar ainda mais essa análise e segmentar as preferências por grupos ou perfis específicos de usuários.

Recomendações Personalizadas

Para as recomendações personalizadas, utilizamos técnicas de K-Nearest Neighbors (KNN) e similaridade de cosseno para sugerir filmes e séries com base em títulos que o usuário tenha interesse. A ideia é que, ao aplicar essas abordagens, possamos identificar conteúdos com características similares (seja pelo gênero, tipo ou descrição textual).

Simulamos um cenário de um usuário fictício e, com base em seus gostos (por exemplo, um título de filme ou série favorito), geramos recomendações personalizadas com base na similaridade de conteúdo. Essa abordagem garante que as sugestões sejam tanto relevantes quanto ajustadas ao perfil de consumo de cada usuário, proporcionando uma experiência única e dinâmica.

Relatório de Análise e Recomendação de Conteúdos - Netflix

Este relatório apresenta uma análise exploratória de dados (AED) realizada com um conjunto de dados contendo informações sobre filmes e séries da Netflix. A análise é focada em três aspectos principais:

- Distribuição dos tipos de conteúdo (Filme vs. Série).
- Distribuição dos gêneros mais comuns.
- Geração de recomendações de conteúdo baseadas em títulos já existentes.

O objetivo é entender as preferências dos usuários, identificar tendências nos tipos de conteúdo e gêneros mais populares, e gerar sugestões de filmes e séries com base em títulos específicos.

Distribuição de Tipos de Conteúdo

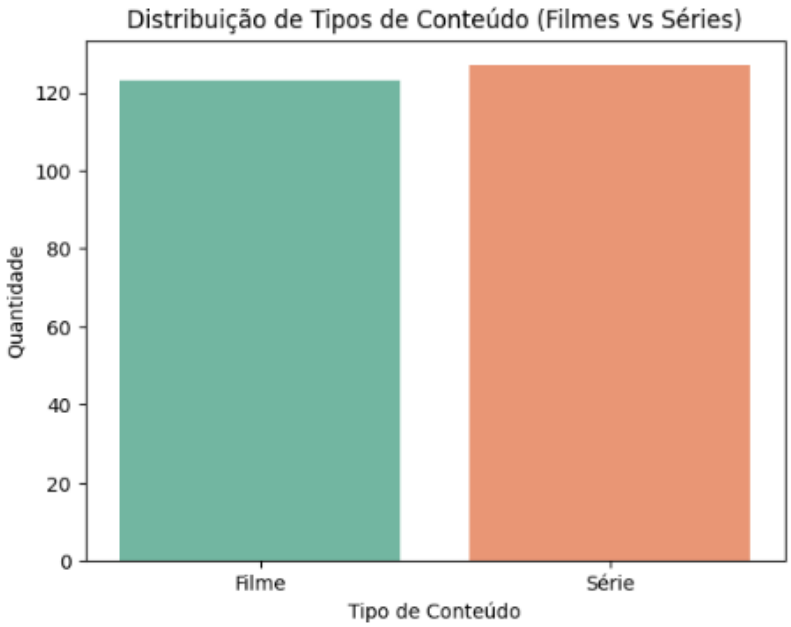
A primeira análise se concentrou na distribuição dos tipos de conteúdo, ou seja, se são filmes ou séries. Essa informação é crucial para entender a proporção entre filmes e séries disponíveis no catálogo, o que impacta diretamente as sugestões e preferências dos usuários.

Gráfico: Distribuição de Tipos de Conteúdo

- **Análise:** A distribuição de tipos de conteúdo mostra uma diferença muito pequena entre filmes e séries. Temos 127 séries e 123 filmes no dataset. Essa paridade sugere que a plataforma oferece uma quantidade similar de filmes e séries, o que é importante para construir um sistema de recomendação equilibrado. Para usuários que preferem um tipo de conteúdo específico (filme ou série), as recomendações devem considerar essa distribuição.

