Testing the quality and accuracy of a recommender system

Introduction

The brief was to use an existing dataset and try to improve its quality and accuracy of recommendations using a machine learning approach.

Dataset:
Movie Lens Small Latest Dataset
https://www.kaggle.com/datasets/shubhammehta21/movie-lens-small-latest-dataset

GitHub: https://git.arts.ac.uk/21035961/Term-3---Personalisation-and-Machine-learning---Mini-Project - Student ID - 21035961

1. Aims of the Project

The report will inspect three models of recommendations

1. Content filtering recommendation
2. Collaborative filtering recommendation
3. Hybrid recommendation

To honestly evaluate the accuracy, the same movie dataset needs to be measured in all forms of models available; due to my restrictions on recommender model creation, the below are basic versions. (See Fig1)

2. What are recommender systems

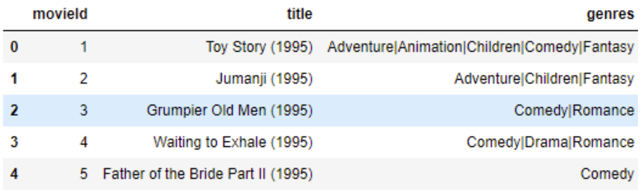
Recommender systems are a form of information filtering used based on predicting a user's preference. Many everyday platforms, such as streaming platforms, use recommender systems. These companies can then provide better or more personalised content, service, and products to the user.

Recommender systems typically produce a list of recommendations through content-based, collaborative filtering, and now, a version that contains both is called a hybrid recommendation.

An example of a recommender system providing accurate results is that of amazon. They use item-item collaborative filtering and produce great results.(*Introduction to recommender systems | by Baptiste Rocca | Towards Data Science*, no date; *What Are Recommendation Systems in Machine Learning? | Analytics Steps*, no date)

3. Outline of methods used (Refer to jupyter notebook)

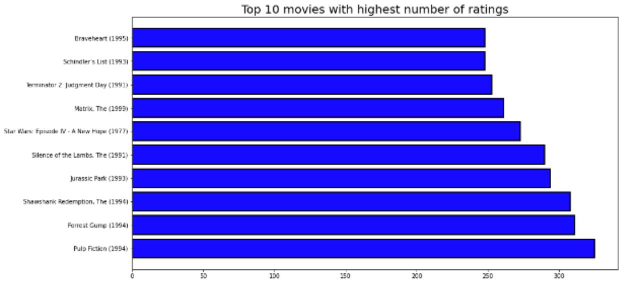
- Load the dataset
- Data Analysis
- Data Cleaning and Merging
- Content Based Filtering
- Collaborative filtering
- A theoretical look into hybrid filtering
- Conclusion and further study



	movieId	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	5	Father of the Bride Part II (1995)	Comedy

Above is a representation of the loaded in dataset.

After the dataset has been loaded and cleaned and all ratings have been categorised correctly, we can then produce a visual representation of the top ten movies are rated the highest. (See Fig 2 below)



After cleaning and merging we can begin to look into our recommender systems

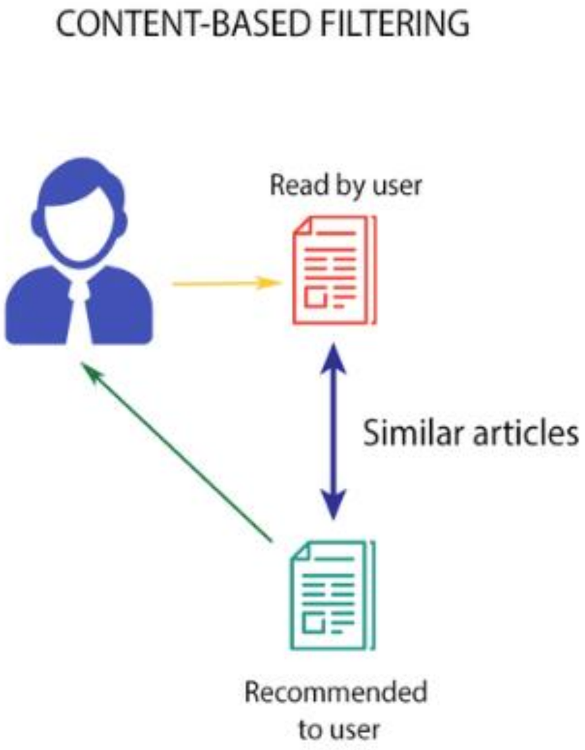
3. Content based filtering

In the industry of recommender systems, many can perform to a significant level of accuracy. However, to understand the level of accuracy needed, we need to compare content vs collaboration since both are the most popular among more prominent companies.

