**Projeto de Pesquisa e Planejamento de Atividades**

|  |  |  |
| --- | --- | --- |
| **Aluno**: Gabriel Felix dos Santos | | **Data início curso**: 25/04/2023 |
| **Orientador**: Prof. Dr. Renato Máximo Sátiro | | **Defesa em:**  Mês/Ano |
| **Curso**: MBA Data Science e Analytics | **Modalidade**: Distância | Turma: 231 |

1. **Título do projeto**

Do demográfico ao híbrido: performances e resultados de algoritmos de recomendação

1. **Introdução**

*(*Nesta seção o autor deve contextualizar o assunto, estabelecendo uma relação com outros trabalhos publicados sobre o tema. Da mesma maneira, deve ser apresentada a justificativa para a realização do trabalho, situando a importância do problema estudado e apoiada em citações bibliográficas de fontes confiáveis. *O texto deverá ser escrito de forma impessoal.,* redigido em no máximo duas páginas e não deve conter subtópicos, figuras ou tabelas)

1. **Objetivo**

Analisar e comparar o desempenho em termos de custo de hardware e tempo de execução das principais abordagens de recomendação – Filtragem Demográfica, Filtragem Baseada em Conteúdo via sinopse e metadados, Filtragem Colaborativa via similaridade de usuários e similaridade de itens, e Filtragem Híbrida. Além disso, avaliar e comparar as recomendações geradas pelos modelo.

1. **Material e Métodos**

Este trabalhoutiliza o método de experimentação e, a fim de facilitar a compreensão dos materiais e métodos aplicados, esta seção foi dividida em seis partes: Ambiente de Desenvolvimento, Coleta de Dados, Dicionário das Variáveis, Limpeza e Transformação dos Dados, Modelos de Recomendação, e Comparação das Performances e dos Resultados.

4.1 Ambiente de Desenvolvimento

Todos os códigos foram desenvolvidos na versão 3.10.9 da linguagem de programação Python nos ambientes de desenvolvimento integrados Jupyter Lab e Jupyter Notebook, ambos na versão 3.5.3. Além do mais, os modelos de Filtragem Colaborativa utilizam o pacote Surprise, que requer a ferramenta Microsoft C++ Built Tools versão 14 para Windows.

Também foram aplicados pacotes do Python para acelerar o processo de desenvolvimento. As listas abaixo constam todos os pacotes com suas respectivas versões e usos neste trabalho:

Pacotes de Processamento de Dados:

* chardet versão 4.0.0 para checagem e transformação do formato de codificação das bases de dados para UTF-8[“Unicode Transformation Format”];
* numpy versão 1.23.5 para a manipulação das bases de dados;
* pandas versão 1.5.3 para a leitura e manipulação das bases de dados;
* re versão 2.2.1 para a aplicação de expressões regulares durante a limpeza e transformação dos dados;

Pacotes de Visualização de Dados:

* matplotlib versão 3.6.3 para a criação de gráficos;
* mplcyberpunk versão 0.7.0 para a estilização padrão dos gráficos;
* seaborn versão 0.12.2 para a criação de gráficos;

Pacotes de Modelagem e Avaliação:

* lightfm versão 1.17 para o treinamento e validação do modelo de Filtragem Híbrida;
* sklearn versão 1.2.1 para o treinamento e validação dos modelos de Filtragem Baseada em Conteúdo;
* spacy versão 3.7.2 para a preparação dos dados nos modelos de Filtragem Baseada em Conteúdo;
* surprise versão 1.1.3 para o treinamento e validação dos modelos de Filtragem Colaborativa;
* wordcloud versão 1.9.2 para a exploração da bolsa de palavras nos modelos de Filtragem Baseada em Conteúdo.

Por fim, vale salientar de que os códigos foram executados em um notebook com processador Intel® Core™ i5-8300H com frequência de 2,30GHz, placa de vídeo NVIDIA GeForce GTX 1050, 12 GB de RAM e sistema operacional Windows 10 baseado em 64 bits.

4.2 Coleta de Dados

As bases de dados encontram-se publicamente disponíveis na plataforma Kaggle e consistem em informações de animes, usuários e avaliações dos usuários sobre os animes assistidos em um dos principais sites do ramo.

A coleta foi realizada por um usuário do Kaggle utilizando a API[“Application Programming Interface”] Jikan durante 06 de jul. de 2023 a 06 de out. de 2023 abrangendo dados até o final da data de coleta.

Foram exportadas três bases em formato CSV[“Comma-Separated Values”]. A primeira contém apenas informações sobre os animes disponíveis, com 24 variáveis e 24.905 observações. A segunda contém apenas informações sobre os usuários disponíveis, com 16 variáveis e 731.290 observações. E a terceira contém as avaliações dos usuários sobre os animes assistidos, com 5 variáveis e 24.325.191 observações.