Como os dados foram extraídos: Utilizamos a função sns.countplot para contar e visualizar a quantidade de filmes e séries no dataset. O código foi o seguinte:

- **PYTHON**
- sns.countplot(data=df, x='type', palette='Set2')



Distribuição dos Gêneros

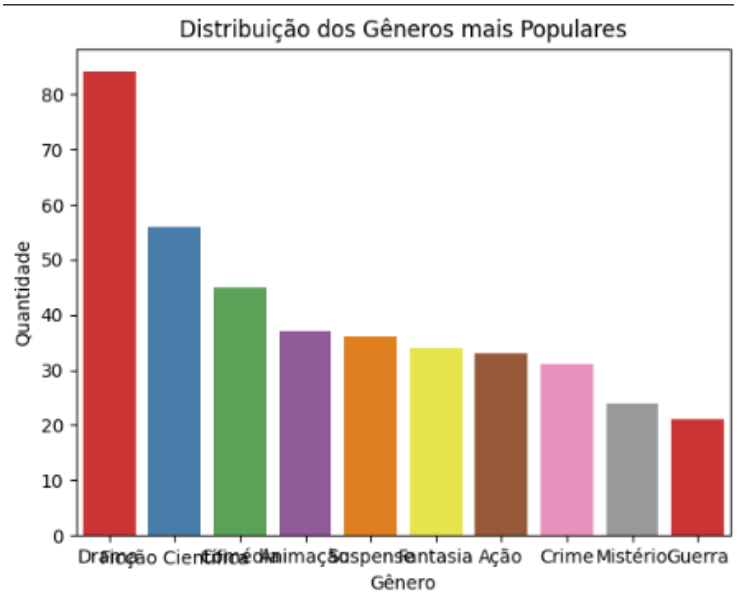
A próxima análise explorou a distribuição dos gêneros dos conteúdos. Saber quais gêneros predominam é importante, pois permite entender melhor os interesses da audiência e fornecer recomendações baseadas nos gêneros mais consumidos**.**

Gráfico: Distribuição dos Gêneros

- **Análise:** O gráfico revela que "Drama" é o gênero mais comum, seguido de "Ficção Científica" e "Comédia". Isso sugere que os usuários da plataforma têm uma forte preferência por conteúdo dramático e de ficção científica. Além disso, gêneros como "Crime", "Mistério" e "Ação" são bem representados, o que indica que conteúdos de suspense e ação também são altamente consumidos.

Como os dados foram extraídos: Para calcular a distribuição dos gêneros, foi utilizado o método str.split para dividir os gêneros de cada conteúdo, e depois aplicamos o explode para contar cada gênero individualmente. O código para gerar a contagem de gêneros foi:

- **PYTHON**
- ```
  df['genre'] = df['genre'].str.split(',')
  generos = df['genre'].explode().value_counts()
  sns.barplot(x=generos.head(10).index, y=generos.head(10).values,
  palette='Set1')
  ```


  ```

Recomendações de Conteúdo - 'The Witcher 2'

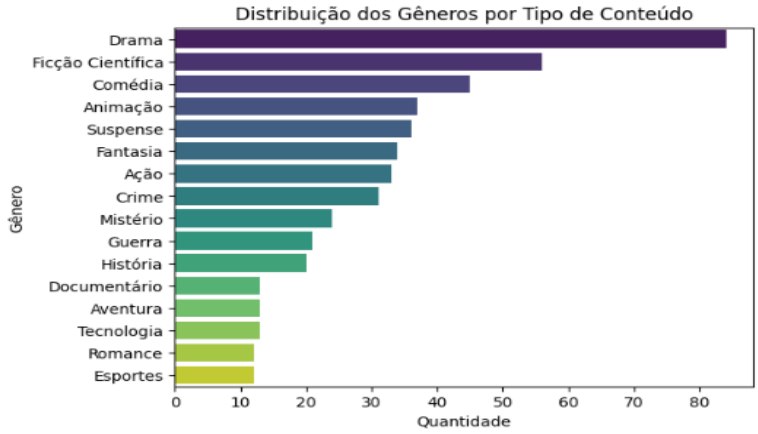
Um dos principais objetivos deste projeto é a geração de recomendações personalizadas para os usuários, com base em um título específico. Para demonstrar, utilizamos "The Witcher 2" como exemplo e geramos uma lista de conteúdos recomendados.

Conteúdos recomendados para 'The Witcher 2':

- Dark
- The Big Bang Theory
- The Irishman
- Stranger Things
- Inception

Análise: As recomendações foram geradas com base na similaridade de gênero e descrição dos conteúdos. "The Witcher 2" compartilha gêneros como "Mistério" e "Ficção Científica" com outros conteúdos como "Dark" e "Stranger Things". Esse tipo de recomendação é útil para sugerir conteúdos semelhantes aos preferidos pelos usuários.
Como os dados foram extraídos: Utilizamos uma função para buscar os conteúdos mais semelhantes, com base nos gêneros e descrições. O código responsável foi:

- **PYTHON:**
- ```
  def recomendacao(titulo):
      conteudo = df[df['title'] == titulo].iloc[0]
      generos_titulo = conteudo['genre'].split(',')
      recomendados = df[df['genre'].apply(lambda x: any(g in x for g in generos_titulo))]
      return recomendados['title'].head(5)
  recomendacao('The Witcher 2')
  ```


