**Part Datacleaning and Visualisation: State of affairs** *(18.05.20)*

**A) Written Introduction:**

Not written yet, postponed to beginning of June.

**B) Data cleaning and visualization:**

*Data cleaning:*  First round is accomplished. Ready to review if wishes or errors occur while implementing ML algorithms.

*Remarks*

Some entries (e.g. keywords) seem to be JSON. However, working with JSON-packages didn’t work (I have invested a lot of time to try it). I applied and extended an approach I have seen in the internet which separates entries to columns, processes the entries and combines them again. This part was quite time consuming. However, I guess we will get points for “individual coding” here.

Spoken\_languages: I haven’t found a satisfying way to process the data yet. Entries are not completely clean but I think nonetheless, they can be used for coding algorithms. I will try to improve this, if we really use this category.

########### Hint: How the strings can be filtered with regards to a specific word:

install.packages('stringr')

library(string)

x <- which(str\_detect(genre, "Adventure"))

movie[x,]

I haven’t revised the data cleaning part completely and nicely (layout, etc.) as I suggest that I have to make adaptions when I receive feedback from the coding team :-). Maybe some columns need to be added or reviewed.

*Credits-dataset:* On my laptop, it takes very long to process the data, at least if my approach of data processing is applied. Maye we better drop this part. That’s also the reasons why I didn’t apply it to the other dataset. Furthermore, I couldn’t clean it up completely yet. Maybe somebody has an idea how to select actors faster.

*To-Dos:*

* improve commenting
* improve layout

Further comments, see code.

*Visualization:* Some graphs are implemented. Will be further elaborated at the beginning of June. I guess this part is independent of the part in which ML algorithms are implemented.