4.3 Dicionário das Variáveis

O próximo passo foi realizar o estudo das bases de dados e entender o que cada variável representa, resultando nos dicionários de variáveis abaixo:

4.3.1 Dicionário das Variáreis da Base de Dados de Animes

* Anime ID: número de identificação do item na plataforma;
* Name: nome original do item;
* English Name: versão do nome em Inglês Americano;
* Other Name: versão do nome em Japonês;
* Score: pontuação média seguindo a equação Bayesiana;
* Genres: gêneros atribuídos;
* Synopsis: breve descrição da história do item;
* Type: formato da animação, como episódios de televisão e filmes;
* Episodes: número de episódios. Filmes são considerados como um único episódio;
* Aired: período do início e final do lançamento do anime no formato “mês dia, ano”;
* Premiered: temporada de lançamento do anime no formato “estação ano”;
* Status: status de lançamento;
* Producers: produtoras envolvidas na produção do anime;
* Licensors: licenciadoras e plataformas de vídeo envolvidas na distribuição do anime;
* Studios: estúdios envolvidos na produção do anime;
* Source: material original da história, como mangás e livros;
* Duration: duração de cada episódio no formato “horas minutos” para aqueles acima de 60 minutos e “minutos” para os demais;
* Rating: classificação etária;
* Rank: posição de classificação do anime na plataforma com base na variável “Score”;
* Popularity: posição de classificação do anime na plataforma com base na popularidade do momento;
* Favorites: quantidade de usuários que marcaram o item como favorito;
* Scored By: quantidade de usuários que avaliaram o item;
* Members: quantidade de usuários que adicionaram o item na lista para assistir;
* Image Url: imagem promocional do anime.

4.3.2 Dicionário das Variáreis da Base de Dados de Usuários

* Mal ID: número de identificação do usuário na plataforma;
* Username: nome do usuário na plataforma;
* Gender: gênero sexual;
* Birthday: data de nascimento;
* Location: localização do usuário, podendo ser expressa em país, estado e cidade ou na combinação dos três;
* Joined: data de quando a conta foi criada na plataforma, seguindo o formato “ano-mês-diaThoras:minutos:segundos+fuso\_horário\_de\_greenwitch”;
* Days Watched: quantidade de dias que o usuário assistiu animes na plataforma;
* Mean Score: pontuação média das avaliações seguindo a equação Bayesiana;
* Watching: quantidade de itens sendo assistidos;
* Completed: quantidade de itens já completamente assistidos;
* On Hold: quantidade de itens que o usuário parou de assistir mas que ainda deixou na lista dele;
* Dropped: quantidade de itens que o usuário parou de assistir e que retirou da lista dele;
* Plan to Watch: quantidade de itens que o usuário adicionou na lista dele mas que ainda não começou a assistir;
* Total Entries: quantidade total de itens na lista do usuário;
* Rewatched: quantidade de itens que foram reassistidos;
* Episodes Watched: quantidade total de episódios assistidos de todos os itens.

4.3.3 Dicionário das Variáreis da Base de Dados de Avaliações

* User ID: número de identificação do usuário na plataforma;
* Username: nome do usuário na plataforma;
* Anime ID: número de identificação do anime na plataforma;
* Anime Title: nome original do anime na plataforma;
* Rating: pontuação que o usuário atribuiu ao anime.

4.4 Limpeza e Transformação dos Dados

Esta etapa foi dividida em quatro partes, sendo a primeira composta por limpezas e transformações aplicadas nas três bases de dados de forma geral, enquanto as demais consistem em processos aplicados individualmente em cada uma delas. Cada parte é explicada individualmente para facilitar o entendimento.

4.4.1 Limpeza e Transformação Gerais

Todas as bases de dados passaram pela checagem e transformação da codificação do arquivo para o formato UTF-8, além de terem as variáveis padronizadas em minúsculas, uso de sublinhados no lugar de espaços e palavras no idioma Inglês Americano. Os nomes finais das variáveis são os abaixo:

* Base de Dados de Animes: “id”, “title”, “english\_title”, “japanese\_title”, “score”, “genres”, “synopsis”, “type”, “episodes”, “aired”, “premiered”, “status”, “producers”, “licensors”, “studios”, “source”, “duration”, “rating”, “rank”, “popularity”, “favorites”, “scored\_by”, “members” e “image\_url”;
* Base de Dados de Usuários: “id”, “name”, “gender”, “birthday”, “location”, “joined”, “days\_watched”, “mean\_score”, “watching”, “completed”, “on\_hold”, “dropped”, “plan\_to\_watch”, “total\_entries”, “rewatched” e “episodes\_watched”;
* Base de Dados de Avaliações: “user\_id”, “username”, “anime\_id”, “anime\_title” e “rating”.