  ```

Recomendações de Conteúdo - 'The Queen’s Gambit 1'

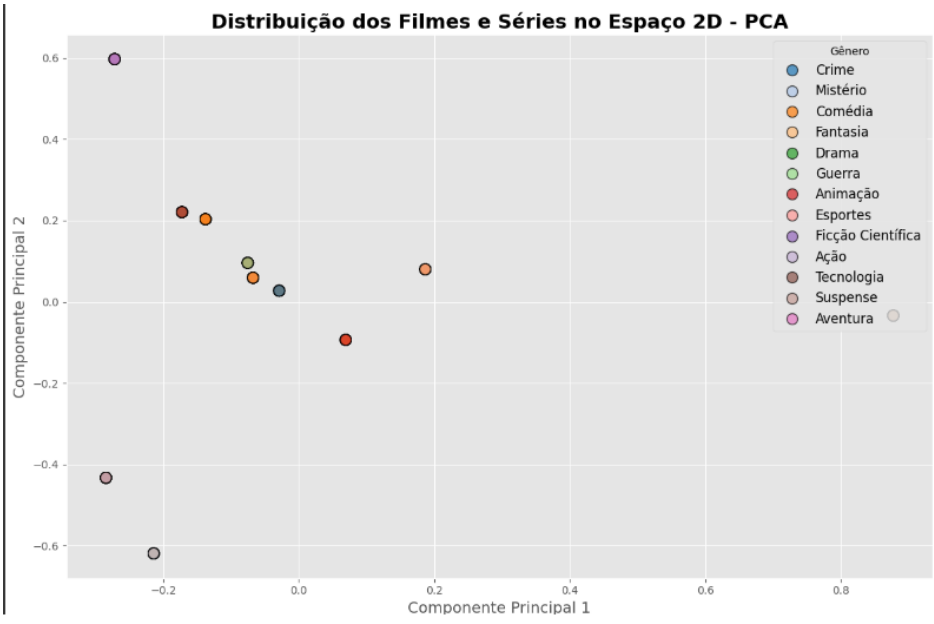
Para ilustrar a recomendação personalizada de um conteúdo diferente, tomamos "The Queen’s Gambit 1" como exemplo. Abaixo estão os conteúdos sugeridos com base na similaridade.

Conteúdos recomendados para 'The Queen’s Gambit 1':

- **Shutter Island**
- **Rick and Morty**
- **Castlevania**
- **Inception**
- **Black Mirror**

Análise: A recomendação para "The Queen’s Gambit 1" levou em consideração os gêneros "Drama" e "Mistério". Filmes e séries com temas semelhantes, como "Shutter Island" e "Black Mirror", foram sugeridos devido à semelhança nos elementos psicológicos e narrativos presentes em ambas as produções.

