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. Thus, this is a very basic project depicting the recommendation system used in the OTT platforms.
- The models used are NLP-based Tfidf-Vectorizer to convert the words or text to vectorized format and then Nearest Neighbors algorithm is used to find the interaction between the content using cosine similarity.

Advantages of TFIDF Vectorizer

Content based recommendation deals with the content i.e., the description/reviews corresponding to the items in the dataset.

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))

- In the given text not, all words are important in terms of vectorizing.
- Since different words have different importance, it is necessary for us to give less important words less importance.
- So, this vectorizer gives unequal importance to different words of the text. This helps in forming efficient vectors from the given text.
- 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.

Advantages of Nearest Neighbors Algorithm (NNA)

- Unsupervised algorithm to find the cosine similarity between the item vectors. Cosine similarity gives the nearest or closest items pertaining to the current item.
- 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.
- Cosine similarity uses the cosine law between vectors. Less is the cosine similarity more is the cosine distance.
- Thus, NNA helps us in finding the best k items pertaining to the current item all based upon cosine similarity between the items.

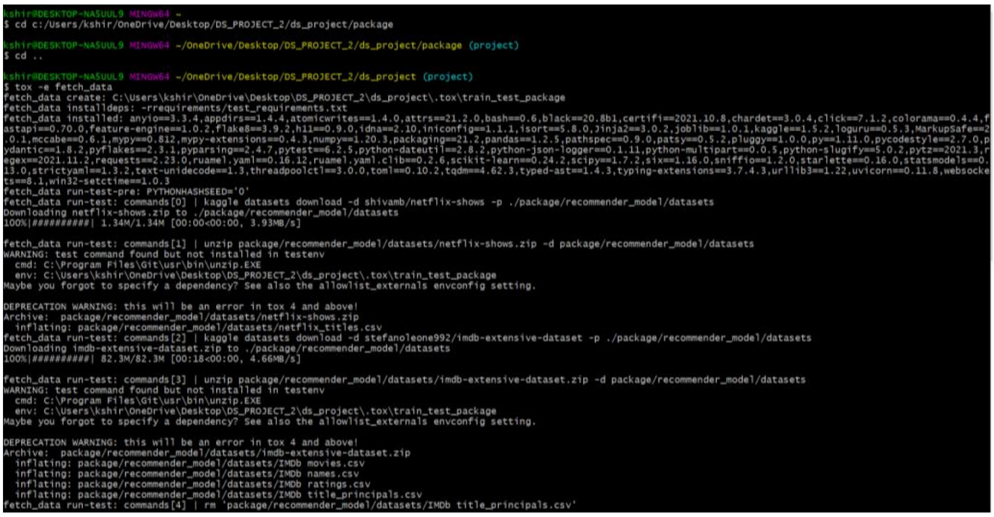
Using the repository (use the commands given in the dataset section of Readme File)

