CURSO: Tecnologia em Ciência de Dados
POLO DE APOIO PRESENCIAL: Higienópolis
SEMESTRE: 4° Semestre - 2/2024
COMPONENTE CURRICULAR: PROJETO APLICADO III
NOME COMPLETO DO ALUNO: Beatriz de Souza Ferreira – RA: 10414697 Eduardo Nogueira Mota – RA: 10414834 Gustavo Castro Sangali – RA: 23023708 Jéssica Clara Da Silva Santos – RA: 10414974
TEMA: Projeto: Sistema de Recomendação de Livros utilizando Dados do Goodreads
NOME DO PROFESSOR: Carolina Toledo Ferraz

Sumário

Resumo

Este projeto visa desenvolver um sistema de recomendação de livros utilizando dados do Goodreads. A partir de técnicas de aprendizado de máquina e ciência de dados, o sistema propõe sugestões personalizadas com base em avaliações de usuários e características dos livros. O projeto adota duas abordagens principais: filtragem colaborativa, que identifica padrões de comportamento entre diferentes usuários, e recomendações baseadas em conteúdo, que utilizam a similaridade entre livros para gerar sugestões. A implementação é realizada em Python, utilizando bibliotecas especializadas como Scikit-learn e Surprise, com o objetivo de criar uma solução que possa ser aplicada em ambientes educacionais e bibliotecas comunitárias, incentivando o hábito da leitura.

Introdução

Nos últimos anos, os sistemas de recomendação têm desempenhado um papel crucial em diversos setores, ajudando usuários a encontrar produtos, serviços e informações de interesse de maneira personalizada. Em plataformas de leitura, como o Goodreads, a vasta quantidade de livros disponíveis torna difícil para o leitor identificar quais obras são mais relevantes de acordo com seu gosto e preferências. Por isso, um sistema de recomendação de livros torna-se uma ferramenta importante para guiar os usuários em suas escolhas de leitura, potencializando a descoberta de novos títulos e autores.

Motivados pela necessidade de auxiliar os leitores a encontrar obras que correspondam aos seus gostos e interesses este projeto tem como objetivo desenvolver um sistema de recomendação de livros utilizando dados disponibilizados pelo Goodreads, através do link Goodreads datasets. O foco será na implementação de um sistema que personalize sugestões de leitura com base no histórico do usuário, similaridade entre livros e classificações realizadas na plataforma.

O caráter extensionista do projeto reflete-se no impacto que ele pode ter na comunidade de leitores e na democratização do acesso à literatura. Um sistema de recomendação de livros pode facilitar o acesso a obras que não estão no mainstream, valorizando autores independentes e promovendo a diversidade literária. Além disso, o projeto poderá ser expandido para incluir recomendações baseadas em perfis de diferentes públicos, promovendo o gosto pela leitura em grupos que talvez não tenham fácil acesso a recomendações personalizadas de qualidade.

Este sistema pode ser utilizado por bibliotecas comunitárias, escolas e organizações que promovem a leitura, tornando mais acessível o processo de descoberta de novos livros e incentivando a continuidade no hábito da leitura, em sintonia com o objetivo 4 do desenvolvimento sustentável da ONU (Organização Das Nações Unidas).

O Objetivo de Desenvolvimento Sustentável (ODS) 4 da ONU, Educação de Qualidade, tem como objetivo garantir que todos tenham acesso a uma educação inclusiva, de qualidade e equitativa, e que tenham oportunidades de aprendizagem ao longo da vida.

Em resumo, este projeto visa desenvolver uma ferramenta que não apenas facilite a descoberta de novos livros, mas também contribua para uma cultura de leitura mais ampla e inclusiva. Ao personalizar as sugestões e promover a diversidade literária, o sistema se torna

um aliado fundamental na busca por obras que despertem o interesse e a curiosidade dos leitores.

1. Desenvolvimento em Python

Link para código no Google Colab:

https://colab.research.google.com/drive/1-eVj7PTUH2vZvnQthkVNtFxI22ykmqET?usp=sharing

O sistema de recomendação começa importando as bibliotecas essenciais, como pandas, numpy, seaborn, scipy, matplotlib, e algumas funções do scikit-learn. Essas bibliotecas são amplamente utilizadas para manipulação de dados, visualização, criação de matrizes esparsas e avaliação de modelos de aprendizado de máquina. Em seguida, os datasets de livros (books.csv) e avaliações (ratings.csv) são carregados usando a função pd.read_csv, que inclui a opção on_bad_lines='skip' para ignorar linhas problemáticas e garantir que os dados sejam carregados corretamente.