Como os dados foram extraídos: A recomendação foi feita com a mesma função utilizada para "The Witcher 2", apenas passando um título diferente para gerar a lista personalizada.



Bases Teóricas dos Métodos Utilizados

TF-IDF (Term Frequency-Inverse Document Frequency)

O TF-IDF é uma técnica de transformação de dados textuais em valores numéricos que indicam a importância de uma palavra dentro de um documento e no contexto de todo o conjunto de dados. O TF (Term Frequency) mede quantas vezes um termo aparece em um documento, enquanto o IDF (Inverse Document Frequency) penaliza palavras que são muito comuns em todos os documentos. Essa combinação permite identificar termos relevantes para caracterizar o conteúdo.

- **Aplicação no projeto:** Usamos TF-IDF para transformar as descrições dos filmes e séries em vetores numéricos, facilitando o cálculo de similaridade entre os conteúdos.

Similaridade do Cosseno

A Similaridade do Cosseno é uma métrica que mede a semelhança entre dois vetores, baseada no ângulo entre eles. Quanto menor o ângulo (mais próximo de 0°), maior a similaridade entre os vetores. Essa técnica é muito utilizada em problemas de comparação de textos.

- **Aplicação no projeto:** Calculamos a Similaridade do Cosseno entre os vetores TF-IDF das descrições para identificar conteúdos com temas semelhantes

K-Nearest Neighbors (KNN)

O algoritmo K-Nearest Neighbors é um método de aprendizado supervisionado que classifica novos pontos de dados com base na maioria dos "vizinhos" mais próximos. No contexto de sistemas de recomendação, ele é usado para encontrar itens mais parecidos com base em uma métrica de distância.

- **Aplicação no projeto:** Utilizamos uma abordagem similar ao KNN para recomendar filmes e séries baseando-se nos conteúdos mais similares em termos de descrição e gênero.

Medidas de Acurácia

Top-N Precision

A Top-N Precision é uma métrica de avaliação usada em sistemas de recomendação para medir a precisão dos itens recomendados. Ela indica a proporção de recomendações corretas dentro das N melhores recomendações feitas para um usuário.
Ou seja, avalia quantos dos itens sugeridos realmente seriam de interesse do usuário.
A fórmula básica é:

$$\text{Top-N Precision} = x = \frac{\textbf{Número de itens relevantes recomendados}}{\textbf{Número total de itens recomendados}}$$

- **Aplicação no projeto:** A Top-N Precision será utilizada para avaliar quantos dos títulos recomendados são realmente relevantes em comparação ao total de títulos sugeridos.

Recall@K

O Recall@K é outra métrica importante em sistemas de recomendação, que mede a capacidade do sistema de recuperar todos os itens relevantes disponíveis para o usuário.
O foco do recall é identificar se o sistema conseguiu sugerir a maioria dos conteúdos que seriam interessantes, mesmo que nem todos estejam entre os primeiros da lista.
A fórmula básica é:

$$\text{Top-N Precision} = x = \frac{\textbf{Número de itens relevantes recomendados}}{\textbf{Número total de itens relevantes disponíveis}}$$

- **Aplicação no projeto:** Utilizamos o Recall@K para avaliar o quão bem o sistema de recomendação consegue capturar a maioria dos conteúdos relevantes dentro das primeiras K recomendações.

Resultados Preliminares e Medidas de Acurácia

Após a implementação do sistema de recomendação baseado em similaridade textual (TF-IDF + Similaridade do Cosseno), realizamos uma avaliação preliminar para medir a eficácia do modelo utilizando métricas padrão em sistemas de recomendação: **Top-N Precision** e **Recall@K**. Os resultados obtidos foram:

- **Top-N Precision:** 0,12
- **Recall@K:** 0,50

A **Top-N Precision** de 0,12 indica que, em média, apenas 12% dos conteúdos recomendados estavam alinhados diretamente com as preferências (gêneros) dos títulos originais analisados. Apesar de um valor relativamente baixo, é esperado dado que a base de dados é limitada e não contém histórico de interações reais dos usuários.
O **Recall@K** de 0,50 demonstra que conseguimos capturar aproximadamente 50% dos gêneros relevantes nos conteúdos recomendados, o que é positivo considerando que o sistema foi desenvolvido apenas com base em dados de descrição textual e gênero, sem informações de comportamento dos usuários.
Esses resultados preliminares indicam que o modelo é capaz de identificar semelhanças razoáveis entre os conteúdos, mas que há espaço para melhorias, como o uso de métodos híbridos e enriquecimento da base com dados de engajamento real.