1] Fetching 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

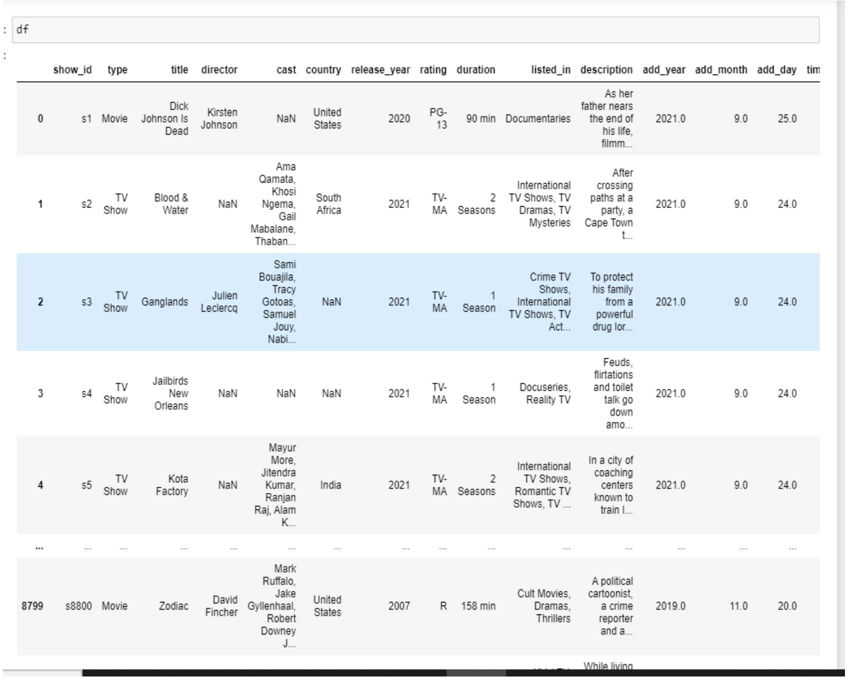
- Netflix is one of the most popular media and video streaming platforms. They have over 8000 movies or tv shows available on their platform.
- IMDb is the most popular movie website and it combines movie plot description, Megastore ratings, critic and user ratings and reviews, release dates, and many 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.



```
kshir@DESKTOP-NA5UUL9 MINGW64 ~
$ cd c:/Users/kshir/OneDrive/Desktop/DS_PROJECT_2/ds_project/package

kshir@DESKTOP-NA5UUL9 MINGW64 ~/OneDrive/Desktop/DS_PROJECT_2/ds_project/package (project)
$ cd ..

kshir@DESKTOP-NA5UUL9 MINGW64 ~/OneDrive/Desktop/DS_PROJECT_2/ds_project (project)
$ tox -e fetch_data
fetch_data create: C:\Users\kshir\OneDrive\Desktop\DS_PROJECT_2\ds_project\.tox\train_test_package
fetch_data installdeps: -rrequirements/test_requirements.txt
fetch_data installed: anyio==3.3.4,appdirs==1.4.4,atomicwrites==1.4.0,attrs==21.2.0,bash==0.6,black==20.8b1,certifi==2021.10.8,chardet==3.0.4,click==7.1.2,colorama==0.4.4,f
astapi==0.70.0,feature-engine==1.0.2,flake8==3.9.2,h11==0.9.0,idna==2.10,iniconfig==1.1.1,isort==5.8.0,Jinja2==3.0.2,joblib==1.0.1,kaggle==1.5.2,loguru==0.5.3,MarkupSafe==2
.0.1,mccabe==0.6.1,mypy==0.812,mypy-extensions==0.4.3,numpy==1.20.3,packaging==21.2,pandas==1.2.5,pathspec==0.9.0,patsy==0.5.2,pluggy==1.0.0,py==1.11.0,pycodestyle==2.7.0,p
ydantic==1.8.2,pyflakes==2.3.1,pyparsing==2.4.7,pytest==6.2.5,python-dateutil==2.8.2,python-json-logger==0.1.11,python-multipart==0.0.5,python-slugify==5.0.2,pytz==2021.3,r
egex==2021.11.2,requests==2.23.0,ruamel.yaml==0.16.12,ruamel.yaml.clib==0.2.6,scikit-learn==0.24.2,scipy==1.7.2,six==1.16.0,sniffio==1.2.0,starlette==0.16.0,statsmodels==0.
13.0,strictyaml==1.3.2,text-unidecode==1.3,threadpoolctl==3.0.0,toml==0.10.2,tqdm==4.62.3,typed-ast==1.4.3,typing-extensions==3.7.4.3,urllib3==1.22,uvicorn==0.11.8,websocke
ts==8.1,win32-setctime==1.0.3
fetch_data run-test-pre: PYTHONHASHSEED='0'
fetch_data run-test: commands[0] | kaggle datasets download -d shivamb/netflix-shows -p ./package/recommender_model/datasets
Downloading netflix-shows.zip to ./package/recommender_model/datasets
100%|##########| 1.34M/1.34M [00:00<00:00, 3.93MB/s]

fetch_data run-test: commands[1] | unzip package/recommender_model/datasets/netflix-shows.zip -d package/recommender_model/datasets
WARNING: test command found but not installed in testenv
  cmd: C:\Program Files\Git\usr\bin\unzip.EXE
  env: C:\Users\kshir\OneDrive\Desktop\DS_PROJECT_2\ds_project\.tox\train_test_package
Maybe you forgot to specify a dependency? See also the allowlist_externals envconfig setting.

DEPRECATION WARNING: this will be an error in tox 4 and above!
Archive:  package/recommender_model/datasets/netflix-shows.zip
  inflating: package/recommender_model/datasets/netflix_titles.csv
fetch_data run-test: commands[2] | kaggle datasets download -d stefanoleone992/imdb-extensive-dataset -p ./package/recommender_model/datasets
Downloading imdb-extensive-dataset.zip to ./package/recommender_model/datasets
100%|##########| 82.3M/82.3M [00:18<00:00, 4.66MB/s]

fetch_data run-test: commands[3] | unzip package/recommender_model/datasets/imdb-extensive-dataset.zip -d package/recommender_model/datasets
WARNING: test command found but not installed in testenv
  cmd: C:\Program Files\Git\usr\bin\unzip.EXE
  env: C:\Users\kshir\OneDrive\Desktop\DS_PROJECT_2\ds_project\.tox\train_test_package
Maybe you forgot to specify a dependency? See also the allowlist_externals envconfig setting.

DEPRECATION WARNING: this will be an error in tox 4 and above!
Archive:  package/recommender_model/datasets/imdb-extensive-dataset.zip
  inflating: package/recommender_model/datasets/IMDb movies.csv
  inflating: package/recommender_model/datasets/IMDb names.csv
  inflating: package/recommender_model/datasets/IMDb ratings.csv
  inflating: package/recommender_model/datasets/IMDb title_principals.csv
fetch_data run-test: commands[4] | rm 'package/recommender_model/datasets/IMDb title_principals.csv'
```