A primeira etapa na análise exploratória dos dados consiste em verificar a estrutura do dataset de livros, visualizando as primeiras linhas, descrevendo estatísticas básicas e identificando valores ausentes. As colunas do dataset de livros incluem informações como book_id, title, authors, average_rating, isbn, isbn13, language_code, num_pages, ratings_count, text_reviews_count, publication_date e publisher. A análise exploratória visual inclui histogramas para a distribuição das avaliações médias e do número de páginas, além de gráficos de dispersão (scatter plot) para analisar a relação entre a avaliação média e o número de páginas. Adicionalmente, a contagem de livros por idioma é calculada para fornecer uma visão geral da diversidade linguística presente no dataset.

A seguir, o dataset de avaliações é carregado e analisado de maneira semelhante, verificando a estrutura, estatísticas descritivas e valores ausentes. A visualização inclui um histograma das classificações para entender a distribuição das avaliações dos usuários. Além disso, são calculadas as classificações médias por livro e o número de classificações por usuário, fornecendo uma visão detalhada do comportamento de avaliação dos usuários.

Para a criação da matriz de usuário-livro, o código utiliza a função pivot_table para transformar os dados de avaliações em uma matriz esparsa, onde as linhas representam usuários e as colunas representam livros, com os valores sendo as classificações. Os valores NaN são preenchidos com 0 para indicar que o usuário não avaliou aquele livro. Essa matriz é fundamental para as técnicas de filtragem colaborativa utilizadas no sistema de recomendação.

O sistema de recomendação simplificado calcula a média das avaliações e o número de avaliações para cada livro. Apenas livros com pelo menos 50 avaliações são incluídos na análise para garantir uma base de dados confiável. Esses livros são classificados pela média

das avaliações, do maior para o menor, e os top N livros são selecionados como recomendação.

Para avaliar o desempenho do modelo, o dataset de avaliações é dividido em conjuntos de treino e teste usando a função train_test_split, com 80% dos dados sendo utilizados para treinamento e 20% para teste. A média das avaliações no conjunto de treino é calculada e usada para predizer as avaliações no conjunto de teste. O desempenho do modelo é avaliado usando a métrica Mean Squared Error (MSE), que mede a diferença média ao quadrado entre as avaliações preditas e as reais. Um MSE menor indica um melhor desempenho do modelo.

Além disso, o código inclui a utilização do modelo SVD (Singular Value Decomposition) da biblioteca Surprise. O SVD é uma técnica de fatoração de matriz que decompõe a matriz de avaliações em três matrizes menores, capturando as relações latentes entre usuários e itens. A validação cruzada é realizada com medidas como RMSE (Root Mean Squared Error) e MAE (Mean Absolute Error) para avaliar a precisão do modelo. O modelo é então treinado com todo o conjunto de dados e utilizado para prever a avaliação de um usuário específico para um livro específico.

Em resumo, o sistema de recomendação utiliza técnicas de filtragem colaborativa para recomendar livros com base nas avaliações dos usuários. O desempenho do modelo é avaliado usando MSE, RMSE e MAE, e a utilização do modelo SVD permite capturar relações latentes entre usuários e itens, proporcionando uma recomendação personalizada e precisa. Este sistema oferece uma base sólida para a implementação de recomendações de livros, combinando análise exploratória, processamento de dados e técnicas avançadas de aprendizado de máquina.

1.1. Bibliotecas Utilizadas

O desenvolvimento do sistema de recomendação envolverá o uso das seguintes bibliotecas:

- **Pandas**: Manipulação e análise de dados. Utilizada para carregar os datasets (books.csv e ratings.csv), realizar operações de agrupamento, cálculo de estatísticas descritivas e criação de tabelas dinâmicas.
- **NumPy**: Operações matemáticas e processamento de arrays. Utilizada para manipulação de dados numéricos e operações matemáticas, complementando as funcionalidades do pandas.
- **Scikit-Learn**: Algoritmos de aprendizado de máquina. Utilizada para dividir os dados em conjuntos de treino e teste (train_test_split), calcular a similaridade de cosseno entre

usuários (cosine_similarity), e avaliar o desempenho do modelo através da métrica Mean Squared Error (MSE).

