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

Gabriel Felix dos Santos¹\*;Prof. Dr. Renato Máximo Sátiro2

1 Universidade Paulista UNIP. Bacharel em Ciência da Computação. Rua Maria Milhan Biazotto, 317 – Jardim Sampaio – CEP 168000-000 – Mirandópolis, São Paulo, Brasil.

2 Escola Superior de Agricultura "Luiz de Queiroz" - USP/ESALQ – Professor do curso de MBA em Data Science e Analytics – Rua Alexandre Herculano 120, sala T4 – CEP 13418-445 – Piracicaba (SP), Brasil.

Universidade Federal de Goiás – UFG. Doutor em Administração pelo Programa de Pós-Graduação em Administração da FACE/UFG. Avenida Esperança s/n, Câmpus Samambaia – Prédio da Reitoria. CEP 74690-900.

\*autor correspondente: [csfelix08@gmail.com](mailto:csfelix08@gmail.com)

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

**Resumo**

Tópico obrigatório para o depósito do TCC, porém opcional para a etapa dos Resultados preliminares.

**Palavras-chave:** (inserir até cinco palavras diferentes das contidas no título, separadas por ponto-e-vírgula).

**Atenção:** antes de enviar o arquivo para o Move - Sistema de TCCs, remova todas as instruções originais que estão abaixo do conteúdo dos tópicos.

**Introdução**

Em nenhum momento da história, a humanidade esteve tão interligada como na era atual, graças aos avanços computacionais e à acessibilidade a rede de internet. Relações pessoais agora estão a apenas um clique de distância e milhões de informações são trocadas a todo momento, desde textos até vídeos e áudios.

Esta abundância de dados favoreceu, e muito, a população mundial. Basta realizar uma pequena pesquisa sobre um determinado assunto que o usuário deparar-se-á com diversos conteúdos e até mesmo pesquisas científicas atuais do ramo. No entanto, também apresenta desafios significativos, pois o ser humano não consegue lidar com muitas informações simultâneas e, consequentemente, acaba tendo dificuldades de filtrar os conteúdos que o interessa Silva (2014), se desgasta e toma péssimas decisões (Ricci et al., 2011).

Devido a estes problemas, os Sistemas de Recomendação [SR] foram desenvolvidos, no qual a pesquisa ministrada por Whittaker e Sidner (1996) na classificação e filtragem de e-mails consiste em uma das primeiras aplicações da abordagem.

Com o passar do tempo, novas formas de SRs foram desenvolvidas e aperfeiçoadas e, atualmente, destacam-se alguns grupos de algoritmos, como os de Filtragem Demográfica, que agrupam os usuários em comunidades e realizam recomendações específicas para cada nicho; os de Filtragem Baseada em Conteúdo, que indicam itens similares aos que os usuários avaliaram positivamente; os de Filtragem Colaborativa, que agrupam os usuários com gostos semelhantes e realizam recomendações de itens de acordo com as similaridades entre estes grupos; e os de Filtragem Híbrida, que combinam duas ou mais abordagens de Filtragem para fornecerem melhores recomendações (Burke, 2002).

O motivo pelo qual há diferentes metodologias de filtragem de conteúdo está relacionado aos problemas encontrados nas aplicações dos SRs. Alguns algoritmos realizam recomendações muito próximas e flexíveis aos gostos dos usuários, mas em troca de uma quantidade de dados necessariamente vasta e um alto custo de processamento e tempo computacional no treinamento dos modelos de Inteligência Artificial. Outros algoritmos são menos custosos, porém fornecem recomendações ruins e generalizadas, como é o que acontece na Filtragem Baseada em Conteúdo, pois os gostos dos usuários não são sempre identificados apenas pela similaridade dos itens (Das et al., 2007).

Portanto, devido a estas problemáticas, o presente trabalho foi desenvolvido a fim de analisar o desempenho em termos de custo de hardware e tempo de execução das principais abordagens de recomendação descritas anteriormente a fim de pontuar os prós e contras de cada uma. Além de, como objetivo secundário, comparar as recomendações geradas pelos modelos.

**Material e Métodos**

Este trabalho utiliza 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.

**Ambiente de Desenvolvimento**

Os códigos foram desenvolvidos na versão 3.10.9 da linguagem de programação Python nos ambientes de desenvolvimento Jupyter Lab e Jupyter Notebook, ambos na versão 3.5.3. Além da aplicação de pacotes do Python para otimizar o processo de desenvolvimento, com suas versões e finalidades listadas na Tabela 1.

Tabela 1. Pacotes da linguagem Python aplicados

|  |  |  |  |
| --- | --- | --- | --- |
| Aplicação | Pacote | Versão | Uso |
| Avaliação de performance | Psutil | 5.9.0 | Medição do uso de “Random-Access Memory [RAM]” e de processador |
| Avaliação de performance | Time | 3.10.9 | Medição do tempo de execução dos códigos |
| Modelagem e validação | SKLearn | 1.2.1 | Treinamento e validação dos modelos de Filtragem Baseada em Conteúdo |
| Modelagem e validação | Spacy | 3.7.2 | Preparação dos dados para os modelos de Filtragem Baseada em Conteúdo |
| Processamento de dados | Chardet | 4.0.0 | Checagem e transformação do formato de codificação das bases de dados para “Unicode Transformation Format 8 [UTF-8]” |
| Processamento de dados | Numpy | 1.23.5 | Manipulação das bases de dados |
| Processamento de dados | Pandas | 1.5.3 | Leitura e manipulação das bases de dados |
| Processamento de dados | Re | 2.2.1 | Aplicação de expressões regulares na limpeza e transformação dos dados |
| Visualização de dados | Matplotlib | 3.6.3 | Criação de gráficos |
| Visualização de dados | Mplcyberpunk | 0.7.0 | Estilização de gráficos |
| Visualização de dados | Plotly | 5.22.0 | Criação de gráficos interativos |
| Visualização de dados | Seaborn | 0.12.2 | Criação de gráficos |

Fonte: Dados originais da pesquisa

Por fim, vale salientar 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.

