**Machine Learning Assignment-2**

**Analysis of Movie Recommendations with content dataset**

A Content-Based Recommender works by the data that we take from the user, either explicitly (rating) or implicitly (clicking on a link). By the data we create a user profile, which is then used to suggest to the user, as the user provides more input or take more actions on the recommendation, the engine becomes more accurate.

# ****User Profile:****

In the User Profile, we create vectors that describe the user’s preference. In the creation of a user profile, we use the utility matrix which describes the relationship between user and item. With this information, the best estimate we can make regarding which item user likes, is some aggregation of the profiles of those items.

# ****Item Profile:****

In Content-Based Recommender, we must build a profile for each item, which will represent the important characteristics of that item.  
For example, if we make a movie as an item then its actors, director, release year and genre are the most significant features of the movie. We can also add its rating from the IMDB (Internet Movie Database) in the Item Profile.

# Procedure:

We can use the cosine distance between the vectors of the item and the user to determine its preference to the user.

For explaining this, let us consider an example:  
 We observe that the vector for a user will have a positive number for actors that tend to appear in movies the user likes and negative numbers for actors user doesn’t like, Consider a movie with actors which user likes and only a few actors which user doesn’t like, then the cosine angle between the user’s and movie’s vectors will be a large positive fraction. Thus, the angle will be close to 0, therefore a small cosine distance between the vectors.

It represents that the user tends to like the movie, if the cosine distance is large, then we tend to avoid the item from the recommendation.

Dataset:

The final data used here consist of 4803 rows and 23 columns.

* credits.csv :
  + Related to user ratings.
  + Consisting 4083 rows and 4 columns
  + Columns : movie\_id,title,cast,crew
* movies.csv :
  + Related to movies characteristics
  + Consisting 4803 rows and 20 columns
  + Columns-- budget,genres,homepage,id,keywords,original\_language,original\_title,overview,popularity,production\_companies,production\_countries,release\_date,revenue,runtime,spoken\_languages,status,tagline,title,vote\_average,vote\_count

# Conclusion:

It is concluded that this model doesn't need any data about other users, since the recommendations are specific to this user. This makes it easier to scale to a large number of users. The model can capture the specific interests of a user, and can recommend niche items that very few other users are interested in.