Capstone Project 2
"Sentiment analysis of Tweets"

Proposal

- **The problem I want to solve**

I would like to perform a sentiment analysis on Tweets about the late blockbuster release: *Joker*, with Joaquin Phoenix.

- **My potential client and why he cares about this problem**

Out in the USA beginning of October 2019 and throughout theaters in the world later on, no doubt that *Joker* has divided its audience. Whether people love or hate it, the reactions were numerous. I saw it myself at the movies in Switzerland as soon as it came out, and felt exactly this way: divided. The movie received a Golden Lion win at the Venice Film Festival, I had high expectations. The complex personality of the Joker was one of the reasons I wanted to see the movie. But after visioning it, I realized it had a strong impact on me, more than expected and not only positive. The critics I read afterwards were also two-folds, both positive and negative. Therefore, I thought it would be interesting to train a classification model on Tweets on the topic of *Joker*.

Different types of people could be interested to know the proportion of positive vs negative tweets on this topic: fans of Joaquin Phoenix and of *The Dark Knight Rises*, owners of movies, producers, script writers, psychologists and psychiatrists.

- **The data I will be using and how I will acquire it**

I will be using the Twitter API to collect a Test set based on keywords. A function will return a list of tweets that contain our keywords selected. Each tweet's text will see itself attributed a label ('positive' or 'negative') to classify each tweet as positive or negative. The Training set will be downloaded because it has to be labelled into 'positive' or 'negative' on a big amount of tweets. The Training set is critical to the success of the model since our model will "learn" how to do create a sentiment analysis based on the Training set.

- **How I will solve this problem**

The steps that we will follow to perform the sentiment analysis are:

- acquiring a Test set
- acquiring and preparing a Training set
- pre-processing tweets in both Test set and Training set (remove punctuation, tokenize)
- build a vocabulary/list of words in our training data set
- match tweet content against our vocabulary
- build our word feature vector
- training the classifier
- testing the model

- **My deliverables**

A jupyter notebook containing the code, a slide deck , and a final report in PDF.