Content-based filtering is more of an insulated recommender system and requires active engagement from every user. Content-based filtering relies on keywords and attributes in the database and matching recommendations based on the user's previous ratings.

(5: *Content based filtering vs Collaborative filtering (Source:... | Download Scientific Diagram*, no date; *How to do a Content-Based Filtering using TF-IDF? | by Ankur Dhuriya | Analytics Vidhya | Medium*, no date;

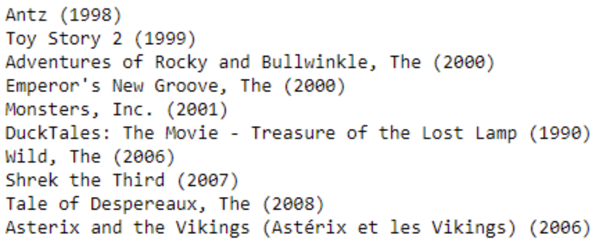
***What is a Content-based Recommendation System in Machine Learning?| Analytics Steps*, no date)**



The cosine similarity is a popular metric. The following step taken in the notebook will then look for similar tf-idf vectors (movies). We've encoded the genre of each film into a tf-idf representation; now, we need to define the proximity measure.

Once we have assigned the dataset to the cosine similarity it produces the following recommender results.

The input: Toy Story (1995)



```
Antz (1998)
Toy Story 2 (1999)
Adventures of Rocky and Bullwinkle, The (2000)
Emperor's New Groove, The (2000)
Monsters, Inc. (2001)
DuckTales: The Movie - Treasure of the Lost Lamp (1990)
Wild, The (2006)
Shrek the Third (2007)
Tale of Despereaux, The (2008)
Asterix and the Vikings (Astérix et les Vikings) (2006)
```

Recommendation breakdown:

- Toy Story Genre - Comedy, Adventure, Fantasy
- Antz Genre - Adventure, Comedy, Fantasy
- Adventures of Rocky and Bullwinkle, The Genre - Comedy, Adventure

As you can see, based on one input that I extracted from the movie’s dataset, the recommendation is pretty good. All the movies are labelled with the above genre. But one insight is that they are also all animated and are traditionally considered to be within the children's film genre.

This information was not provided in the dataset, so there is a good level of accuracy in a content-based model.

However, with content-based filtering, it is easier for the recommendations to be made, there is much more transparency in your suggestions, and you, as the viewer, everything is insulated—your choice. You are in control of the outcome. In contrast, collaborative filtering is precisely that, and you would need to collaborate with many others, including companies, to get newer and even more diverse movie recommendations.

Further exploration: Acquire a larger dataset

Below is an article with great insight into why content-based filtering might not be the right choice for this task, but is an area:

(https://www.upwork.com/resources/what-is-content-based-filtering)

4. Collaborative filtering

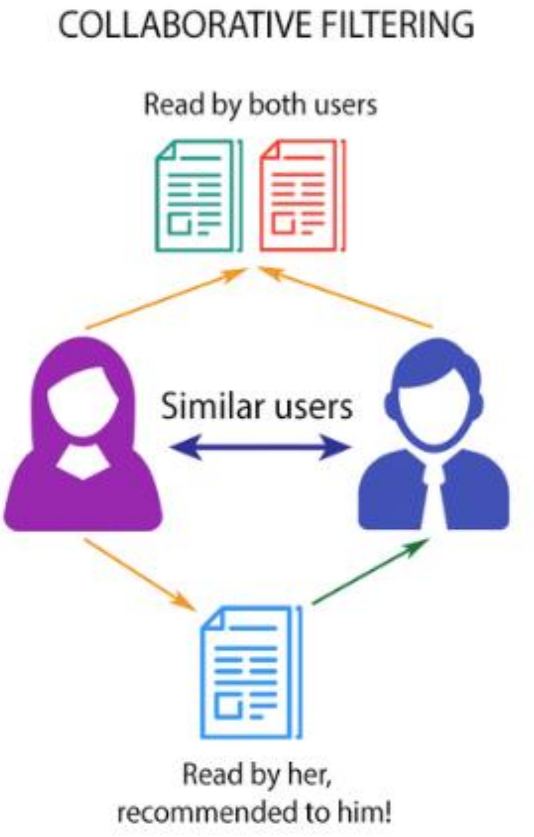
Collaborative filtering is precisely what it means. It is a collaboration and requires many users to provide their input. It will then use a process of matching people's similar interests and make recommendations as a whole to either people within the exact geo-location, same likes and interests and same previously watched content.

