MBA USP ESALQ

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

Aluno: Gabriel Felix dos Santos

Orientador: Prof. Dr. Renato Máximo Sátiro

Introdução

- Abundância de dados e Big Data
- Dificuldades em filtrar conteúdos; Desgaste dos usuários; Decisões equivocadas
- Sistemas de Recomendação [SRs]
- Reconhecimento de gostos pessoais
- Bases de dados robustas
- Poder computacional

Objetivos

1. Algoritmos de recomendação

   - Filtragem Demográfica: 2 modelos
   - Filtragem Baseada em Conteúdo: 2 modelos
   - Filtragem Colaborativa: 2 modelos
   - Filtragem Híbrida: 1 modelo

2. Desempenhos

   - Tempo de execução
   - Consumo médio de CPU
   - Consumo médio de RAM

3. Recomendações

   - Adaptação aos gostos individuais

Material e Métodos

1. Base de Dados

- Coleta: Plataforma Kaggle - Stream de animes - dados até 06 de out. de 2023
- animes.csv: 24 variáveis; 24.905 observações; nome, gênero, sinopse, score, estúdio...
- users.csv: 16 variáveis; 731.290 observações; nome, sexo, score médio, episódios assistidos...
- ratings.csv: 5 variáveis; 24.325.191 observações; id do usuário, nome do usuário, id do anime, nome do anime e avaliação
- benchmarkings.csv: 9 variáveis; 70 observações; nome do modelo, tempo de execução, consumo médio de CPU, consumo médio de RAM...

Material e Métodos

2. Transformações nas Bases de Dados

- Padronização dos nomes das variáveis
- Descarte de variáveis
- Limpeza dos dados
- Remoção de animes não anunciados e não lançados
- Remoção de usuários desativados
- Remoção de avaliações desativadas

Tabela 6. Quantidade de variáveis e observações antes e depois da Limpeza e Transformação dos dados

	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
Variáveis removidas (%)	12,5000%	12,5000%	0,0000%
Observações iniciais	24.905	731.290	24.325.191
Observações finais	23.748	731.282	23.796.586
Observações removidas (%)	4,6456%	0,0011%	2,1731%

Fonte: Dados originais da pesquisa

Material e Métodos

3. Filtragem Demográfica (Modelo A e Modelo B)

- Modelo A: Avaliação dos Itens - Média Bayesiana $\frac{(c \cdot m) + (n \cdot r)}{c + n}$
- Modelo B: Popularidade
- 1 Comunidade: Usuários da Plataforma (+75% dos usuários estão com localização inexistente ou não informada)

Material e Métodos

4. Filtragem Baseada em Conteúdo (Modelo C e Modelo D)

- Sinopses dos itens (Modelo C) e Gêneros, Tipos e Fontes Originais (Modelo D)
- Padronização dos textos em minúsculo, sem quebras de linha e sem pontuações
- * Remoção das Palavras de Parada e de Substantivos Próprios
- Tokenização por palavra
- * Lematização
- Bolsa de Palavras
- Frequência do Termo - Frequência Inversa do Documento [FT-FID]
- Similaridade do Cosseno

* processo efetuado apenas no Modelo C

Material e Métodos

5. Filtragem Colaborativa (Modelo E)

- Itens bem avaliados por usuários semelhantes
- Itens acima do percentil 75 na quantidade de avaliações
- Separação dos itens para treino e validação
- Tabela Pivô
- Padronização subtraindo pela média aritmética dos itens
- Correlação de Pearson
- Cálculo tabela de predições
- Erro Quadrático Médio da Raiz: 1.2704

Material e Métodos

6. Filtragem Colaborativa (Modelo F)

- Itens semelhantes aos bem avaliados pelo usuário, sendo estes mesmos itens bem avaliados por usuários semelhantes
- Itens acima de 75 mil avaliações
- Separação dos itens para treino e validação
- Tabela Pivô
- Padronização subtraindo pela média aritmética dos itens
- Correlação de Pearson
- Cálculo tabela de predições
- Erro Quadrático Médio da Raiz: 5.7741