• **Surprise**: Especializada para sistemas de recomendação. Utilizada para implementar e avaliar o modelo SVD (Singular Value Decomposition), facilitando a criação de recomendações e a realização de validação cruzada para avaliar o desempenho do modelo.

• **Scipy**: Extensão do Numpy para computação científica. Utilizada para criar e manipular matrizes esparsas (csr_matrix), o que é essencial para armazenar eficientemente a matriz de usuário-livro, especialmente quando muitos valores são zero.

1.2. Modelos de Aprendizado de Máquina

O modelo utilizado é um modelo baseado em conteúdo e utiliza a técnica de filtragem colaborativa. A ideia central é encontrar itens (livros) similares a outros que o usuário já interagiu (gostou).

- Baseado em Conteúdo: As recomendações são feitas com base nas características dos itens (no caso, informações sobre os livros como título, autor, avaliação média, etc.).
- Filtragem Colaborativa: Utiliza informações sobre as interações dos usuários com os itens (no caso, as avaliações) para fazer recomendações.

2. Repositório

Será utilizado um repositório no GitHub para organizar os materiais do projeto, como códigos, documentos e planejamentos.

O link do repositório é: https://github.com/EduNogueiraMota/MACK---projeto-aplicado-III.git

3. Cronograma do projeto:

Input	EDT	Nome da tarefa	Duração	Início	Término	Nomes dos recursos
25%	**1**	**PROJETO - SISTEMA DE RECOMENDAÇÃO DE LIVROS UTILIZANDO DADOS DO GOODREADS**	**96 Dias**	**19/08/2024**	**23/11/2024**	
100%	**1.1**	**FASE I – CONCEPÇÃO DO PRODUTO**	**23 Dias**	**19/08/2024**	**11/09/2024**	
100%	**1.1.1**	**Encontro Síncrono A1**	**1 Dia**	**19/08/2024**	**19/08/2024**	Prof. Carolina Toledo Ferraz
100%	**1.1.2**	Organização do grupo e repositório Github	4 Dias	20/08/2024	24/08/2024	Beatriz; Eduardo; Gustavo; Jéssica
100%	**1.1.3**	Escolha do tema e da base de dados	4 Dias	25/08/2024	29/08/2024	Beatriz; Eduardo; Gustavo; Jéssica
100%	**1.1.4**	Elaboração do documento inicial	11 Dias	30/08/2024	10/09/2024	Beatriz; Eduardo; Gustavo; Jéssica
100%	**1.1.5**	**Envio Fase I AVA**	**1 Dia**	**11/09/2024**	**11/09/2024**	Jéssica
0%	**1.2**	**FASE II – DEFINIÇÃO DO PRODUTO**	**23 Dias**	**12/09/2024**	**05/10/2024**	
0%	**1.2.1**	**Encontro Síncrono A2**	**1 Dia**	**30/09/2024**	**30/09/2024**	Prof. Carolina Toledo Ferraz
0%	**1.2.2**	Análise e limpeza da base de dados	6 Dias	12/09/2024	18/09/2024	Beatriz; Eduardo; Gustavo; Jéssica
0%	**1.2.3**	Escolha da técnica para treinamento do modelo	4 Dias	19/09/2024	23/09/2024	Beatriz; Eduardo; Gustavo; Jéssica
0%	**1.2.4**	Construção da prova de conceito	5 Dias	24/09/2024	29/09/2024	Beatriz; Eduardo; Gustavo; Jéssica
0%	**1.2.5**	Definição da nota de avaliação de desempenho	2 Dias	01/10/2024	03/10/2024	Beatriz; Eduardo; Gustavo; Jéssica
0%	**1.2.7**	**Envio Fase II AVA**	**1 Dia**	**05/10/2024**	**05/10/2024**	Jéssica
100%	**1.3**	**FASE III – METODOLOGIA**	**27 Dias**	**06/10/2024**	**02/11/2024**	
0%	1.3.1	**Encontro Síncrono A3**	**1 Dia**	**28/10/2024**	**28/10/2024**	Prof. Carolina Toledo Ferraz
0%	1.3.2	Implementação da técnica proposta	14 Dias	06/10/2024	20/10/2024	Beatriz; Eduardo; Gustavo; Jéssica
0%	1.3.3	Ajuste do pipeline de dados	4 Dias	21/10/2024	25/10/2024	Beatriz; Eduardo; Gustavo; Jéssica
0%	1.3.4	Documentação dos passos implementados	6 Dia	26/10/2024	01/11/2024	Beatriz; Eduardo; Gustavo; Jéssica
0%	1.3.6	**Envio Fase III AVA**	**1 Dia**	**02/11/2024**	**02/11/2024**	Jéssica
0%	**1.4**	**FASE IV – RESULTADOS E CONCLUSÃO**	**20 Dias**	**03/11/2024**	**23/11/2024**	
0%	1.4.1	**Encontro Síncrono A4**	**1 Dia**	**18/11/2024**	**18/11/2024**	Prof. Carolina Toledo Ferraz
0%	1.4.2	Organização e documentação dos resultados	8 Dias	03/11/2024	11/11/2024	Beatriz; Eduardo; Gustavo; Jéssica
0%	1.4.3	Finalização da documentação do projeto	4 Dias	12/11/2024	16/11/2024	Beatriz; Eduardo; Gustavo; Jéssica
0%	1.4.4	Produção do vídeo de apresentação	5 Dias	17/11/2024	22/11/2024	Beatriz; Eduardo; Gustavo; Jéssica
0%	1.4.5	Disponibilização do repositório no GitHub	1 Dia	23/11/2024	23/11/2024	Beatriz; Eduardo; Gustavo; Jéssica
0%	1.4.6	**Envio Fase IV AVA**	**1 Dia**	**23/11/2024**	**23/11/2024**	Jéssica

