|  |  |  |  |
| --- | --- | --- | --- |
| **Step** | **Work** | **Theory** | **Done** |
| **Step 1a: X variables** [week 2] | * Fetch the data using the **twitter API** or any other API or website that you are interested in. * Beware of the rate limits and organize your program so as to overcome them if needed * The data should include some text, but might also have other interesting variables (retweets, favorites...) * Create some (non-text based) X\_num variables that you will use for the prediction | W2 | OK |
| **Step 1b: continuous y variables** [week 2] | * Fetching the data:   + if you work on the suggested idea, you can easily access daily stock prices using the [yfinance package](https://pypi.org/project/yfinance/) (see below)   + otherwise, you can find some interesting data listed in the syllabus | W2 | OK |
| **Step 1c: merge X\_num and y** [week 2] | * Beware of the temporality: in the case of the proposed study on stock market prices, you will have to deal with the fact that the X is at the tweet level while y is daily. | W2 | OK |
| ***Milestone 1 - March 24th*** | | | |
| **Step 2a: estimate different regression models using X\_num and y** [week 3] | | W3 | OK |
| ***Milestone 2 - March 31th*** | | | |
| **Step 3: text analysis** [week 4] | * Featurize tweets (or another text dataset related to your subject): transform the text into a standard document-level dataset X\_doc | W4 | OK |
| **Step 2b: estimate different regression models using X\_doc and y** [week 3] | | W3 | OK |
| **Step 4: estimate a classification models** [week 5] | * propose a categorical variable y\_calc that you can compute from the continuous one (y) (e.g. positive or negative growthin stock prices). For the X dimension, you can use X\_doc or X\_num or both. * you can use any other categorical variable that you find relevant | W5 | OK |
| ***Milestone 3 - April 21th*** | | | |
| **Step 6: Dimension reduction** [week 6] | * Use one of the dimension reductions methods to dimension-reduce the features   + PCA or topic model (LDA or STM) or k-means clustering on the featurized text X\_doc * Run another classifier | W6 | OK |
| **Milestone 4 - May 17th** | | | |