Material e Métodos

7. Filtragem Híbrida (Modelo G)

- Modelo A (Filtragem Demográfica pela Média Bayesiana)
- Modelo D (Filtragem Baseada em Conteúdo - Gêneros, tipos e fontes originais)
- Modelo E (Filtragem Colaborativa – Itens bem avaliados por usuários semelhantes)

Material e Métodos

8. Medição das Performances

- Treinamento, validação e geração de recomendações
- Dez iterações
- Tempo da iteração
- Consumo mínimo, máximo e médio de CPU (%)
- Consumo mínimo, máximo e médio de RAM (%)

Resultados

1. Performances

- DF - Bayesian Mean: Modelo A
- DF - Popularity: Modelo B
- CBF - Plots: Modelo C
- CBF - Metadatas: Modelo D
- CF - User-Based: Modelo E
- CF - Item-Based: Modelo F
- Hybrid Filtering: Modelo G

Figura 1: Tempo médio de execução dos modelos

Fonte: Resultados originais da pesquisa

Resultados

1. Performances

- DF - Bayesian Mean: Modelo A
- DF - Popularity: Modelo B
- CBF - Plots: Modelo C
- CBF - Metadatas: Modelo D
- CF - User-Based: Modelo E
- CF - Item-Based: Modelo F
- Hybrid Filtering: Modelo G