4. Definição da base de dados:

A base de dados escolhida para o projeto vem da coleção de datasets do **Goodreads**, conforme disponibilizada no repositório Goodreads datasets. Esses datasets contêm informações detalhadas sobre livros, avaliações e interações de usuários, tornando-se uma excelente fonte para criar um sistema de recomendação.

O conjunto de dados é dividido em duas partes, incluindo:

- **Books (Livros)**: Contém metadados sobre os livros disponíveis na plataforma.
- **Ratings (Avaliações)**: Contém as avaliações feitas pelos usuários em livros.

Principais Campos dos Datasets

4.1. Dataset de Livros (books.csv)

Este dataset contém informações sobre os livros cadastrados no Goodreads. Os principais campos incluem:

- **book_id**: Identificador único do livro.
- **title**: Título do livro.

- **authors**: Nome(s) do(s) autor(es) do livro.
- **average_rating**: Avaliação média do livro (de 0 a 5) com base nas classificações dos usuários.
- **isbn**: Código ISBN do livro (identificação internacional de obras publicadas).
- **isbn13**: Versão de 13 dígitos do ISBN.
- **language_code**: Código do idioma em que o livro está publicado.
- **num_pages**: Número de páginas do livro.
- **ratings_count**: Quantidade total de avaliações recebidas pelo livro.
- **text_reviews_count**: Quantidade de resenhas textuais sobre o livro.
- **publication_date**: Data de publicação.
- **publisher**: Editora do livro.

4.2. Dataset de Avaliações (ratings.csv)

Esse dataset contém as classificações numéricas feitas por usuários nos livros disponíveis. Os principais campos são:

- **user_id**: Identificador único do usuário que fez a avaliação.
- **book_id**: Identificador do livro avaliado.
- **rating**: Classificação dada pelo usuário, geralmente de 1 a 5 estrelas.

5. Referencial Teórico

O desenvolvimento de sistemas de recomendação é amplamente estudado na área de ciência de dados e inteligência artificial. Existem três abordagens principais:

- Filtragem Colaborativa: Essa técnica identifica padrões de comportamento entre usuários com preferências similares. A filtragem colaborativa pode ser dividida em duas categorias: baseada em usuários e baseada em itens. No primeiro caso, o sistema recomenda livros que foram bem avaliados por outros usuários com gostos semelhantes. No segundo, livros similares àqueles que o usuário já leu e avaliou positivamente são sugeridos.

- Recomendações Baseadas em Conteúdo: Nessa abordagem, o foco está nas características dos livros, como gênero, autor e descrições. O sistema analisa as preferências do usuário com base em livros que ele já avaliou e sugere novos títulos com características semelhantes.

