Identificação do Gênero de um Filme através de sua sinopse utilizando Naïve Bayes

Daniel Nascimento da Silva

Centro de Informática (CIn)
Universidade Federal de Pernambuco
(UFPE)
Recife, Brasil
dns3@cin.ufpe.br

João Miguel Francisco de Souza

Centro de Informática (CIn)
Universidade Federal de Pernambuco
(UFPE)
Recife, Brasil
jmfs@cin.ufpe.br

Lucas Ferreira Alves

Centro de Informatica (CIn)
Universidade Federal de Pernambuco
(UFPE)
Recife, Brasil
lfa4@cin.ufpe.br

Miguel Pereira de Lemos

Centro de Informatica (CIn)
Universidade Federal de Pernambuco
(UFPE)
Recife, Brasil
mpl4@cin.ufpe.br

Millena Ferreira Marçal das Neves

Centro de Informatica (CIn)
Universidade Federal de Pernambuco
(UFPE)
Recife, Brasil
mfmn@cin.ufpe.br

***Abstract*— O objetivo deste trabalho é utilizar uma base de dados para a criação de um modelo capaz de caracterizar o gênero de um filme com base na sua sinopse, utilizaremos a linguagem de programação python. Com esse objetivo, utilizaremos a base de dados 'The Movie Database' (TMDB), contendo dados sobre diversas informações de filmes.**

Keywords—filme, identificador de gênero cinematográfico, classificador probabilístico, teorema de bayes, naive bayes

I. Introdução

Filmes são uma forma universal de entretenimento, capazes de transmitir emoções, contar histórias e refletir aspectos da sociedade. Cada obra cinematográfica se destaca por características únicas que a conectam a determinados gêneros, como ação, comédia, drama, ficção científica, entre outros. Identificar esses gêneros é uma tarefa importante tanto para produtores, que buscam entender seu público-alvo, quanto para plataformas de streaming, que precisam recomendar filmes com maior precisão.

A sinopse de um filme, geralmente breve e direta, contém informações essenciais sobre sua temática e enredo. Essa descrição é uma ferramenta poderosa para análise, pois nela estão presentes elementos que podem ser associados a categorias específicas de gênero.

Com o avanço das tecnologias, a possibilidade de identificar o gênero de um filme a partir de sua sinopse representa um desafio que combina estatística, probabilidade e aprendizado de máquina, e determinar, com base em padrões textuais, o gênero ao qual pertencem.

II. Objetivo

Por meio da análise de um banco de dados coletado e disponibilizado pela plataforma TMDB (The Movie Database), que contém informações detalhadas sobre filmes, incluindo sinopses e seus gêneros correspondentes, nosso objetivo é construir um modelo de classificação automática de gêneros cinematográficos.

Utilizando o conteúdo textual das sinopses, serão extraídas características relevantes, como palavras-chave, frequência de termos e padrões linguísticos, para que o classificador ingênuo de Bayes possa identificar, com base nos atributos definidos, a qual gênero cada filme pertence. O modelo será ajustado conforme necessário e avaliado utilizando validação cruzada, garantindo resultados consistentes e confiáveis.

III. Justificativa

A identificação de gêneros de filmes é essencial para aprimorar a experiência dos usuários em plataformas de streaming, serviços de recomendação e curadoria de conteúdos. Utilizando o classificador ingênuo de Bayes, é possível identificar padrões nas sinopses e associá-los aos gêneros correspondentes, promovendo uma categorização eficiente e precisa. Esse método probabilístico, fundamentado no Teorema de Bayes, é simples de implementar e tem ótimo desempenho em tarefas de classificação de texto.

O projeto será desenvolvido em Python, com o uso de bibliotecas como Pandas e Scikit-Learn, garantindo praticidade e eficiência. Essa abordagem contribui para otimizar sistemas de busca e recomendações, além de fornecer insights sobre as características linguísticas relacionadas a cada gênero cinematográfico, fortalecendo tanto a experiência dos usuários quanto a análise de dados no setor.

IV. METODOLOGIA

Neste projeto, almejamos classificar os gêneros de filmes presentes no banco de dados a partir das sinopses, utilizando o classificador ingênuo de Bayes. Para isso, realizaremos uma análise exploratória dos dados, visando identificar quais informações das sinopses são mais relevantes para a classificação e descartando dados que não contribuam para o modelo.

A implementação será realizada na plataforma Google Colab, utilizando a linguagem Python e bibliotecas essenciais para análise, visualização e construção do modelo.

A. Banco de Dados

A base dados que utilizaremos é a plataforma TMDB (The Movie Database)

Atributos do Dataset:

1. ID: Identificação única do filme
2. Título do Filme: Nome do filme para referência
3. Sinopse: Texto descritivo do enredo, utilizado como principal fonte de análise
4. Gênero: Rótulo categórico com os gêneros possíveis, como ação, comédia, drama, romance, etc. (atributo-alvo)
5. Ano de Lançamento: Ano em que o filme foi lançado (opcional, mas pode ajudar a categorizar tendências ao longo do tempo)
6. Duração: Duração do filme em minutos (opcional, caso o dataset forneça essa informação)
7. Popularidade (Rating): Média das avaliações ou nota atribuída (se disponível)
8. Número de Avaliações: Quantidade de avaliações registradas (opcional).