In a nutshell, it is about casting a wide net strung together by various people hoping that like-minded individuals will be drawn and caught within it when it is thrown.

Collaborative filtering has one considerable limitation in its approach, and that is, it relies on the user's ratings, and if not, all users are rating movies, it means we will have been given a cold start which is a common problem in the industry. However, major companies have overcome this challenge, which we will discuss later.

(*5: Content based filtering vs Collaborative filtering (Source:... | Download Scientific Diagram*, no date;

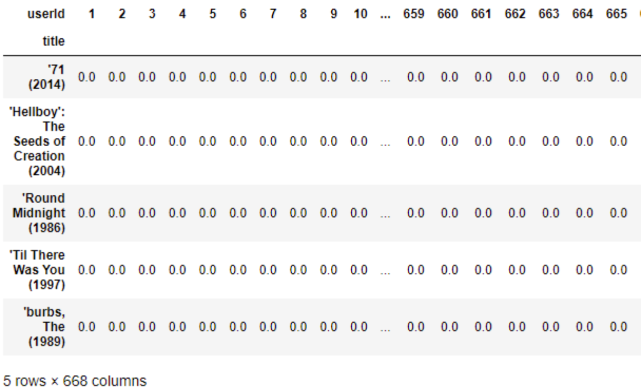
***Limitations of Collaborative Recommender Systems | by Dawn Graham | Towards Data Science*, no date)**



The process in which we create collaborative filtering requires more data manipulation for easier processing.

The dataset is then loaded into a pivot table, The use of pivot tables is generally considered a powerful tool in analysing your data, it does a great job combining and summarising, and the loss of data is minimal. **(*Use of Pivot Table in Pandas – Regenerative*, no date)**

Pivot table example:



userId	1	2	3	4	5	6	7	8	9	10	...	659	660	661	662	663	664	665
title																		
'71 (2014)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
'Hellboy': The Seeds of Creation (2004)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
'Round Midnight (1986)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
'Til There Was You (1997)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
'burbs, The (1989)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0

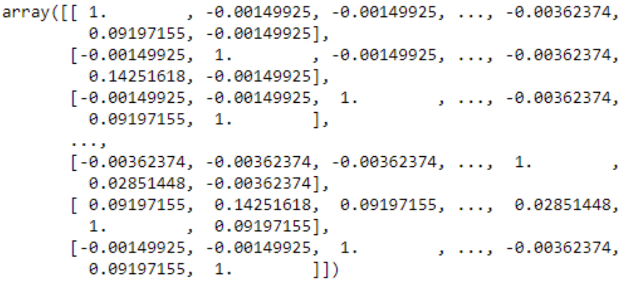
5 rows × 668 columns

We will need to see the correlation/relationship between movies and ratings, and to do this; we will need to use a correlation matrix with the help of our pivot table. It will do this by determining how one variable linearly changes to another.

If a user likes a movie, we can take the column of that movie, find the correlation between that and all other films within that column, and then provide the ones with the strongest correlation.

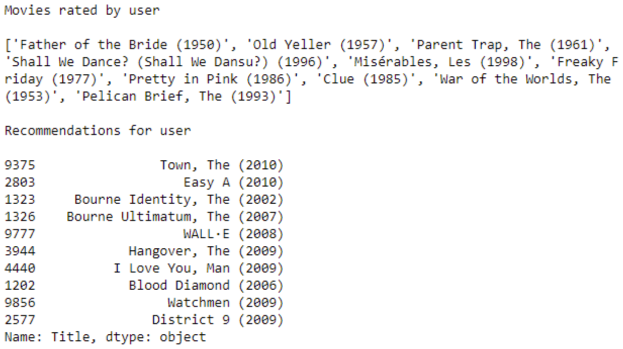
This mathematical method helps will the recommender to accurately churn out recommendations based on the information it has provided. **(*Simple Movie Recommender System with Correlation Coefficient with Python | by Ashwin Prasad | Analytics Vidhya | Medium*, no date)**

