Research Project (Netflix Big Data Analysis & Recomendation/Filtering System)

Carlos Gutierrez PID: 6248381

CNT 4147 - Big Data Analytics for ECE

Professor Dr. Jayesh Soni

Demonstration of the project:

https://youtu.be/mYV3ArXXipQ

GitHub Repository (data, notebooks and Python Scripts used):

https://github.com/carlosgutierrez1412/CNT4147.git

I. Abstract

In this project I will be using the concepts and tools learned in CNT 4147, specifically Apache Hadoop and Apache Spark in order to collect, process, clean, analize and deploy two Netflix Movies and Shows open datasets; a titles dataset that includes the following features :

Field	Description
id	The title ID on JustWatch.
title	The name of the title.
show_type	TV show or movie.
description	A brief description.
release_year	The release year.
age_certification	The age certification.
runtime	The length of the episode (SHOW) or movie.
genres	A list of genres.
production_countries	A list of countries that produced the title.
seasons	Number of seasons if it's a SHOW.
imdb_id	The title ID on IMDB.
imdb_score	Score on IMDB.
imdb_votes	Votes on IMDB.
tmdb_popularity	Popularity on TMDB.
tmdb_score	Score on TMDB.

and a credits dataset with the following features:

Field	Description
person_ID	The person ID on JustWatch.
id	The title ID on JustWatch.
name	The actor or director's name.
character_name	The character name.
role	ACTOR or DIRECTOR.
production_countries	A list of countries that produced the title.
seasons	Number of seasons if it's a SHOW.
imdb_id	The title ID on IMDB.

this, with the ultimate purpose, to create a GUI where the user can interact with two different systems: a filtering system which will take different parameters selected by the user and return all the different Netflix Movies or Shows that match those specific parameters, and a content recommendation system where the user will input the name of a Netflix Movie or Show and the model will return the top 3 most similar titles to the one the user selected as well as a similarity score provided by the use of a cosine similarity approach.

II. Languages, Tools and Techniques

> Languages: For this project I mainly used python for the data analysis and the GUI deployment, but CMD was also used in order to communicate with Apache Hadoop HDFS and Apache Spark

> Tools and Techniques:

1- **Apache Hadoop**: In this project, I utilized the Apache Hadoop framework, specifically the Hadoop Distributed File System (HDFS), to efficiently store and manage my two datasets across a distributed cluster environment which also gave me the possibility to seamlessly interact with Apache Spark for my analysis.

2- **Apache Spark:** For this project I specifically used Pyspark in order to perform all the necessary data collecting, cleaning and analysis utilizing techniques like creating a SparkSession, loading my datasets from HDFS, grouping by certain features like the names in the credits dataset to have a unique cast with all members of a title, inner join: to combine both datasets into one by using their common ID keys, converted certain features from categorical to numerical data, displaying statistical graphs like correlation matrices and boxplots in order to explore the relations between features and look for outliers and finally using pyspark.ml to build my content recommendation model using encoders, vector assemblers, scalers and cosine similarity.

3- **Jupyter Notebooks:** Used as my main web application to perform the data analysis.

4- **Pycharm IDE:** Used as primary IDE to develop, deploy and test my GUI as well as importing the recommendation model from Jupyter.

5- **Tkinter:** In order to create the GUI I used the Tkinter python library, in order to create the main pages, labels and buttons as well as output boxes to display the results from bith the Filter System and Recomendation System.

6- **Pandas and Scikit-Learn:** In order to use tkinter to create the GUI I had to convert my program from Pyspark in a cluster to Pandas locally and perform loading and filtering of my data as well as using Scikit-Learn to develop the recommendation model again so it would be compatible with the GUI.

III. Data Source

The data source used for this project is called "Netflix TV Shows and Movies" and it is a open public dataset in kaggle.com that contains a list of all shows and movies available on Netflix Streaming up to July 2022, the datasource offers two main datasets:

1- A titles dataset which contains more than 5000 titles on Netflix

2- A credits dataset with more than 77000 credits of actors and directors on Netflix

IV. Results

The results of this project is a fully function GUI that incentivizes the user to use two different functionalities made possible by the analysis and processing of the datasource. One called "Filtering System" where the user can search up an already processed and cleaned dataset using filters like: Cast (name of a desired cast member), Type (movie or show), minimum release year, age certification, genre (name of desired genres), minimum IMDB and TMDB score, minimum and maximum number of seasons, and production country.The other functionality called "Recomendation Model" where the user can type the name of a Netflix movie or show and using the cosine similarity model, the program will display the top 4 titles that are most similar to the one the user provided.

V. References

1- *Netflix TV shows and movies*. (2022,July,26).Kaggle.https://www.kaggle.com/datasets/victorsoeiro/netflix-tv-shows-and-movies

2- Murthy, A. K. (2017). Big Data Analysis Using Hadoop and Spark (Master's thesis). Retrieved from https://digitalcommons.memphis.edu/etd/1700

3- *Module 4 Assignment*. (n.d.). FIU Canvas. https://fiu.instructure.com/courses/185727/assignments/2537381?module_item_id=8097032

4- *How to upload a file to HDFS and download a file from HDFS* -. (2022, December 22). ProjectPro. https://www.projectpro.io/recipes/upload-file-hdfs-and-download-file-from-hdfs

5- *PySpark Overview — PySpark master documentation*. (n.d.). https://spark.apache.org/docs/latest/api/python/index.html

6- freeCodeCamp.org. (2021, July 14). *PySpark Tutorial* [Video]. YouTube. https://www.youtube.com/watch?v=_C8kWso4ne4

7- *Graphical User Interfaces with Tk*. (n.d.). Python Documentation. https://docs.python.org/3/library/tk.html

8- OpenAI. (n.d.). GPT-3.5 [Computer software]. https://openai.com/research/gpt-3