4.4.2 Limpeza e Transformação na Base de Dados de Animes

Foram removidos animes não anunciados ou lançados ainda, identificados pelos valores “Not yet aired” na variável “status” e “Not available” na variável “aired” respectivamente. Em seguida, foram descartadas as variáveis “english\_title”, “japanese\_title”, “aired” e “premiered”, pois somente a variável “title” foi utilizada para identificar os itens pelo nome, e os itens não anunciados e/ou lançados já foram removidos.

Depois, foi necessário tratar os valores ausentes, substituindo todos os “UNKNOWN” por -1 nas variáveis “score”, “episodes”, “rank” e “scored\_by” e converter estas de texto para número inteiro; enquanto que os “UNKNOWN” nas variáveis “genres”, “producers”, “licensors”, “studios” e “’rating” foram substituídos por um hífen “-”. Já na variável “synopsis”, os valores ausentes estavam representados como “No description available for this anime” e foram também substituídos por um hífen “-”.

Após filtrar as observações e substituir os valores ausentes, foi realizada a padronização das variáveis “title” e “synopsis”, convertendo os valores para minúsculo, sem acentos, e consistidos em apenas letras do alfabeto arábico, números, vírgulas, espaços e hifens, sendo "◯" o único caractere japonês mantido e substituído pela sua forma em "Rōmaji", "Maru". Ele foi mantido pelo fato de que há um anime no qual o título original é apenas o caractere e, consequentemente, ficaria com o nome em branco após o descarte de todos os caracteres japoneses.

Por fim, foi necessário verificar a existência de observações duplicadas. Alguns itens possuem o mesmo título, mas após checá-los na plataforma, foi constatado de que eles consistem em sequências das obras. Já sobre os gêneros, produtoras, licenciadoras e estúdios, foi realizada a exclusão de valores duplicados. Tomando os gêneros como exemplo, o valor “mistery, sci-fi, mistery, action” tornar-se-ou “mistery, sci-fi, action”.

4.4.3 Limpeza e Transformação na Base de Dados de Usuários

Foram removidos usuários que desativaram as contas, identificados pelos valores nulos nas variáveis “days\_watched”, “mean\_score”, “watching”, “completed”, “on\_hold”, “dropped”, “plan\_to\_watch”, “total\_entries”, “rewatched” e “episodes\_watched”. Em seguida, foram descartadas as variáveis “birthday” e “episodes\_watched”, pois não foram utilizadas no estudo.

Depois, foi necessário tratar os valores ausentes, substituindo-os por hifens “-” na variável “gender”, além de forçar o nome de usuário “None” para ser do tipo de dado texto, pois este valor é considerado como nulo dependendo da versão do Python.

Após isto, todas as variáveis numéricas foram convertidas de ponto flutuante para inteiro, pois todos os valores contidos nelas são e podem ser apenas inteiros.

Verificado se todos os usuários possuem pontuação média de avaliação entre zero e dez, sendo constatado apenas um com valor acima do esperado e, portanto, tendo o valor truncado para dez.

Por fim, os valores das variáveis “name”, “joined” e “gender” foram padronizados para minúsculo.

4.4.4 Limpeza e Transformação na Base de Dados de Avaliações

Foram removidas observações que contêm os animes e/ou os usuários que foram removidos nas duas bases de dados anteriores, além de forçar o nome de usuário “None” para ser do tipo de dado texto.

Em sequência, foi necessário realizar a padronização da variável “anime\_title” da mesma forma que foi feito nas variáveis “title” e “synopsis” da base de dados de animes.

Por fim, os valores da variável “username” foram padronizados para minúsculo.

A tabela abaixo indica a quantidade de variáveis e observações antes e depois da limpeza e transformação dos dados em cada base.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Base de Dados de Animes | Base de Dados de Usuários | Base de Dados de Avaliações |
| Variáveis Iniciais | 24 | 16 | 5 |
| Variáveis Finais | 21 | 14 | 5 |
| Observações Iniciais | 24.905 | 731.290 | 24.325.191 |
| Observações Finais | 23.748 | 731.282 | 23.796.586 |

4.5 Modelos de Recomendação

Neste trabalho, sete modelos de recomendação foram criados, sendo dois de Filtragem Demográfica, dois de Filtragem Baseada em Conteúdo, dois de Filtragem Colaborativa e um de Filtragem Híbrida. O processo de preparação dos dados, treinamento e validação foram descritos nas subseções abaixo.