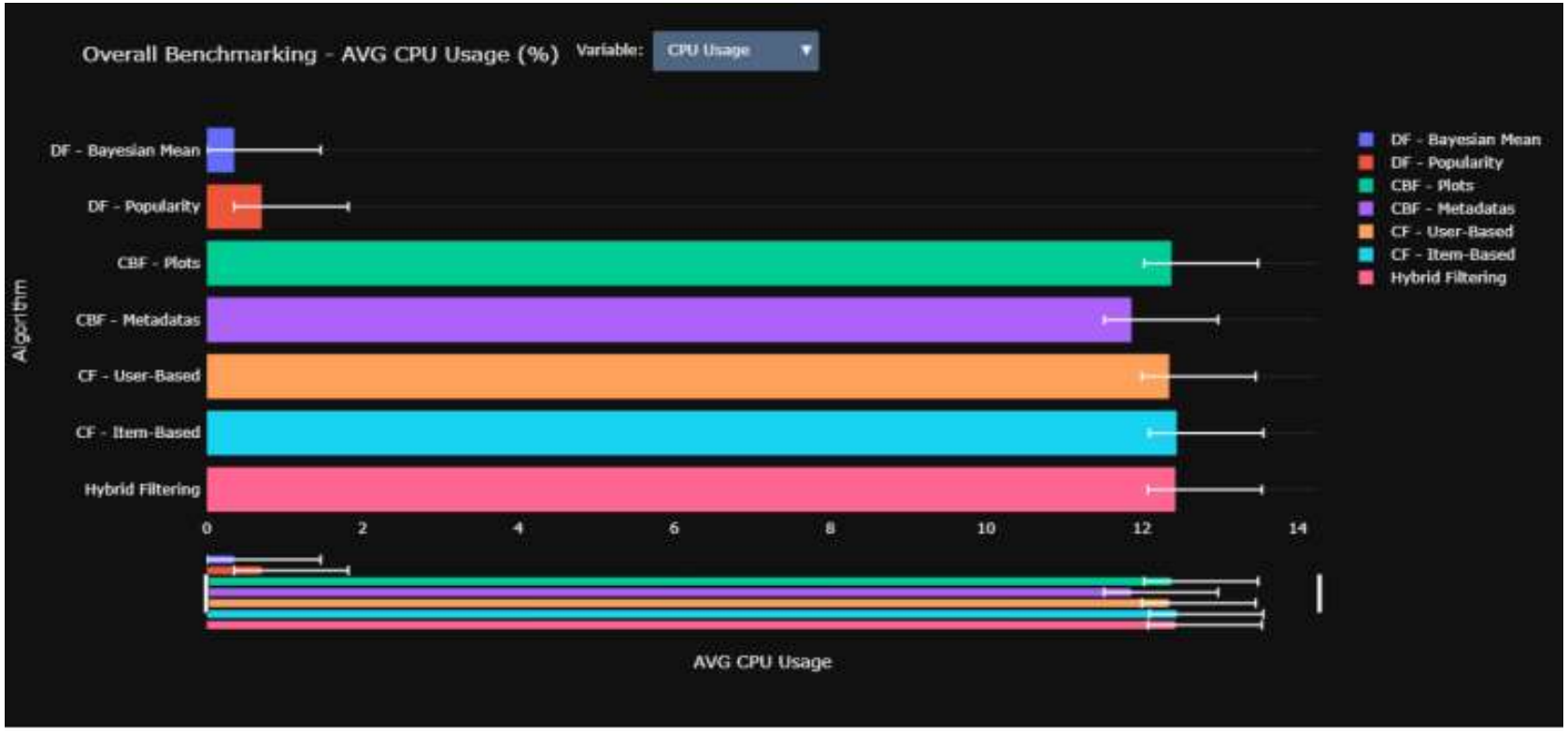


Figura 2: Consumo médio de processamento de CPU dos modelos

Fonte: Resultados originais da pesquisa

Resultados

1. Performances

- DF - Bayesian Mean: Modelo A
- DF - Popularity: Modelo B
- CBF - Plots: Modelo C
- CBF - Metadatas: Modelo D
- CF - User-Based: Modelo E
- CF - Item-Based: Modelo F
- Hybrid Filtering: Modelo G