Correlation matrix:



```
array([[ 1.        , -0.00149925, -0.00149925, ..., -0.00362374,
         0.09197155, -0.00149925],
       [-0.00149925,  1.        , -0.00149925, ..., -0.00362374,
         0.14251618, -0.00149925],
       [-0.00149925, -0.00149925,  1.        , ..., -0.00362374,
         0.09197155,  1.        ],
       ...,
       [-0.00362374, -0.00362374, -0.00362374, ...,  1.        ,
         0.02851448, -0.00362374],
       [ 0.09197155,  0.14251618,  0.09197155, ...,  0.02851448,
         1.        ,  0.09197155],
       [-0.00149925, -0.00149925,  1.        , ..., -0.00362374,
         0.09197155,  1.        ]])
```

After which the collaborative filtering can be built and we provide it with a prompt, which in this case will be a random users number, to receive the following output.



```
Movies rated by user

['Father of the Bride (1950)', 'Old Yeller (1957)', 'Parent Trap, The (1961)',
'Shall We Dance? (Shall We Dansu?) (1996)', 'Misérables, Les (1998)', 'Freaky F
riday (1977)', 'Pretty in Pink (1986)', 'Clue (1985)', 'War of the Worlds, The
(1953)', 'Pelican Brief, The (1993)']

Recommendations for user

9375                  Town, The (2010)
2803                     Easy A (2010)
1323      Bourne Identity, The (2002)
1326     Bourne Ultimatum, The (2007)
9777                     WALL·E (2008)
3944               Hangover, The (2009)
4440             I Love You, Man (2009)
1202               Blood Diamond (2006)
9856                   Watchmen (2009)
2577                 District 9 (2009)
Name: Title, dtype: object
```

Recommendation breakdown:

The user we have chosen, was at random and below are the recommendations for that user based on the process of collaboration.

The users’ preferences:

- Father of the Bride (1950) Genre - Romance, Comedy, Drama
- Old Yeller (1957) Genre - Drama, Western, Adventure
- Parent Trap, The (1961) Genre - Romance, Comedy, Romantic comedy

Collaborative filtering recommendations:

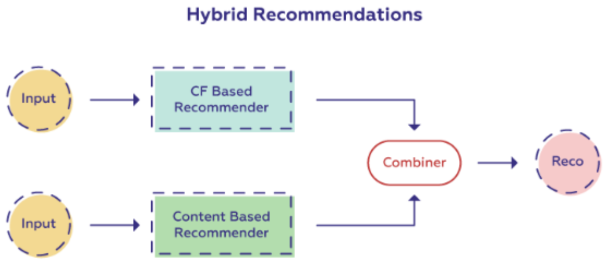
From highest rating to lowest and based on our user's preference.

- Town, The (2010) Genre - Heist, Crime, Drama, Thriller
- Easy A (2010) Genre - Comedy, Drama, Romance, Romantic comedy
- Bourne Identity, The (2002) - Adventure, Action Thriller, Drama

In terms of diversity and novelty, you are provided with a rich panoply of choices. Each recommendation is not like the other but hits the mark in terms of keeping your user engaged and on your platform.

Netflix, YouTube, TikTok and many other video streaming services rely heavily on collaborative filtering recommendations.

5. Hybrid recommendations



(*Inside recommendations: how a recommender system recommends | by Sciforce | Sciforce | Medium*, no date)

Combining collaborative and content-based filtering may help overcome the limitations we face when using them separately, and it may also be more effective in some cases. We can implement hybrid recommender system approaches in various ways, such as generating predictions separately using content and collaborative methods, then combining the forecasts or simply adding collaborative-based method capabilities to a content-based approach (and vice versa).

Several studies compare the performance of conventional approaches to hybrid methods and conclude that using mixed methods results in more accurate recommendations.

(Aksel and Birtürk, no date; *A Guide to Building Hybrid Recommendation Systems for Beginners*, no date; Ghazanfar and Prugel-Bennett, 2010)

6. Conclusion

The methods used to improve quality and accuracy range from data merging and pivoting, cosine similarity, Pearson correlation, TFIDF and a few others to improve the accuracy and quality of the recommenders.

The question of your recommender's quality and accuracy depends on its usage. Companies such as IMDB, Rotten tomatoes, and pandora use content filtering, while others like Amazon, Google, Twitter, LinkedIn, Spotify and much more use collaborative filtering. And the most prominent user and most successful user of hybrid filtering are Netflix, which has invested over $150 million each year to improve functionality.

Using the well-known movie dataset, exploring different recommendation systems has included Collaborative Filtering, Content-Based Filtering and Hybrid recommendation systems(in theory). Conducting a mixed analysis quantitatively and qualitatively comes from the need that content-Based Filtering systems cannot be easily quantifiable.