4.5.1 Filtragem Demográfica

Dois modelos foram criados utilizando a metodologia de Filtragem Demográfica, enquanto o primeiro (modelo A) realiza recomendações dos itens com maiores pontuações de Média Bayesiana, o segundo (modelo B) realiza recomendações levando em consideração a popularidade dos itens entre os usuários.

Como a plataforma onde os dados foram coletados utiliza a Média Bayesiana para as pontuações dos animes, o modelo A precisou apenas ordenar os itens pela variável “score” de modo descendente e retornar os dez com maiores valores.

Já o modelo B trabalhou com a variável “popularity”, ordenando os itens de modo ascendente e retornando os dez com menores valores, ou seja, os dez mais acima na classificação de popularidade.

Caso fosse necessário realizar o cálculo da Média Bayesiana no primeiro modelo, a equação a ser aplicado é a abaixo:

Onde “v” consiste na quantidade de avaliações o item obteve, “m” na quantidade mínima de avaliações que o item necessita obter para ser considerado na Média Bayesiana, “A” na Média Aritmética do item e “C” na Média Aritmética de toda a base de dados.

4.5.2 Filtragem Baseada em Conteúdo

1. **Resultados Esperados**

Espera-se que o modelo de Filtragem Demográfica apresente o menor custo de hardware e tempo de execução devido à somente aplicação da Média Bayesiana para calcular as pontuações dos animes. No entanto, prevê-se que este modelo gere as recomendações menos precisas entre os demais, pois tenderá a recomendar os mesmos produtos a todos os usuários, ignorando os gostos individuais destes e as similaridades daqueles.

Em relação aos modelos A e B de Filtragem Baseada em Conteúdo, estima-se um aumento no custo tanto de hardware quanto de tempo em comparação com o modelo de Filtragem Demográfica. Esse aumento será devido à consideração da similaridade dos itens, seja por meio da sinopse (modelo A), seja por meio dos metadados de gênero, tipo e fonte original (modelo B). Contudo, espera-se que as recomendações sejam mais precisas, embora fiquem limitadas a animes do mesmo universo ou com temáticas semelhantes, respectivamente.

Sobre os modelos A e B de Filtragem Colaborativa, prevê-se um custo alto, talvez exigindo uma redução significativa na base de dados para tornar o treinamento do modelo viável na máquina do autor. Isso pode resultar em recomendações mais precisas e baseadas nos gostos individuais dos usuários, mas sem atingir todo o potencial dos modelos devido à redução da base de dados.

Quanto ao modelo de Filtragem Híbrida, estima-se que apresente um problema semelhante de custo, porém com uma menor redução na base de dados aplicando a técnica de Decomposição em Valores Singulares. Com isso, os resultados tendem a ser melhores do que os da Filtragem Colaborativa, mas também pode não ter o potencial total atingido.

1. **Cronograma de Atividades**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Atividades planejadas** | **Mês** | | | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| Iniciar conversa sobre o TCC |  | **x** |  |  |  |  |  |  |  |  |
| Fechar o tema e o objetivo |  |  | **x** |  |  |  |  |  |  |  |
| Escrever a Metodologia Pretendida |  |  | **x** |  |  |  |  |  |  |  |
| Entregar o projeto ao orientador |  |  | **x** |  |  |  |  |  |  |  |
| Entregar o projeto ao MBA USP/ESALQ |  |  |  | **x** |  |  |  |  |  |  |
| Aplicar a metodologia/coletar e estruturar os dados e iniciar a discussão dos resultados |  |  |  |  |  | **x** |  |  |  |  |
| Fechar introdução, material e métodos, iniciar os resultados e discussões e enviar ao orientador |  |  |  |  |  | **x** |  |  |  |  |
| Entregar os Resultados Preliminares ao MBA USP/ESALQ |  |  |  |  |  |  | **x** |  |  |  |
| Fechar resultados e iniciar a conclusão |  |  |  |  |  |  |  |  | **x** |  |
| Fechar a conclusão e iniciar o resumo |  |  |  |  |  |  |  |  | **x** |  |
| Fechar o trabalho como um todo e enviar ao orientador |  |  |  |  |  |  |  |  | **x** |  |
| Fechar o trabalho como um todo e enviar ao MBA USP/ESALQ (entrega do TCC e agendamento da defesa) |  |  |  |  |  |  |  |  |  | **x** |
| Entrega da versão final revisada |  |  |  |  |  |  |  |  |  | **x** |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

Projeto de Pesquisa; Resultados Preliminares; Entrega do Trabalho de Conclusão de Curso; Entrega da Apresentação da Defesa

1. **Referências Bibliográficas**

*(Listagem das bibliografias citadas no projeto de pesquisa, seguindo rigorosamente as Normas do MBA USP/Esalq. Consulte o manual de “Normas para Elaboração do Trabalho de Conclusão de Curso” disponível no Sistema TCC)*