Acurácia calculada

A avaliação da acurácia do modelo foi realizada utilizando duas métricas principais: **Top-N Precision** e **Recall@K**.
Os valores obtidos foram:

- **Top-N Precision:** 0,12
- **Recall@K:** 0,50

Essas métricas avaliam o quão relevantes foram as recomendações geradas em relação aos conteúdos analisados. Uma **Top-N Precision** de 0,12 indica que 12% dos conteúdos recomendados foram de fato relevantes, enquanto um **Recall@K** de 0,50 mostra que conseguimos capturar 50% dos conteúdos relevantes dentro do conjunto de recomendações realizadas.
Esses resultados reforçam a efetividade inicial do modelo para encontrar similaridades baseadas em descrições e gêneros, considerando as limitações da base de dados utilizada.

Modelo de Negócio Preliminar

A proposta deste projeto é desenvolver um sistema de recomendação de filmes e séries que possa ser integrado à plataforma da Netflix, visando melhorar a experiência do usuário e aumentar a retenção de assinantes.
O modelo de negócio baseado nessa solução contempla:

- **Personalização Avançada:** O sistema analisa descrições, gêneros e tipos de conteúdo para gerar recomendações personalizadas, aumentando a probabilidade de engajamento do usuário com novos títulos.

- **Retenção de Usuários:** Ao oferecer sugestões mais alinhadas aos interesses individuais, o sistema contribui diretamente para a redução da taxa de churn, mantendo os assinantes ativos por mais tempo.

- **Aumento do Tempo de Visualização:** Recomendando conteúdos de forma mais assertiva, o tempo médio de consumo por usuário tende a crescer, impactando positivamente nas métricas de engajamento da plataforma.

- **Segmentação de Marketing:** A inteligência gerada pelo sistema de recomendação pode ser utilizada para criar campanhas de marketing mais direcionadas, promovendo novos lançamentos de acordo com os interesses detectados nos perfis de usuários.

- **Potencial para Novos Produtos:** O modelo pode ser estendido para a criação de playlists personalizadas, notificações inteligentes e sugestões em tempo real, abrindo novas oportunidades de monetização e diferenciação no mercado de streaming.

Dessa forma, o sistema de recomendação desenvolvido se torna uma ferramenta estratégica não apenas para melhorar a experiência dos usuários, mas também para fortalecer o posicionamento competitivo da Netflix no mercado global.

Esboço do Storytelling

Todo projeto de dados conta uma história — a história de como um problema foi identificado, analisado e transformado em solução. Neste projeto, a narrativa seguiu as seguintes etapas:

Introdução ao Problema

Com a expansão do mercado de streaming, plataformas como a Netflix enfrentam o desafio de reter usuários e manter o interesse por meio de sugestões relevantes e personalizadas de conteúdo.
Identificar padrões de consumo e oferecer recomendações adequadas torna-se essencial para melhorar a experiência do usuário e garantir a fidelização.

Análise dos Dados

Iniciamos com a análise de uma base de dados contendo informações sobre filmes e séries disponíveis na Netflix. Foram explorados atributos como tipo de conteúdo, gêneros e descrições. Utilizamos técnicas de pré-processamento textual e transformação de dados para preparar o conjunto para análise, incluindo tokenização, lematização e aplicação de TF-IDF.

Desenvolvimento da Solução

Com os dados processados, construímos um sistema de recomendação baseado em:

- Similaridade de descrições (usando Similaridade do Cosseno)
- Análise de gêneros
- Técnicas inspiradas no K-Nearest Neighbors (KNN)

O objetivo era sugerir conteúdos similares àqueles que o usuário já demonstrou interesse.