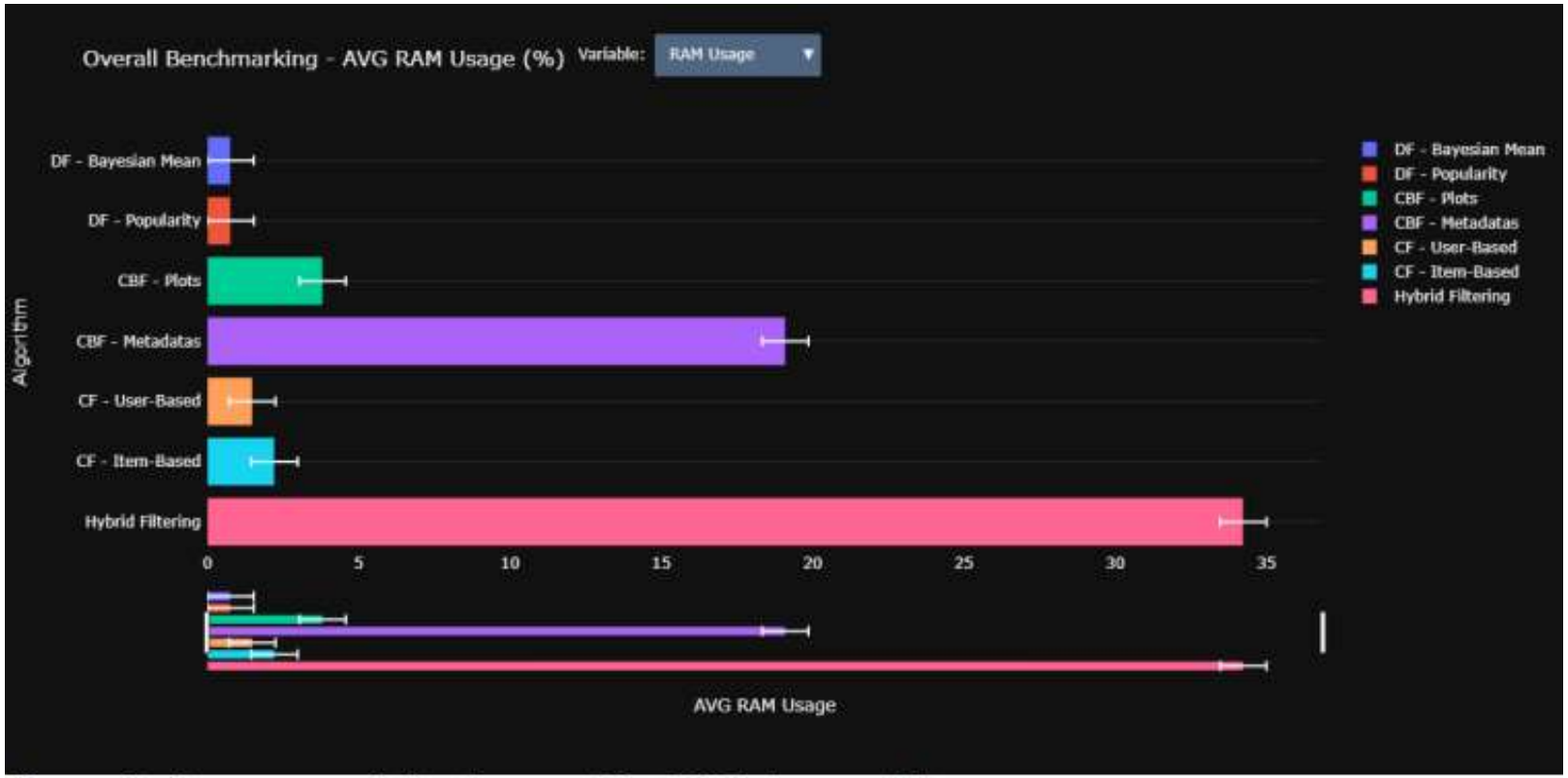


Figura 3: Consumo médio de memória RAM dos modelos

Fonte: Resultados originais da pesquisa

Resultados

1. Performances

- DF - Bayesian Mean: Modelo A
- DF - Popularity: Modelo B
- CBF - Plots: Modelo C
- CBF - Metadatas: Modelo D
- CF - User-Based: Modelo E
- CF - Item-Based: Modelo F
- Hybrid Filtering: Modelo G

