Project Proposal.

Data Mining & Machine Learning 2019.

Team Orange:

Blanck, Constant
Oswald, Cyrill
Tsareva, Svetlana
Losey, Sylvain

For the semester project we have decided to choose a topic of sentiment analysis for movie reviews from the movie database IMDb. We want to write a program that indicates if a movie review is positive (user liked the movie) or negative (user didn't like the movie). For this we will explore more thoroughly how we can apply the text analysis and classification techniques we have studied in class. The criteria we will use to determine if the review is indeed positive or negative is the rating given by the user: If the rating is equal or greater than 7, that indicates that the user liked the movie. If it's equal to 4 or smaller he didn't like it. We will exclude neutral reviews (with ratings between 4 and 6) from analysis since we only want to differentiate between positive and negative reviews

In terms of the dataset we have chosen to use a dataset from Stanford which includes 50'000 different movie reviews from IMDb, you can find it here: http://ai.stanford.edu/~amaas/data/sentiment/ . The reviews are highly polarized, which should help us to have a more reliable model.

Thus, we would use the dataset in order to perform a sentiment classification task using either or both '*Bag of Words'* or '*Term Frequency-Inverse Document Frequency model (TFIDF)*' methods. Finally, a Logistic Regression model will be constructed to predict the classification of ratings for the movies based on their reviews.