Atributos Textuais Extraídos da Sinopse

9. Frequência de Palavras-Chave: Contagem de palavras associadas a determinados gêneros (exemplo: "ação", "amor", "batalha", "risos")
10. Comprimento da Sinopse: Número total de palavras
11. Presença de Emoções ou Tons: Palavras que indiquem emoções específicas (como "triste", "alegre", "suspense")
12. Gramática e Estrutura: Uso de adjetivos, substantivos, e formas verbais predominantes (pode ser extraído via processamento de linguagem natural)
13. Temas Principais: Classificação automática de temas principais baseados em tópicos extraídos

B. Teorema de Bayes

Apesar de ser um conceito específico da probabilidade estatística, o Teorema de Bayes está mais presente no cotidiano do que muitos imaginam. Em situações que envolvem análise de dados e tomada de decisões, problemas podem surgir devido a "armadilhas intuitivas", que nos levam a conclusões equivocadas. Nesses casos, o Teorema de Bayes se mostra uma ferramenta poderosa, capaz de evitar esses erros ao fornecer uma abordagem mais estruturada e precisa para o cálculo de probabilidades.

O Teorema de Bayes é uma fórmula matemática utilizada para calcular a probabilidade condicional, ou seja, a probabilidade de um evento ocorrer dado que outro evento já aconteceu. Esse conceito é amplamente aplicado em diversas áreas, como aprendizado de máquina, medicina, economia e outras disciplinas que exigem análise preditiva e otimização de processos.

A essência do Teorema de Bayes está em utilizar informações prévias ou evidências conhecidas para atualizar a probabilidade de um evento. Assim, ele permite calcular a probabilidade da interseção de dois eventos ou avaliar cenários com eventos mutuamente exclusivos.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

- P(A|B): Probabilidade do evento A acontecer;
- P(B|A): Probabilidade de B acontecer, dado que A já ocorreu;
- P(A): Probabilidade de A ocorrer;
- P(B): Probabilidade de B acontecer.

Para profissionais que buscam otimização de processos e decisões fundamentadas em dados, compreender e aplicar o Teorema de Bayes oferece uma vantagem significativa. Sua versatilidade o torna um recurso indispensável em diferentes ferramentas de análise, contribuindo para soluções mais eficazes e decisões mais bem informadas.

C. Classificador Ingênuo de Bayes

O *Bayes Ingênuo* é um algoritmo de classificação amplamente utilizado em problemas de aprendizado de máquina, especialmente em tarefas de classificação de texto. Baseado no *Teorema de Bayes*, ele faz uma suposição simplificadora, porém eficiente: todas as características são independentes entre si, dado o rótulo da classe. Essa premissa, apesar de irrealista em muitos casos, simplifica os cálculos e torna o algoritmo muito rápido e eficiente. O termo "ingênuo" refere-se justamente a essa suposição de independência.

No contexto de classificação de texto, o Bayes Ingênuo assume que cada palavra contribui de forma independente para a probabilidade de um documento pertencer a uma determinada classe. Por exemplo, na análise de uma sinopse, o algoritmo ignora possíveis correlações entre palavras, tratando-as como variáveis isoladas. Apesar dessa limitação, o modelo frequentemente apresenta bons resultados em aplicações práticas.

A partir de dados de treinamento, o Bayes Ingênuo calcula as probabilidades de cada classe (gênero, no caso do nosso projeto) com base nas características observadas (as palavras presentes na sinopse). Ele então classifica uma nova entrada ao escolher a classe com a maior probabilidade. Este algoritmo é rápido, eficiente em termos computacionais e requer menos recursos quando comparado a métodos mais complexos.

No contexto do nosso projeto, cujo objetivo é determinar o gênero de um filme com base em sua sinopse, o Bayes Ingênuo se encaixa perfeitamente. Cada palavra da sinopse pode ser tratada como uma característica usada para prever o gênero do filme. Por exemplo, palavras como "amor" ou "relacionamento" podem ser indicadores de dramas ou romances, enquanto "espaço" e "futuro" são mais associadas à ficção científica. O modelo aprende esses padrões a partir dos dados fornecidos e utiliza esse aprendizado para classificar novas sinopses.

Embora o Bayes Ingênuo não capture relações mais complexas entre palavras, ele oferece uma solução inicial robusta e funcional. Isso o torna uma excelente escolha para o desenvolvimento do nosso sistema, com a possibilidade de evoluir para modelos mais avançados, caso necessário.

Cronograma de atividades

Data	Atividades
20/01	Seleção do Dataset
21/01 - 22/01	Elaboração de ideias
23/01 - 26/01	Elaboração da Proposta
27/01	Entrega da Proposta
01/02 - 28/02	Desenvolvimento do projeto
01/03	Finalização do projeto
01/03 - 02/03	Escrita do Relatório
03/03	Elaboração dos slides
04/03	Elaboração da apresentação
05/03 - 11/03	Possíveis reajustes
12/03 - 17/03	Entrega do Projeto
26/03	Apresentação do projeto

Referências

[1] ˜M. Paul, “Probabilidade: Aplicações ˜á Estatıstica”. 2 Edição. livros Técnicos e Científicos Editora.

[2] https://www.suno.com.br/artigos/teorema-de-bayes/

[3] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[4] https://didatica.tech/entenda-o-teorema-de-bayes/

[5] https://medium.com/data-hackers/o-teorema-de-bayes-descomplicado-entenda-e-aplique-na-pr%C3%A1tica-ea3429a407d3

[6] https://developer.imdb.com/non-commercial-datasets/

[7] https://developer.themoviedb.org/reference/intro/getting-started

[8] Gabriel Sacramento, “Naive Bayes: como funciona esse algoritmo de classificação” “https://blog.somostera.com/data-science/naive-bayes”

[9] Lauro Becker, “Algoritmo de Classificação Naive Bayes” - “https://www.organicadigital.com/blog/algoritmo-de-classificacao-naive-bayes/”