Also, the qualitative approach holds much importance for a movie recommendation system. That is why the method of evaluation follows the more traditional path.

Further studies:

There are opportunities for further analysis following this work.

For example, demographic-based information about the user could significantly improve the accuracy of recommendations. However, considering this can add another layer of refinement to the hybrid recommendation system.

Also, the Content-Based Filtering recommendation only had a few categories, such as genre. If the dataset included cast, crew, and reviews of the movies, this could provide further similarity. Furthermore, comparing different Collaborative Filtering based methods and similarity measures could be interesting.

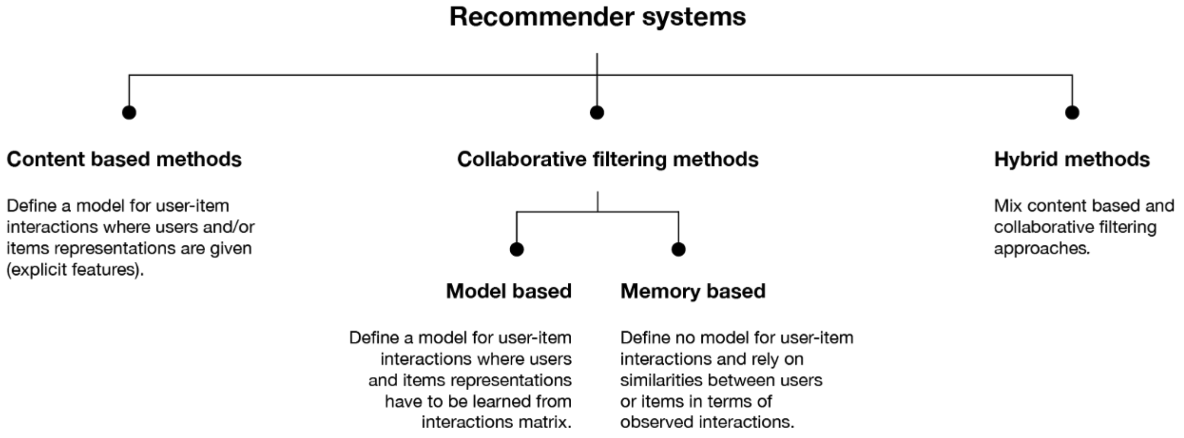
Further reading:

https://www.researchgate.net/publication/346218759_Improving_Accuracy_of_Recommender_Systems_using_Social_Network_Information_and_Longitudinal_Data

https://towardsdatascience.com/4-ways-to-supercharge-your-recommendation-system-aeac34678ce9

https://techcrunch.com/2015/09/28/the-evolving-landscape-of-recommendation-systems/?guccounter=1&guce_referrer=aHR0cHM6Ly93d

3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAASh3FMAeydwxSRbpQ8wM_0CvLQScxzJ5YOUzySl6OlP78lJHyfZ1qHMAR2me1yzno0jQJn2wTcou_g8PNlZAgtz4b6lLdNjxERkg1IGvI0kAhD1h8bbBwlqNxi0XIjHl_h_BGqfrVn3RB1sJPRQPrn8ebg3cdwSq_kWEkGQkfeM



Recommender systems

Content based methods

Define a model for user-item interactions where users and/or items representations are given (explicit features).

Collaborative filtering methods

Model based

Define a model for user-item interactions where users and items representations have to be learned from interactions matrix.

Memory based

Define no model for user-item interactions and rely on similarities between users or items in terms of observed interactions.

Hybrid methods

Mix content based and collaborative filtering approaches.

Figure 1 – (https://towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada)

