CDS 492: Deliverable 1

* **Name**: Avantika Prativadhi
* **Title**: CineMatch: A Content-Based Recommendation System for Movies and TV Shows
* **Research Question:** Can a content-based recommendation system utilizing features such as actors, directions, plot, number of seasons, age certification, production country, and genre, along with the popularity and score data from TMDB and IMDb, accurately predict a user’s rating of a movie or a TV Show.
* **Motivation:** Recommendation systems are widely used in many industries today. From e-commerce to social media, and to various entertainment platforms, companies implement these systems to improve user experience and increase the engagement on the content.
* **Importance:** I heavily rely on recommendation algorithms on platforms like Spotify (for music), LinkedIn (for job posts), and Netflix or Hulu (for TV shows and movies) to discover new content. While the algorithm is very impressive sometimes, there have also been times where the show I see in the “you might like this” tab is nowhere close to anything I would like to watch. This grew my curiosity on how these systems are made. I wanted to use this opportunity to learn more about the systems while also adding a new project to my portfolio. I’ve always been interested in joining the entertainment field and I think that having a project like this will be very beneficial.
* **Hope to Learn:** Aside from gaining more insight on the recommendation systems, I think this will be challenging and rewarding task where I can learn more about applying the various analysis methods and algorithms on multiple datasets. While I have briefly worked on k-means clustering in the past, it was done on R and not Python. I hope to gain familiarity with other clustering methods and get a deeper understanding on them. I also plan on using various similarity measures and feature training on the variables that I have never done before. Overall, this project is a learning experience while also applying various techniques that I’ve learnt in the past.
* **Data:** The different datasets that I will be using can be found on Kaggle and was scrapped from the streaming guide called “JustWatch”. The collaborator is Victor Soeiro. The first dataset is the Hulu data. This has two csv files: credits and titles. The other dataset is Amazon Prime data with the two csv files as well.
* **Metadata**: The three datasets shown below were combined into one major data frame called df.
  + **Hulu dataset**: There are 20 columns with 31,802 entries. The titles data frame has information regarding the movie like the title, description, release year, runtime, genre, etc.… and the credits file has information on the name of the actor/ director and their role in the show/movie. Both these csv files were combined into one with a new column called “service” added which specifies which of the three platforms (Prime, Hulu, Netflix). The type of variables can be seen in Figure 1.

A picture containing text, plaque, screenshot

Description automatically generated

Figure 1 Hulu Dataset Info

* + **Prime dataset**: There are 20 columns with 124,347 entries. The type of the variables can be seen below in Figure 2.

Graphical user interface, text

Description automatically generated with medium confidence

Figure 2 Prime Dataset Info

* + **Netflix dataset:** There are 20 columns with 1,217 entries. The type of the variables can be seen below in figure 3.

A picture containing text, plaque

Description automatically generated

Figure 3 Netflix Dataset Info

* **Method of Analysis:** Statistical Modeling and Machine Learning
  + **Feature Engineering**: After the process of cleaning the data and fixing any discrepancies, the plan is to transform and preprocess the features in my dataset. For the text feature extraction, I plan on using Bag-of-words approach. Next, for the genre, age, production country, and director/actor variables, I will create a dummy variable for each of the different types.
  + **Similarity measures**: To calculate the similarity between the features of the movies/shows and recommend the ones with the highest score, I will use cosine similarity and Euclidean distance.
  + **Clustering**: To group shows/movies by similar features, I’ll use k-means clustering. This is what I will use for the actual recommendation system.
  + **Matrix Factorization**: Non-negative matrix factorization (NMF) to factorize matrix of movie or TV show features and their overall ratings.
* **Software:** For my Project, I will be using Python to conduct the analysis and build the model. Initially, I also used Tableau to create some simple plots. But I was able to recreate similar ones on python that I will be using. The IDE where I will be writing the code is Jupyter Notebook. All the code written on here will be pushed onto a GitHub repository which contains all the relevant files including the datasets.
* **Workflow Diagram:**

Timeline

Description automatically generated

* **Progress:** So far, I have selected the datasets, combined all three of them, cleaned & transformed them, and conducted exploratory data analysis. I have also begun the feature engineering section where I have created the dummy variables. In the next week, I plan to focus on finishing the similarity measures and clustering sections. I should be done with a majority of the project by the end of march. Then, the final stage of the project will be the matrix formation and creating a user-friendly interface that will output the results. Below are all the plots generated while exploring the data.

A picture containing calendar

Description automatically generated Chart, pie chart

Description automatically generated

Figure 4 Word Cloud of common words in summary Figure 5 Distribution of shows and movies.

Chart

Description automatically generated Chart

Description automatically generated

Figure 6 Movie and Shows Distribution by Genre. Figure 7 Distribution of Movies and Shows by Production Country

Chart

Description automatically generated with medium confidence

Figure 8 Distribution of Ratings based on Age Certification Grouped by Platforms

* **References:** 
  + Choi, S.-M., Han, Y.-S.: A content recommendation system based on category correlations. In: The Fifth International Multi-Conference on Computing in the Global Information Technology, pp. 1257–1260 (2010)
  + D. Das, H. T. Chidananda and L. Sahoo, "Personalized Movie Recommendation System Using Twitter Data" in Progress in Computing Analytics and Networking, Singapore:Springer, vol. 710, 2018.
  + SK. Ko et al., "A Smart Movie Recommendation System", *Lecture Notes in Computer Science*, vol. 6771, 2011.
  + N. Mishra, S. Chaturvedi, V. Mishra, R. Srivastava and P. Bargah, "Solving Sparsity Problem in Rating-Based Movie Recommendation System" in Computational Intelligence in Data Mining, Singapore:Springer, vol. 556, 2017.