**K-Drama Recommender System**

Kankana Ghosh

**Introduction:**

The Kdrama Recommender System project aimed to design and implement a data-driven solution for suggesting TV series to users based on their preferences. This report outlines the methodology, data preprocessing, model selection, evaluation, and results of the project. This project leverages cosine similarity and TF-IDF vectorization to provide personalized recommendations.

**Methodologies:**

This project uses content-based filtering techniques, i.e., it suggests Kdrama to users based on the features or content of the series given as input by the user. This approach focuses on understanding the attributes, properties, or characteristics of Kdrama and then making recommendations that align with the user's preferences, as inferred from them. The dataset contains 250 observations on 18 variables. The dataset typically contains South-Korean series. Content-Rating, Director, Screenwriter and Production company have missing values. After removing the missing values the analysis is conducted on the new dataset containing 242 rows on 18 variables. The boxplot for Year of Release, Number of Episodes and Ratings have outliers present in them . From the histogram of ‘Year of Release’ it is seen that the distribution is negatively skewed which is obvious. The production of series started increasing from the year 2017. The histogram for Number of Episodes also exhibit a positively skewed distribution with highest frequency of Kdrama have 16 number of episodes . From the histogram of Rating given to series we see that it is positively skewed i.e., only few series gets higher rating and it's evident. Maximum series have rating 8.3. From the barplot of Original Network, tvN has highest frequency for series release. The heatmap shows a moderate negative correlation between Number of Episodes and Year of Release which means that older series had more number of episodes. Applied TF-IDF Vectorizer to convert text data into numerical form. Firstly, combined all the relevant text columns like Synopsis, Genre, Tags, Director, Screenwriter, Cast and Production companies into a single text column. Then used this vectorization technique to convert text data into numerical features. Implemented the cosine similarity to get the similarity matrix over the vectorized column. As it is a content based filtering technique so I asked the user to input his/her choice and the number of recommendations that he/she want. Got the index of the series in the dataset, the similarity scores for the series with other series then sorted the Kdrama based on similarity scores in descending order. Hence got the recommendations as desired as by user based on his/her choice of Kdrama.

**Conclusion:**

While this algorithm has done a decent job of finding series with similar features of the series given by the user but the quality of recommendations is not that great. Content-based filtering tends to recommend series that are closely related to a user's past interactions. This could result in over-specialization, where users receive recommendations that are too narrow in scope and don't expose them to a variety of content.