Youtube : Fame predictor

Benoît Steinmann, Cyrille Rolland, Tanguy Rossel

Introduction

This document shortly presents our semester project for the course Network Tour of Data Science. It should involve the basic data science processing pipeline, i.e. raw data collecting, data processing, data cleaning, models & algorithms and data product. Our goal is to be able to predict if a Youtube video gets famous (gets a lot of views) in function of his title, image and tags.

1 Data acquisition

To minimize the quantity of videos required to get interesting results we decided to focus on the french-speaking videos.
We will use the Youtube API to get information about the videos on the platform. The collected information is : title, tags, thumbnail, number of views and number of likes/dislikes. We will also get information about the channel : name, number of total views on the videos and number of subscribers.

2 Data exploration

We will list the most and least popular videos and the most and least popular channels of our data set. We will use a histogram to see the distribution of the number of videos we collected for different view count intervals. The pattern should be symmetric. We will also plot the ratio likes/dislikes of the videos in function of the number of views to see if we can find a pattern.

3 Data exploitation

To train our neural network we will define classes for different views ranges. For instance we will use a class for 0 to 99 views, one for 100 to 999, etc. The number of classes and there ranges are to be determined. The training data will be : the video name, the video image, the video tags, the number of subscribers of the channels hosting the video.

4 Evaluation

A test set will be extracted from our data set to evaluate the precision of our algorithm.