- Modelos Híbridos: Combina as abordagens de filtragem colaborativa e recomendação baseada em conteúdo para maximizar a precisão das sugestões.

Além dessas abordagens, a análise de dados e a implementação de técnicas de aprendizado de máquina, como a decomposição em valores singulares (SVD) e algoritmos de vizinhos mais próximos (KNN), são amplamente utilizadas para melhorar a performance dos sistemas de recomendação.

6. Metodologia

O desenvolvimento deste sistema de recomendação segue uma abordagem estruturada em cinco etapas principais:

6.1. Coleta de Dados

A coleta de dados foi realizada utilizando os datasets do Goodreads, que contêm informações detalhadas sobre livros, avaliações de usuários e resenhas. Os dados foram obtidos por meio de downloads dos arquivos CSV, garantindo a inclusão de campos essenciais como book_id, título, autor, avaliação média, e avaliações de usuários. Esta etapa assegura a base necessária de dados para treinamento e teste do modelo de recomendação.

6.1. Pré-processamento

Os dados coletados passaram por um rigoroso processo de limpeza para garantir a qualidade e consistência dos mesmos. Durante esta fase, valores nulos foram identificados e removidos, duplicatas foram eliminadas e inconsistências foram corrigidas. Além disso, os nomes das colunas foram normalizados para garantir uma manipulação de dados mais eficiente. Este pré-processamento é crucial para evitar problemas durante as fases subsequentes de análise e modelagem.

6.3. Modelos de Aprendizado de Máquina

A criação do sistema de recomendação baseia-se na técnica de Filtragem Colaborativa, que utiliza as avaliações de diferentes usuários para identificar padrões de comportamento e sugerir livros que outros usuários com gostos semelhantes também gostaram. Foram empregados algoritmos como o SVD (Singular Value Decomposition) para decompor a matriz

de avaliações em componentes que capturam a relação latente entre usuários e itens. O SVD é particularmente eficaz em sistemas de recomendação, permitindo a descoberta de fatores subjacentes que influenciam as avaliações dos usuários.

6.4. Implementação

A implementação foi realizada utilizando Python e diversas bibliotecas como Pandas, NumPy, Scikit-learn e Surprise. O código desenvolvido manipula e analisa os dados, calcula similaridades entre usuários e itens, e treina o modelo de recomendação. A biblioteca Surprise foi utilizada para a implementação do algoritmo SVD e validação cruzada do modelo, garantindo a robustez e eficácia das recomendações geradas. O sistema foi implementado como uma API utilizando Flask, facilitando a integração com outras aplicações e sistemas.

6.5. Avaliação de Desempenho

Para avaliar o desempenho do modelo de recomendação, os dados foram divididos em conjuntos de treino e teste utilizando train_test_split. O modelo foi avaliado utilizando métricas como Mean Squared Error (MSE), Root Mean Squared Error (RMSE) e Mean Absolute Error (MAE). A validação cruzada foi realizada com a biblioteca Surprise, fornecendo uma visão detalhada da precisão do modelo. Um MSE menor indica um melhor desempenho, validando a eficácia do sistema de recomendação.

6.7. Resultados

Os resultados demonstram a capacidade do sistema de fornecer recomendações de livros personalizadas com um alto grau de precisão. O modelo foi capaz de identificar e sugerir livros com base nas preferências dos usuários, mostrando-se aplicável em bibliotecas comunitárias, plataformas educacionais e outras aplicações relacionadas. As avaliações de desempenho indicam que o sistema é robusto e eficaz, proporcionando uma experiência de usuário aprimorada através de recomendações precisas e relevantes. Esta abordagem não apenas melhora a satisfação dos usuários, mas também pode ser expandida para incluir outras métricas e algoritmos, aumentando ainda mais a precisão e a aplicabilidade do sistema de recomendação.

Referências

1. MCAULEY, Julian. **Dados do Steam**. Disponível em: https://cseweb.ucsd.edu/~jmcauley/datasets.html#steam_data. Acesso em: 06 set. 2024.
2. WAN, Mengting. **Dados Goodreads**. Disponível em: https://mengtingwan.github.io/data/goodreads . Acesso em: 06 set. 2024.