2] Training of the models:

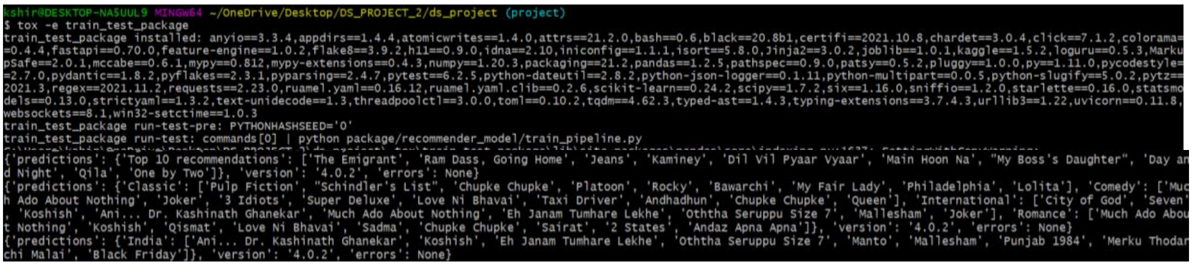
- There are three datasets and so we need to pre-process and clean them individually. After they are in proper format then we have to merge the movies and ratings datasets on the Netflix dataset to obtain a complete dataset corresponding to the content.



```
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	

- Then vectorize the required column(s) using the Tfidf Vectorizer and then use the vectors to find the cosine similarity between them. Lastly apply the Nearest Neighbour Algorithm to find the cosine similarity between the vectors.
- As we can see here the results because of content i.e., taking only description into consideration are poor. If we were to take other features into consideration like directors, cast, etc. we would have to use Count Vectorizer by giving space between those features and not between text of a single feature.