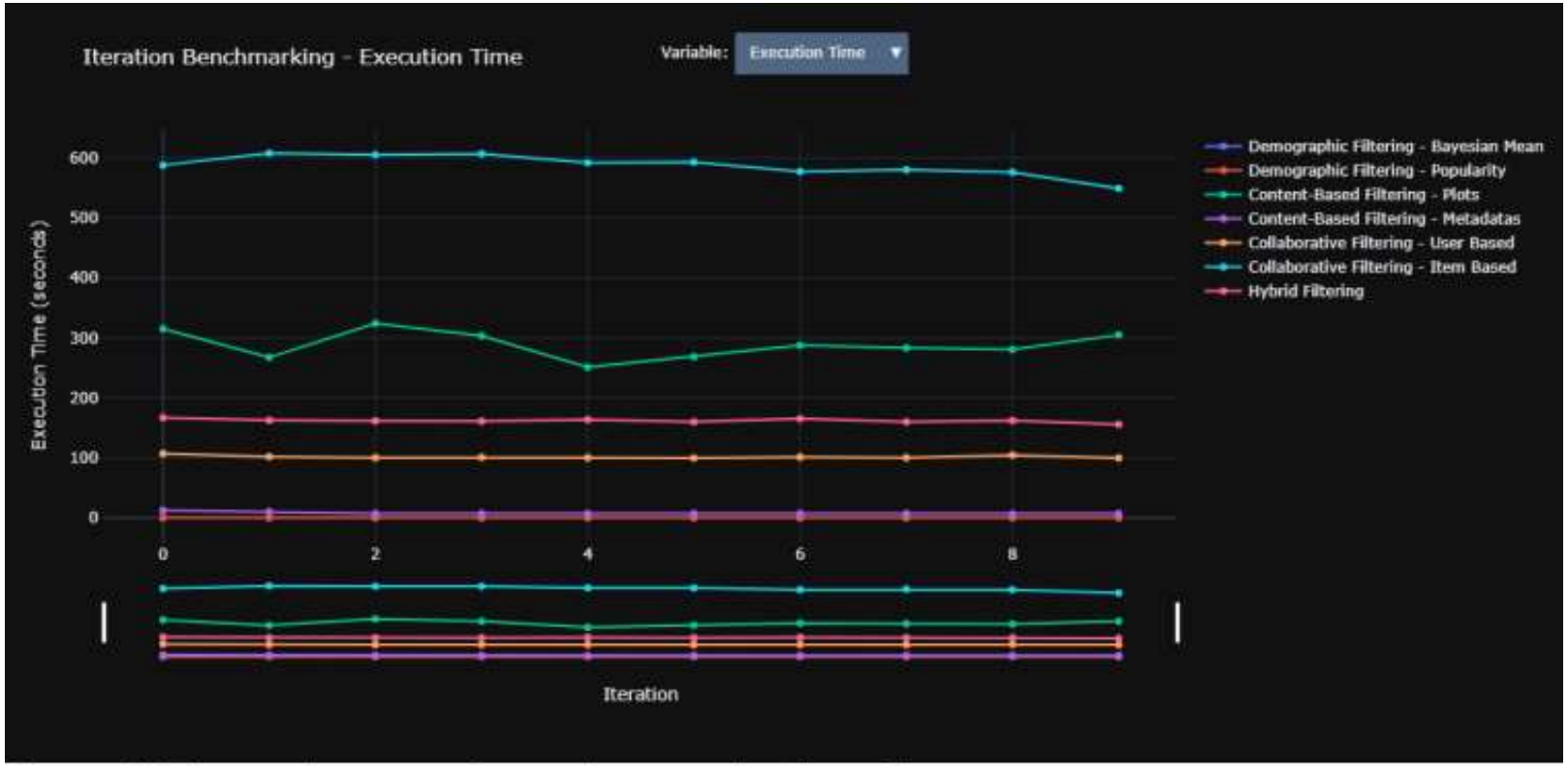


Figura 4: Tempo de execução ao decorrer das iterações

Fonte: Resultados originais da pesquisa

Resultados

1. Performances

- DF - Bayesian Mean: Modelo A
- DF - Popularity: Modelo B
- CBF - Plots: Modelo C
- CBF - Metadatas: Modelo D
- CF - User-Based: Modelo E
- CF - Item-Based: Modelo F
- Hybrid Filtering: Modelo G

