Project Plan

1. Team Member

- Nicolas Lassaux

2. Project Description

A stock evolution is linked to the general feeling about one company or similar companies. It can be interesting to know before a seance if potential buyers trust enough the company or to forecast the fear of actual stock owners. Twitter can permit us to get this global sentiment evolution. By analysing tweets one day earlier than the stock value, I would like to estimate the evolution of the next day.

This kind of analysis is already used in high frequency trading, by banks to prevents their customers and are largely studied. I could't see any plain description about the state of the art. I would like to explore what is feasable at this scale, and try to understand the limits of this model. The difficulty is that independant variables can't be find in a file, ready to be used. Its a real challenging subject where I will have to find my own variables.

3. Description of the Data

I would like to use the data both from Twitter and from Yahoo Finance :

- https://dev.twitter.com/rest/public
- http://finance.yahoo.com/q/hp?s=%5EIXIC+Historical+Prices

I can gather some popular and relevant tweets about a company with the Twitter API. For each tweet, I can obtain : - polarity (positive, negative, neutral) - emotion (joy, anger, sadness, etc.)

I will have to congregate these values to obtain a chart for each value that summerize sentiments evolution about a stock.

Yahoo Finance exposes an API and CSV files with historical values of a stock :

- Date : possibility to go get more than 1 year in past
- Open
- High
- Low
- Close
- Volume
- Adj Close

I would like to estimate as output the evolution of the next day : (Close - Open) / Open

The input would be at least a feeling polarity sum for many tweets (sum of positive values, sum of negative values), and also about the sum for emotions (number of tweets carrying joy, anger...).

4. Methodology

I have already experienced such data-mining jobs, I will focus for this applied-statistics course on the analysis after gathering data. For sentiment analysis I will use the R package "sentiment" or a Python equivalent.

I will focus on several stocks. Firstly, I will have to determine some "key users", like financial actors or newspapers. For each stock I will collect relevant tweets from these users. Next I will compute input data and output data with some simple operations (sum, divide).

I will have to analyse the role of my different variables to select a model. I could have to insert new data or to change followed users if all variables are irrelevant. If some days my model has a very bad accuracy, I could focus on these days to understand the reason.

5. Project Management

As I'm alone, I will do everything by myself.