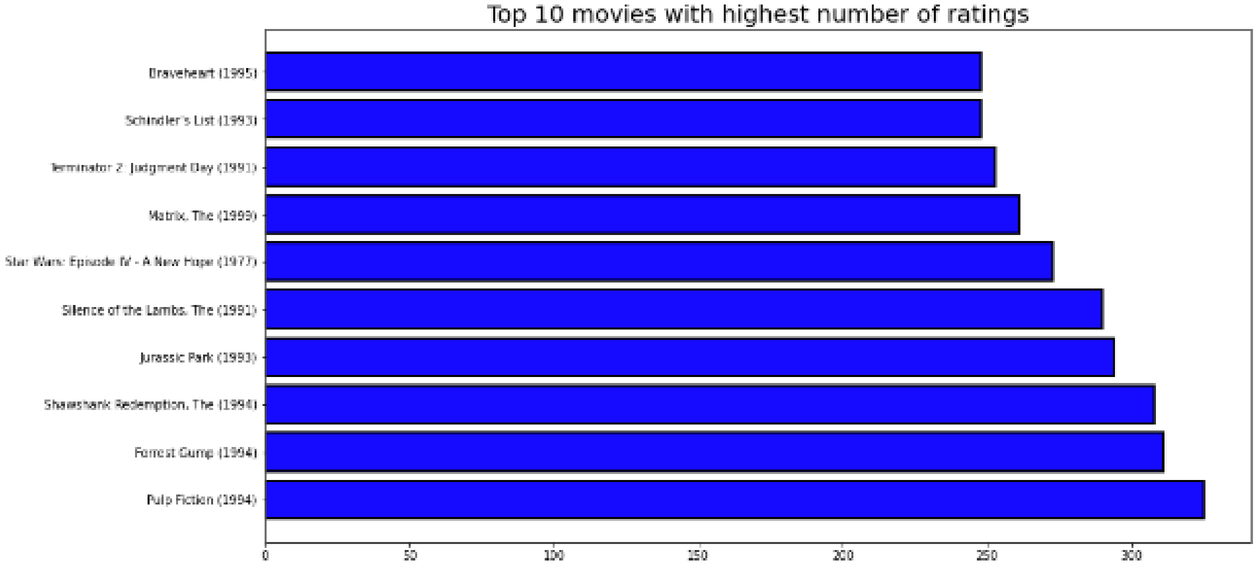


Figure 2 – Data representation of highest rated movies in the given dataset

Bibliography:

5: Content based filtering vs Collaborative filtering (Source:... | Download Scientific Diagram (no date). Available at: https://www.researchgate.net/figure/Content-based-filtering-vs-Collaborative-filtering-Source_fig5_323726564 (Accessed: June 21, 2022).

A Guide to Building Hybrid Recommendation Systems for Beginners (no date). Available at: https://analyticsindiamag.com/a-guide-to-building-hybrid-recommendation-systems-for-beginners/ (Accessed: June 21, 2022).

Aksel, F. and Birtürk, A.A. (no date) "Enhancing Accuracy of Hybrid Recommender Systems through Adapting the Domain Trends." Available at: http://duineframework.org/ (Accessed: June 21, 2022).

Ghazanfar, M.A. and Prugel-Bennett, A. (2010) "A scalable, accurate hybrid recommender system," *3rd International Conference on Knowledge Discovery and Data Mining, WKDD 2010*, pp. 94–98. doi:10.1109/WKDD.2010.117.

How to do a Content-Based Filtering using TF-IDF? | by Ankur Dhuriya | Analytics Vidhya | Medium (no date). Available at: https://medium.com/analytics-vidhya/how-to-do-a-content-based-filtering-using-tf-idf-f623487ed0fd (Accessed: June 21, 2022).

Inside recommendations: how a recommender system recommends | by Sciforce | Sciforce | Medium (no date). Available at: https://medium.com/sciforce/inside-recommendations-how-a-recommender-system-recommends-9afc0458bd8f (Accessed: June 21, 2022).

Introduction to recommender systems | by Baptiste Rocca | Towards Data Science (no date). Available at: https://towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada (Accessed: June 21, 2022).

Limitations of Collaborative Recommender Systems | by Dawn Graham | Towards Data Science (no date). Available at: https://towardsdatascience.com/limitations-of-collaborative-recommender-systems-9801036941b3 (Accessed: June 21, 2022).

Simple Movie Recommender System with Correlation Coefficient with Python | by Ashwin Prasad | Analytics Vidhya | Medium (no date). Available at: https://medium.com/analytics-vidhya/simple-movie-recommender-system-with-correlation-coefficient-with-python-e6cb31dae01e (Accessed: June 21, 2022).

Use of Pivot Table in Pandas – Regenerative (no date). Available at: https://regenerativetoday.com/use-of-pivot-table-in-pandas/ (Accessed: June 21, 2022).

What Are Recommendation Systems in Machine Learning? | Analytics Steps (no date). Available at: https://www.analyticssteps.com/blogs/what-are-recommendation-systems-machine-learning (Accessed: June 21, 2022).

What is a Content-based Recommendation System in Machine Learning?| Analytics Steps (no date). Available at: https://www.analyticssteps.com/blogs/what-content-based-recommendation-system-machine-learning (Accessed: June 21, 2022).