```
kshir@DESKTOP-NA5UUL9 MINGW64 ~/OneDrive/Desktop/DS_PROJECT_2/ds_project (project)
$ tox -e train_test_package
train_test_package installed: anyio==3.3.4,appdirs==1.4.4,atomicwrites==1.4.0,attrs==21.2.0,bash==0.6,black==20.8b1,certifi==2021.10.8,chardet==3.0.4,click==7.1.2,colorama=
=0.4.4,fastapi==0.70.0,feature-engine==1.0.2,flake8==3.9.2,h11==0.9.0,idna==2.10,iniconfig==1.1.1,isort==5.8.0,Jinja2==3.0.2,joblib==1.0.1,kaggle==1.5.2,loguru==0.5.3,Marku
pSafe==2.0.1,mccabe==0.6.1,mypy==0.812,mypy-extensions==0.4.3,numpy==1.20.3,packaging==21.2,pandas==1.2.5,pathspec==0.9.0,patsy==0.5.2,pluggy==1.0.0,py==1.11.0,pycodestyle=
=2.7.0,pydantic==1.8.2,pyflakes==2.3.1,pyparsing==2.4.7,pytest==6.2.5,python-dateutil==2.8.2,python-json-logger==0.1.11,python-multipart==0.0.5,python-slugify==5.0.2,pytz==
2021.3,regex==2021.11.2,requests==2.23.0,ruamel.yaml==0.16.12,ruamel.yaml.clib==0.2.6,scikit-learn==0.24.2,scipy==1.7.2,six==1.16.0,sniffio==1.2.0,starlette==0.16.0,statsmo
dels==0.13.0,strictyaml==1.3.2,text-unidecode==1.3,threadpoolctl==3.0.0,toml==0.10.2,tqdm==4.62.3,typed-ast==1.4.3,typing-extensions==3.7.4.3,urllib3==1.22,uvicorn==0.11.8,
websockets==8.1,win32-setctime==1.0.3
train_test_package run-test-pre: PYTHONHASHSEED='0'
train_test_package run-test: commands[0] | python package/recommender_model/train_pipeline.py
C:\Users\kshir\OneDrive\Desktop\DS_PROJECT_2\ds_project\.tox\train_test_package\lib\site-packages\pandas\core\indexing.py:1637: SettingWithCopyWarning:
{'predictions': {'Top 10 recommendations': ['The Emigrant', 'Ram Dass, Going Home', 'Jeans', 'Kaminey', 'Dil Vil Pyaar Vyaar', 'Main Hoon Na', "My Boss's Daughter", 'Day an
d Night', 'Qila', 'One by Two']}, 'version': '4.0.2', 'errors': None}
{'predictions': {'Classic': ['Pulp Fiction', "Schindler's List", 'Chupke Chupke', 'Platoon', 'Rocky', 'Bawarchi', 'My Fair Lady', 'Philadelphia', 'Lolita'], 'Comedy': ['Muc
h Ado About Nothing', 'Joker', '3 Idiots', 'Super Deluxe', 'Love Ni Bhavai', 'Taxi Driver', 'Andhadhun', 'Chupke Chupke', 'Queen'], 'International': ['City of God', 'Seven'
, 'Koshish', 'Ani... Dr. Kashinath Ghanekar', 'Much Ado About Nothing', 'Eh Janam Tumhare Lekhe', 'Oththa Seruppu Size 7', 'Mallesham', 'Joker'], 'Romance': ['Much Ado Abou
t Nothing', 'Koshish', 'Qismat', 'Love Ni Bhavai', 'Sadma', 'Chupke Chupke', 'Sairat', '2 States', 'Andaz Apna Apna']}, 'version': '4.0.2', 'errors': None}
{'predictions': {'India': ['Ani... Dr. Kashinath Ghanekar', 'Koshish', 'Eh Janam Tumhare Lekhe', 'Oththa Seruppu Size 7', 'Manto', 'Mallesham', 'Punjab 1984', 'Merku Thodar
chi Malai', 'Black Friday']}, 'version': '4.0.2', 'errors': None}
```

3] For running of the API

- First, we build the API using fastapi module leveraging the python async.io web framework.
- Then we run this API with the uvicorn ASGI web server.
- Also, we perform logging using loguru module.

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

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