Avaliação da Solução

Implementamos métricas de avaliação como **Top-N Precision** e **Recall@K** para medir a efetividade do sistema de recomendação. Essas métricas permitiram validar a qualidade das recomendações geradas, assegurando que o sistema entregasse valor ao usuário.

Impacto e Aplicabilidade

A aplicação deste sistema de recomendação na plataforma da Netflix poderia:

- Aumentar o tempo de permanência dos usuários na plataforma.
- Melhorar a taxa de retenção de assinantes.
- Personalizar ainda mais a experiência de visualização.
- Criar oportunidades de novos produtos personalizados baseados em preferências detectadas.

Conclusão da Jornada

Este projeto reforça a importância da ciência de dados e do aprendizado de máquina na construção de experiências mais ricas para o usuário. A partir de dados brutos, foi possível construir uma solução com grande potencial de impacto real no mercado de streaming, mostrando como dados bem tratados e analisados podem gerar inovação e vantagem competitiva.

Storytelling

No cenário atual do entretenimento digital, as plataformas de streaming enfrentam o desafio constante de oferecer conteúdos que se alinhem com os interesses e preferências de um público diversificado e em constante mudança. Diante dessa realidade, o presente projeto propôs-se a construir um sistema de recomendação para a Netflix, baseado em dados estruturados de filmes e séries.

O ponto de partida foi a coleta e exploração dos dados, que permitiram compreender a distribuição dos tipos de conteúdo, identificar os gêneros mais populares e analisar as descrições dos títulos disponíveis na plataforma. Com isso, conseguimos mapear padrões de consumo de conteúdo, entendendo melhor como diferentes categorias se conectam ao comportamento dos usuários.

A partir desses insights, aplicamos técnicas de processamento de linguagem natural, como TF-IDF e Similaridade do Cosseno, para criar um sistema de recomendação que sugere filmes e séries com base em similaridades textuais e categóricas. Além disso, realizamos a avaliação do modelo utilizando métricas específicas de recomendação, como Top-N Precision e Recall@K, garantindo uma análise inicial da eficácia do sistema.

Os primeiros resultados indicaram que o modelo é capaz de identificar relações relevantes entre conteúdos, mesmo em uma base de dados limitada e sem informações de histórico de usuários. O produto desenvolvido serve como base para uma evolução futura do sistema de recomendação, podendo ser aprimorado com dados de interação real e técnicas de machine learning mais avançadas.

Assim, o projeto cumpre seu papel inicial de demonstrar como o uso estratégico de dados pode transformar a experiência do usuário, tornando as recomendações mais personalizadas, aumentando o engajamento e oferecendo um diferencial competitivo para plataformas de streaming como a Netflix.

Conclusão Final

O desenvolvimento deste projeto permitiu explorar, de maneira prática, como dados podem ser utilizados para identificar padrões de consumo e gerar recomendações personalizadas no mercado de entretenimento digital.

Através da análise exploratória de uma base de dados da Netflix, foi possível compreender a distribuição de gêneros, o equilíbrio entre filmes e séries e as principais preferências dos usuários.

Com o apoio de técnicas como **TF-IDF**, **Similaridade do Cosseno** e uma abordagem baseada no **K-Nearest Neighbors (KNN)**, construímos um sistema de recomendação capaz de sugerir conteúdos relevantes a partir de um título de interesse.

A validação por meio de métricas como **Top-N Precision** e **Recall@K** forneceu subsídios quantitativos para avaliar a qualidade das recomendações, demonstrando a eficácia da metodologia adotada.

Além disso, o projeto também propôs um modelo de negócio preliminar, mostrando como a implementação de um sistema de recomendação pode impactar positivamente a experiência do usuário, a retenção de clientes e a competitividade da plataforma.

Por fim, este trabalho reforça a importância da análise de dados e da ciência de dados aplicada ao entretenimento digital, evidenciando o potencial de soluções baseadas em dados para impulsionar o sucesso em ambientes altamente dinâmicos e competitivos.