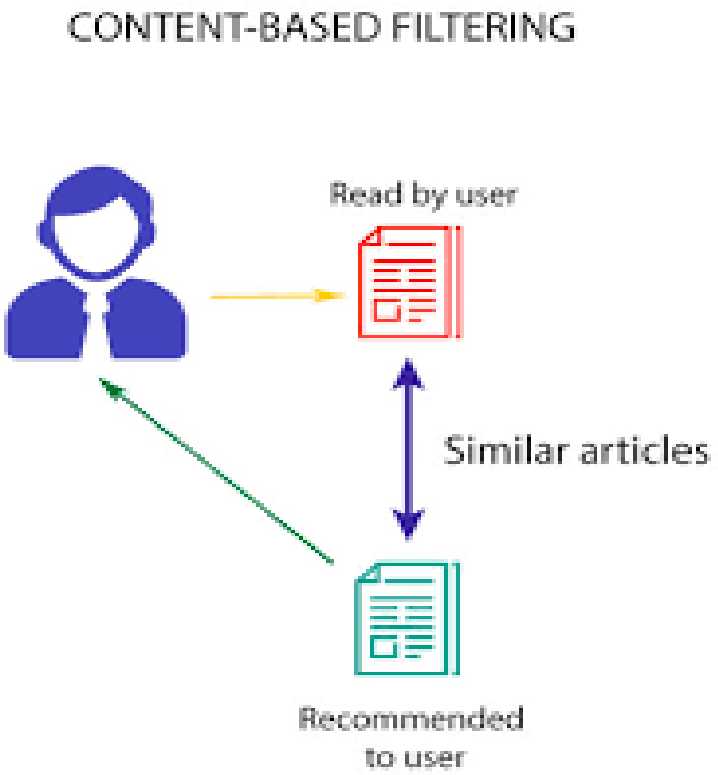
CONTENT BASED RECOMMENDER SYSTEM

Description

- This project as the title suggests focuses on recommending based upon content present in the dataset i.e., the item-item interaction.
- This dataset uses the Netflix dataset and so the items here are the movies and the TV shows. Item-item interaction essentially refers matching users based upon their item history.
- Let's say a user performs an action on items A, B, C and another user performs actions on items A, C. Then automatically the user will be recommended item B by the algorithm.

Content Based Filtering



- Not based upon user ratings and preferences. Instead, solely based upon the context of the movie or the TV show.
- Here we have used the text preprocessing techniques of punctuation marks removal, stripping and removal of stop-words along with making the words in the text lowercase.
- After this use the Tfidf vectorizer as it gives importance to the words based upon their value in the text to convert the text into vectorized format.
- TFIDF Vectorizer stands for term frequency inverse document frequency vectorizer. It is given by the formula:

WORD (W) = (Number of times the word repeats in a sentence/Total number of words in that sentence) *(log (Total number of sentences/Total number of sentences containing that word))

- Then each movie/ TV show will have its context vectors corresponding to the content present in their description.
- Since different words have different importance, it is necessary for us to give less important words less importance.
- Present in scikit learn and no separate downloading required. So, a matrix is formed with columns as the words and rows as the items with row of the item being its corresponding vector.

- Finally use the Nearest Neighbors unsupervised to find the k nearest movies or TV shows based on the cosine similarity.

- Cosine similarity between item vector v1 and v2 is (v1.v2)/(mod(v1) *mod(v2)). Thus, the items closer to each other have more cosine similarity and so less cosine distance which is 1-cosine similarity.

- Item-item interaction is all about comparing between the items and then recommending based upon the item. Content based recommendation uses comparing and finding the relation between items based upon vectors of the description or the reviews.

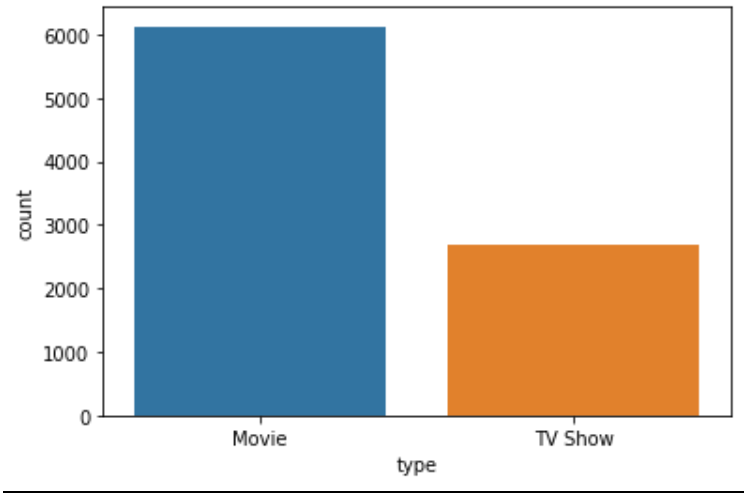
Using the repository (Use the commands given in the Readme file)

1] For fetching of the dataset from Kaggle:

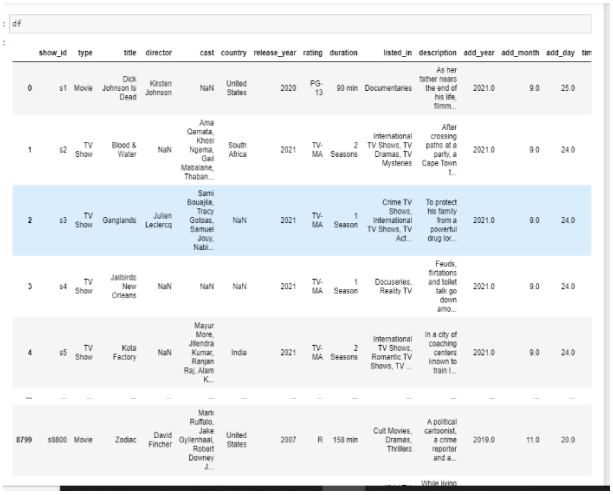
Active Kaggle account is required. Datasets used in the repo are

Netflix Movies and Shows: https://www.kaggle.com/shivamb/netflix-shows
IMDB Movies and Shows: https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset
IMDB Ratings: https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset

- Netflix is one of the most popular media and video streaming platforms. They have over 8000 movies or tv shows available on their platform.
- IMDb is the most popular movie website and it combines movie plot description, Megastore ratings, critic and user ratings and reviews, release dates, and many more aspects.
- Run the tox command mentioned and your datasets will be ready to use. For live data just input the name of the Movie or TV show in string format in the text box.
- After merging the datasets, we can infer those movies are in excess compared to tv shows.



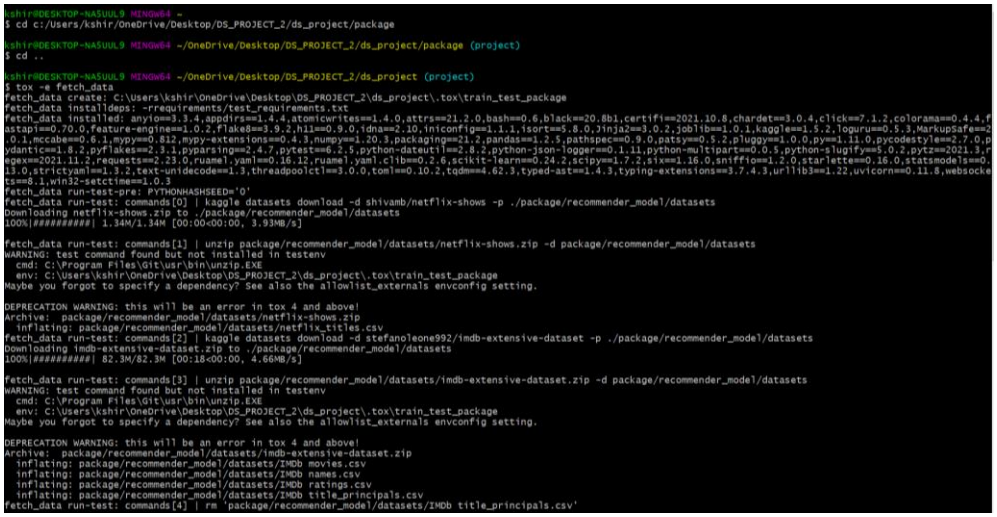
- Final dataset after merging will look like this



df

	show_id	type	title	director	cast	country	release_year	rating	duration	listed_in	description	add_year	add_month	add_day	tim
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...	2021.0	9.0	25.0	
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...	2021.0	9.0	24.0	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...	2021.0	9.0	24.0	
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...	2021.0	9.0	24.0	
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...	2021.0	9.0	24.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
8799	s8800	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	2007	R	158 min	Cult Movies, Dramas, Thrillers	A political cartoonist, a crime reporter and a...	2019.0	11.0	20.0	

- The above dataset uses the Netflix dataset which has important features like the type of content, director, cast, description (extremely important), country of origin, etc.
- On this we have merged the IMDB dataset for ratings and genre of the Netflix content. For live data just input the name of the Movie or TV show in string format in the text box.



2] For training the MODEL:

- First of all, we need to pre-process and clean the data in all the three datasets. Then we need to use the left join and join the ratings and movies dataset on the Netflix dataset using left join.
- After the merging we need to drop the duplicates and the extract the important columns like ‘description’ from the dataset. The final dataset has 8804 entries with 69.6% of them being movies while the rest being TV shows.
- Then Tfidf vectorizer is used with each content vector being of (1,19183) dimensions as the corresponding to the context of the items using the above-mentioned formula for the text separated by spacing. Thus, the data frame comprises of words in columns and items in the index.
- With these content vectors we can then find out the cosine similarity. More is the cosine similarity more likely is the movie or the TV show i.e., the content to be recommended.
- Algorithm used is the Nearest Neighbours which is an unsupervised algorithm with the content being the index of the data frame and the rows being the content vectors. The nearest neighbours taken here are 10 i.e., top 10 recommendations are to be made.

3] For running the API:

- First, we build the API using fastapi module leveraging the python async.io web framework.

- FastApi module is compatible with pydantic module which helps in input data validation and type checking automatically using schemas.
- Then we run this API with the uvicorn ASGI web server on http://localhost:8001/recommendersystem
- Also, we perform logging using loguru module to display the status messages. Live data for recommendation is only the name of the movie or the TV show.

https://drive.google.com/file/d/1V24rJx_ih04pNahCqhbxJcoR5XR--YP_/view?usp=sharing

Refer to the above video.

FOR RUNNING THE REPOSITORY REFER TO THE COMMANDS IN README FILE.