**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 “Application Programming Interface [API]” Jikan durante 06 jul. 2023 a 06 out. 2023 abrangendo dados até o final do processo.

Foram exportadas três bases de dados em formato “Comma-Separated Values [CSV]”, no qual a primeira contém informações sobre os animes disponíveis, com 24 variáveis e 24.905 observações. A segunda contém informações sobre os usuários disponíveis, com 16 variáveis e 731.290 observações. E a terceira é constituída pelas avaliações dos usuários sobre os animes assistidos, com cinco variáveis e 24.325.191 observações.

Por fim, uma quarta base de dados consistida nos resultados de performances dos algoritmos de recomendação foi extraída durante as experimentações, possuindo nove variáveis e 70 observações.

**Dicionário das Variáveis**

O próximo passo foi realizar o estudo das bases e entender o que cada variável representa, resultando nos dicionários de variáveis abaixo representados pela Tabela 2, Tabela 3, Tabela 4 e Tabela 5.

**Resultados Preliminares**

O título da seção Resultados Preliminares deve ser alinhado à esquerda, grafado em negrito com as primeiras letras das palavras em letras maiúsculas. É permitido que a seção seja dividida em subtópicos, seguindo a de acordo com a descrição feita no item 1.1 Formato e margens e apresentados na mesma ordem da seção Material e Métodos. Nesta seção devem ser apresentados os resultados parciais obtidos na pesquisa, ou seja, os resultados obtidos até o momento.

**Atenção:** antes de enviar o arquivo para o Move - Sistema de TCCs, remova todas as instruções originais que estão abaixo do conteúdo dos tópicos.

**Nos resultados preliminares, é importante apresentar uma descrição detalhada dos dados coletados, incluindo a forma como estão organizados e estruturados. Isso pode envolver tabelas, gráficos ou outros recursos visuais que ajudem a visualizar as informações de maneira clara e concisa. Além disso, se aplicável, é recomendável esboçar os primeiros modelos ou hipóteses derivadas dos dados, fornecendo uma visão inicial do padrão ou tendência identificados. Uma análise descritiva dos dados também deve ser incluída, destacando medidas de centralidade, dispersão e forma da distribuição, conforme apropriado para o tipo de dados coletados. Essa análise preliminar dos resultados serve como ponto de partida para uma investigação mais aprofundada, ajudando a orientar a interpretação dos dados e a identificação de padrões ou relações significativas.**

**Conclusão(ões) ou Considerações Finais**

Tópico obrigatório para o depósito do TCC, porém opcional para a etapa dos Resultados preliminares. A seção deve conter frases curtas, apresentando as conclusões e inferências elaboradas a partir da discussão dos resultados obtidos até o momento. Os autores não devem, em hipótese alguma, mencionar, citar ou reproduzir resultados de outros estudos na(s) Conclusão(ões) ou Considerações Finais do TCC. Além disso, esta seção não deve conter tabelas ou figuras, sendo redigida de forma sucinta.

**Atenção:** antes de enviar o arquivo para o Move - Sistema de TCCs, remova todas as instruções originais que estão abaixo do conteúdo dos tópicos.

**Agradecimentos** (opcional, 1 parágrafo sucinto)

O título da seção Agradecimentos deve ser alinhado à esquerda e grafado em negrito, com a primeira letra da palavra grafada em letra maiúscula. Trata-se de uma seção opcional, de no máximo três linhas, na qual o autor agradece aqueles que contribuíram de maneira relevante para o desenvolvimento do trabalho e elaboração do TCC, mas que não tiveram o envolvimento intelectual necessário à atribuição de coautoria do mesmo, abstendo-se totalmente da menção ou citação de nomes de empresas, instituições ou pessoas que permitiram ou contribuíram com o desenvolvimento do trabalho, a menos que esteja documentalmente autorizado a fazê-lo.

**Atenção:** antes de enviar o arquivo para o Move - Sistema de TCCs, remova todas as instruções originais que estão abaixo do conteúdo dos tópicos.

**Referências**

Neste tópico deverá conter todas as referências dos trabalhos citados no texto e formatadas seguindo rigorosamente as normas do MBA USP/Esalq. Para mais informações, vide o manual de “Normas para Elaboração do Trabalho de Conclusão de Curso” disponível no Move - Sistema de TCC.

**Atenção:** antes de enviar o arquivo para o Move - Sistema de TCCs, remova todas as instruções originais que estão abaixo do conteúdo dos tópicos.

**Apêndice ou Anexo** (opcional)

Apêndices são textos e/ou documentos que foram elaborados pelo autor e que são importantes para complementar a argumentação do trabalho. Anexos são textos ou documentos que ilustram, mas que não foram elaborados pelos autores. Apêndices deverão seguir as mesmas normas de formatação do restante do texto, inclusive para figuras e tabelas.

O TCC deverá conter no máximo 30 páginas, incluindo o(s) Apêndice(s) e/ou Anexo(s).

**Atenção:** antes de enviar o arquivo para o Move - Sistema de TCCs, remova todas as instruções originais que estão abaixo do conteúdo dos tópicos.