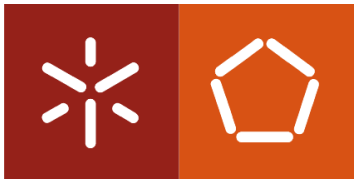


Universidade do Minho
Mestrado em Engenharia Informática
SBS

Trabalho Prático Nº2

Pedro Ribeiro - PG42848

Professor Doutor César Analide Freitas Silva Costa Rodrigues
Professor Doutor Bruno Filipe Martins Fernandes

Mestrado em Engenharia Informática
2021

Conteúdo

Capítulo 1

Introdução

No âmbito na unidade curricular "Sistemas Baseados em Similaridade", do perfil "Machine Learning: Fundamentos e Aplicações"foi desenvolvido este trabalho pratico que se foca na exploração de um conjunto de dados escolhido pelo autor do trabalho pratico.

MyAnimeList é um web site que se assemelha muito a *IMDB.com* porém este como o nome indica foca-se em *animes* e *mangas*, o dataset trabalhado neste projeto é a base de dados do *MyAnimeList*.[2]

Consta com 17562 *anime* e com 325 772 comentários/classificações, é pretendido que o projecto resulte em um sistema de recomendação para os seus utilizadores, o conjunto de dados está presente na plataforma *Kaggle*.[1]

O presente relatório começa fazer uma análise geral aos dados considerados, passando por um exploração dos dados disponíveis e quando necessário o seu devido tratamento; a metodologia utilizada consiste numa análise aprofundada dos métodos em questão, como este foram aplicados no *KNIME* e por fim a respetiva representação dos resultados obtidos.

Capítulo 2

Conjunto de Dados

Foram utilizados dois ficheiros para a resolução deste projeto, estes podem ser observados com mais detalhes na tabela 2

Ficheiro	Definição	N de colunas	N de linhas
anime.csv	Informação sobre o anime	35	17.6m
rating_complete.csv	Contém a classificação dada a um anime de um respectivo utilizador	3	57M

O ficheiro *anime.csv* é constituido pelas seguintes colunas e o seu respectivo tratamento quando aplicado:

- MAL_ID; Name; English name
- Genres; Studios
  transformação em uma lista
- Score; Type TV, Movie, OVA, Special, ONA, Music
  Tratamento de valores não existentes
- Japanese name
  eliminado, não constitutiva um aumento de qualidade para o modelo
- Episodes: numero de episódios.
- Aired: data de estreia.
- Source: Manga, Light novel, Book, etc. (e.g Original)
  Tratamento de valores não existentes
- Rating: Classificação etária (e.g. R - 17+ (violence & profanity))
  Tratamento de valores não existentes
- Ranked: Posição classificativa
  Tratamento de valores não existentes
- Members: numero de utilizadores na comunidade
- Favorites: numero de utilizadores que tem favorito
- Watching: numero de utilizadores assistindo
- Completed: numero de utilizadores completaram
- On-Hold: numero de utilizadores congelaram o anime

- Dropped: numero de utilizadores desistiram de assistir
- Plan to Watch: numero de utilizadores que planeiam em assistir
- Licensors; Duration; Producers; Premiered; Popularity
- 10 colunas com o nome "Score-X"onde X esta compreendido entre 1 e 10
  numero de utilizadores que classificaram como X.

O Segundo conjunto de dados, *rating_complete.csv* apenas consta com 3 colunas, o ID do anime, o ID do utilizador e a respetiva pontuação dada ao anime pelo utilizador, de referir que este conjunto tem uma quantidade de dados muito elevada, constando com 57 milhões de linhas.

2.1 Filtragem: Top-N Não Personalizado

Foram desenvolvidos duas soluções que o seu resultado são das melhores pontuações segundo um critério de exclusividade.

