**Progress report**

So far we have worked on gathering all the necessary data for our project. This proved to provide an initial challenge as some of the datasets were large and hard to work with. We gathered data from the following sources: Netflix viewing history of Gijs, IMDB open data, and we scraped a table from finder.com which provides us with a current list of all available movies in Netflix in France. We were successful in creating a small dataset of roughly 2000 movies. It is still being blocked by us not being able to match the entries of the 3 datasets on the movie title alone. We are looking into ways to improve this. We believe we could have at least 50% more movies in our dataset if we do that successfully. We will include the latest dataset we have with this report.

Our next challenge is in labelling our dataset for the watchability of the movies. As my Netflix viewing history does not contain a rating on the watched movies or the degree at which movies are made available to me through the platform. The current dataset represent only watched movies without ambiguity. This is, however not the case for unwatched movies. Unwatched movies can represent the following: movies that I would not watch, movies I haven’t seen yet but would watch or anywhere in between. This requires some way to identify availability which can then be related to a watched vs unwatched label. For this reason we have thought of a couple of ways to create the availability label:

1. The longer ago the movie has been released the better its availability is. Movies released longer than 6 months ago start gaining availability at a rate of 1 every year up to 10 years. This availability rating can then be used to create a decision boundary based on the watched movie average availability. The assumption this analysis is based on is that the longer a movie has been released the lower the chance is of it not being watched by me.
2. Using unsupervised clustering we can determine the most watched clusters and assume all movies in these clusters still have a chance of being watched. All other movies will be considered unwatchable. This requires a lot of experimentation to build an appropriate size of clusters. Another aspect which makes this very challenging is the categorical features in our dataset. We are looking to use k-modes which was introduced by Zhehue Huang in 1998 for this.[1]
3. A random sample is chosen from the unwatched movies that is the same size as the watched movies. This will be used to train a predictor of the Boolean label; watched/unwatched. This will the prediction algorithm will be chosen to allow for qualitative analysis on the prediction, like SVM etc.

The abovementioned methods will serve as the first part of our algorithm. Now we have unambiguous labels we can train a classification algorithm for a binary class. Another option is presented to us by some of the abovementioned methods, namely a regression methods, which aims at getting a quality rating.

We would really appreciate your input on the three methods proposed above as only two of them can be considered ML and they have varying degrees of complexity and effectiveness. Can you comment on which of these methods has the best feasibility of getting usable results?

**References:**

1. Huang, Z. Data Mining and Knowledge Discovery (1998) 2: 283. https://doi.org/10.1023/A:1009769707641