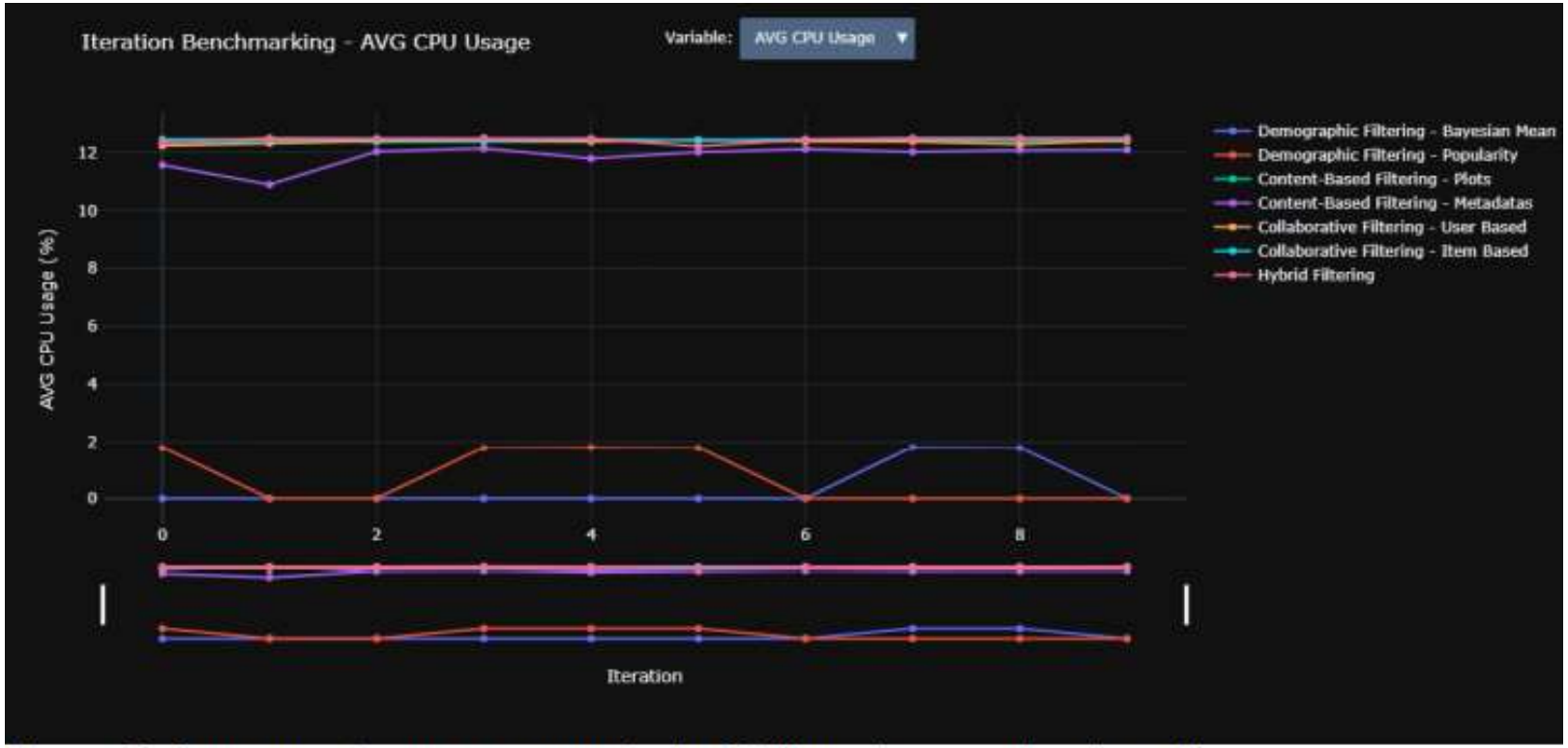


Figura 5: Consumo de processamento de CPU ao decorrer das iterações

Fonte: Resultados originais da pesquisa

Resultados

1. Performances

- DF - Bayesian Mean: Modelo A
- DF - Popularity: Modelo B
- CBF - Plots: Modelo C
- CBF - Metadatas: Modelo D
- CF - User-Based: Modelo E
- CF - Item-Based: Modelo F
- Hybrid Filtering: Modelo G

Figura 6: Consumo de memória RAM ao decorrer das iterações

Fonte: Resultados originais da pesquisa

Resultados

2. Recomendações

Tabela 8. Recomendações realizadas pelo modelo A – Filtragem Demográfica pela Média Bayesiana

Id	Título	Média Bayesiana	Popularidade
5114	Fullmetal Alchemist Brotherhood	9,10	3
41467	Bleach Sennen Kessen-Hen	9,07	464
9253	Steins Gate	9,07	13
28977	Gintama 2	9,06	331
38524	Shingeki no Kyojin Season 3 Part 2	9,05	24

Fonte: Resultados originais da pesquisa

Tabela 9. Recomendações realizadas pelo modelo B – Filtragem Demográfica pela Popularidade

Id	Título	Média Bayesiana	Popularidade
16498	Shingeki no Kyojin Season 1	8,54	1
1535	Death Note	8,62	2
5114	Fullmetal Alchemist Brotherhood	9,10	3
30276	One Punch Man	8,50	4
11757	Sword Art Online	7,20	5