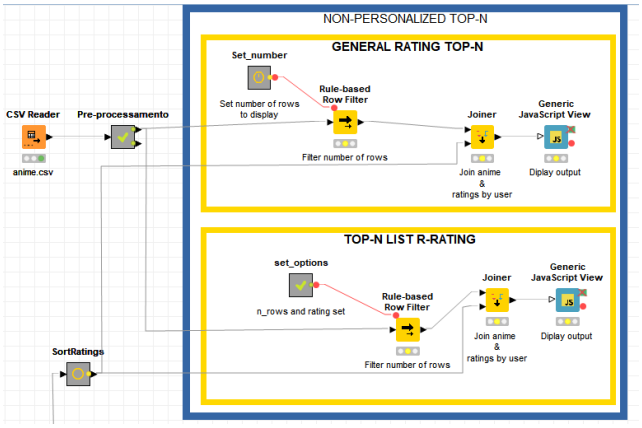


Figura 2.1: Top-N Não Personalizado

O *Genral Rating TOP-N* começa pela declaração do numero que delimita o total de linhas a serem calculadas, de seguida é filtrado segundo esse valor e resulta no top-N segundo a sua pontuação .

Figura 2.2: Numero de total animes no TOP

O *TOP-N List R-Rating* resulta no top de animes segundo a pontuação e segundo a filtragem da Classificação etária escolhida pelo utilizador.

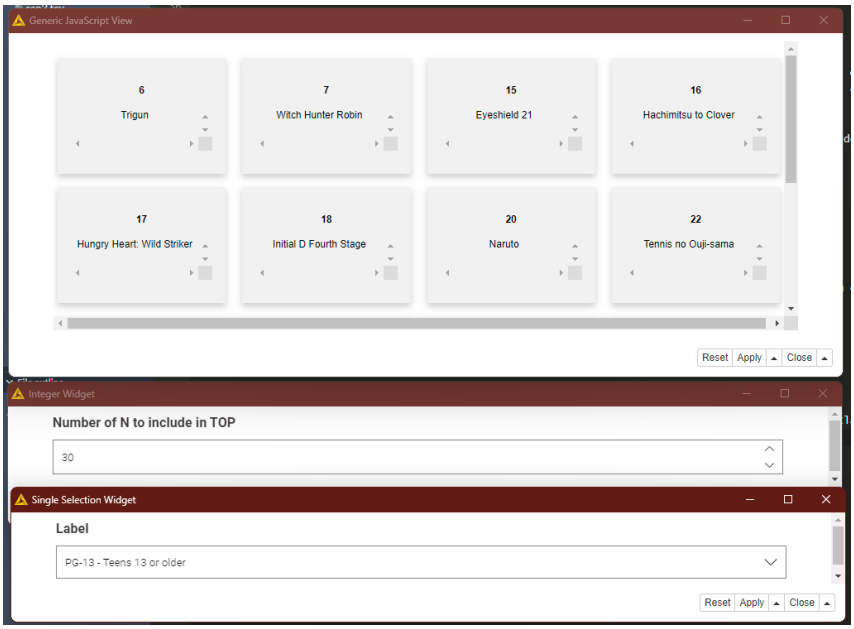
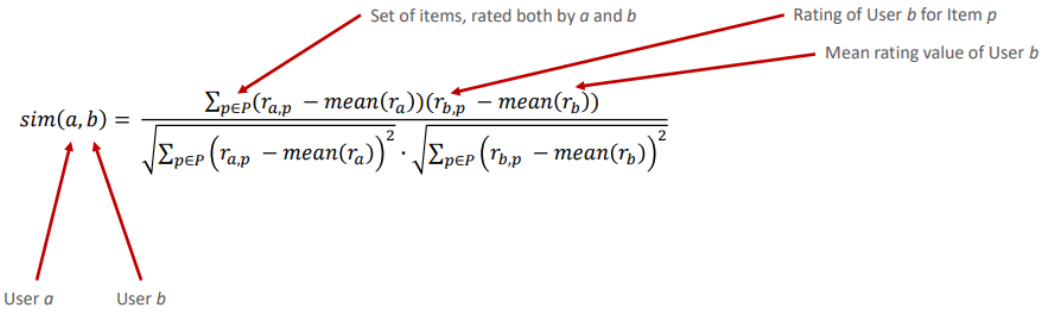