Fonte: Resultados originais da pesquisa

Resultados

2. Recomendações

Tabela 10. Recomendações realizadas pelo modelo C – Filtragem Baseada em Conteúdo pela sinopse considerando “Fullmetal Alchemist Brotherhood” como item de contexto

Id	Título	Similaridade
121	Fullmetal Alchemist	0,5001
10842	Fullmetal Alchemist – The Sacred Star of Milos – Specials	0,2818
430	Fullmetal Alchemist – The Conqueror of Shamballa	0,2635
9135	Fullmetal Alchemist – The Sacred Star of Milos	0,2624
6421	Fullmetal Alchemist Brotherhood – Specials	0,2092

Fonte: Resultados originais da pesquisa

Tabela 11. Recomendações realizadas pelo modelo D – Filtragem Baseada em Conteúdo pelos metadados considerando “Fullmetal Alchemist Brotherhood” como item de contexto

Id	Título	Gêneros	Similaridade
28249	Arslan Senki	Adventure, action, fantasy, drama	1,0000
31821	Arslan Senki Fuujin Ranbu	Adventure, action, fantasy, drama	1,0000
589	Ginga Nagareboshi Gin	Adventure, action, drama	0,9217
2243	Karasu Tengu Kabuto	Adventure, action, drama	0,9217
37521	Vinland Saga	Adventure, action, drama	0,9217

Fonte: Resultados originais da pesquisa

Resultados

2. Recomendações

Tabela 12. Recomendações realizadas pelo modelo E – Filtragem Colaborativa baseada no usuário considerando o usuário com id 609.917 como contexto

Id	Título	Gêneros	Avaliação Predita
4181	Clannad After Story	Supernatural, romance, drama	9,7688
1575	Code Geass Season 1	Action, sci-fi, award winning, drama	9,7496
2001	Tengen Toppa Gurren Lagann	Adventure, action, sci-fi, award winning	9,6618
9989	Anohana	Supernatural, drama	9,6324
11741	Fate Zero Season 2	Action, fantasy, supernatural	9,6239

Fonte: Resultados originais da pesquisa

Tabela 13. Recomendações realizadas pelo modelo F – Filtragem Colaborativa baseada no item considerando o usuário com id 1.129.199 como contexto

Id	Título	Gêneros	Avaliação Predita
20	Naruto	Adventure, action, fantasy	9,6196
269	Bleach	Adventure, action, fantasy	9,5821
1535	Death Note	Suspense, supernatural	9,5687
1575	Code Geass Season 1	Action, sci-fi, award winning, drama	9,4935
2904	Code Geas Season 2	Action, sci-fi, award winning, drama	9,4912

Fonte: Resultados originais da pesquisa

Resultados

3. Limitações

- Performances medidas: Limitadas aos hardwares da máquina do autor
- Recomendações: Limitadas ao universo de animes da plataforma
- Periodicidade: Dados pertencentes até 06 de out. de 2023

Conclusões

1. Complexidade dos algoritmos e acurácia das recomendações

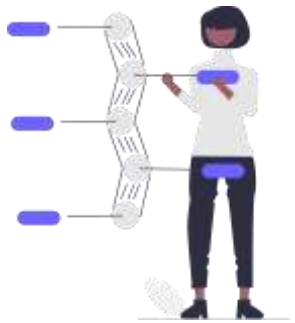
    - Complexidade dos algoritmos e acurácia das recomendações

2. Maior acurácia das recomendações

    - Banco de dados mais vastas e variados

3. Trabalhos futuros

    - Expansão de plataformas
    - Expansão de conteúdos (filmes, séries, músicas...)
    - Aprendizagem Profunda ("Deep Learning")



MBA USP ESALQ

Obrigado!

csfelix.github.io

github.com/CSFelix

linkedin.com/in/csfelix