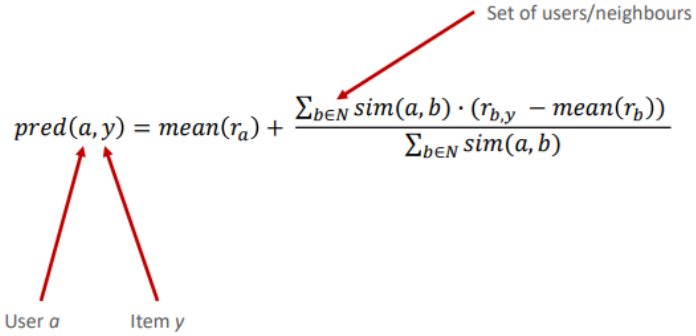
Figura 2.3: TOP-N de Animes segundo Classificação etária

2.2 Memory Based : User-based Nearest-Neighbour

Esta técnica permite uma comparação entre um utilizador e um grupo de vizinhos, os vizinhos são definidos como todos os utilizadores que também classificaram os mesmos animes que o utilizador. Sendo P o conjunto de animes avaliados por ambos os utilizadores a e b (ra,rb). É possivel calcular o coeficiente de correlação Pearson da seguinte forma:



Através da ideia de semelhança, é possível prever a classificação que o cliente daria a um anime que ainda não o avaliou, baseando-se na semelhança do conjunto N de vizinhos que avaliaram; esta ideia pode ser caracterizado da seguinte forma:



2.2.1 WorkFlow desenvolvido

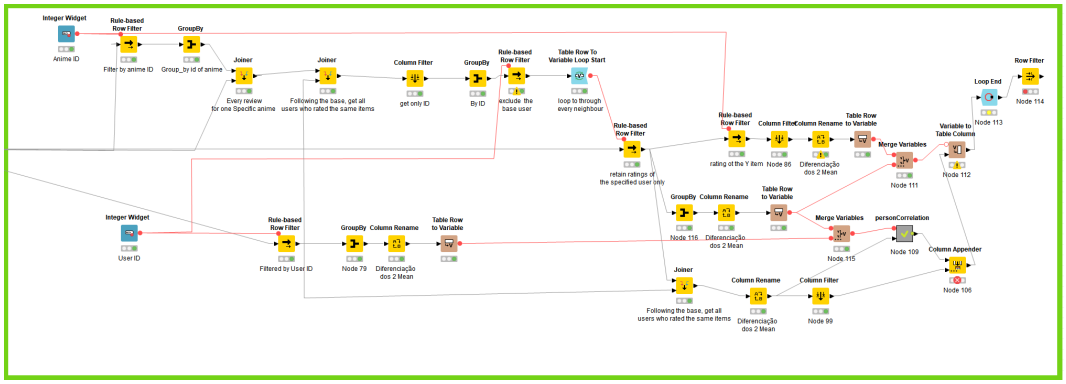


Figura 2.4: WorkFlow do *Memory Based : User-based Nearest-Neighbour*

Capítulo 3

Conclusões

Através da realização de este trabalho pratico o autor aumentou o conhecimento sobre sistemas de recomendações, onde anteriormente não tinha nenhuma experiência neste sector. De salientar a importância da ferramenta KNIME, onde foi possível perceber a tua perseverança em um mundo que as ferramenta são facilmente trocadas por outras; a possibilidade de exploração são extremamente amplas, de referir que o KNIME HUB para além de trazer muita informação e módulos com explicação para alguns problemas mas nem sempre é o mais útil.

O dataset continha muitas colunas o levou a um processamento mais demorado, porém, para escapar aos típicos datasets do IMDB (que são muito utilizados em trabalhos praticos), mesmo este sendo quase o mesmo conceito levou o autor a ter outra visão do problema.

Contudo, o autor considera o sistema desenvolvido insatisfatório em relação ao que era pedido pelos docentes, consta com duas respostas solidas e uma que com um pouco mais de trabalho poderia ser completa e possível ser uma resposta óptima.

De uma forma geral, considera-se que os objectivos académicos de exploração da ferramenta KNIME, estratégias de pré-processamento e novas visões de resolução de problemas foi satisfatório.

Bibliografia

[1] Anime Database
https://www.kaggle.com/hernan4444/anime-recommendation-database-2020

[2] MyAnimeList.met - https://myanimelist.net/