# Challenge 3 – Sentiment Analysis

You are given a table with filtered *stocktwits.com* data such that it only contains tweets with tagged stock symbols. This table has been run through HANA based entity extraction. What this basically does is identify keywords in tweets and amongst other identifies sentiments in them (like “good” or “bad”). What you are supposed to do in the following is take advantage of these extracted sentiments and perform an analysis comparing them to the actual stock prices. To that purpose you will calculate correlation coefficients, which are floats between 0 and 1, indicating whether or not there is a correlation between a symbol’s stock prices and tweets about it.

### TASK 1

* The first thing to do is to for each symbol extract all tweets about that symbol[[1]](#footnote-1) including their sentiments.
* Next you need to normalize all time stamps to full hours. Once done, aggregate all stock data per hour and subsequently compute the percentage difference for each hour with respect to the previous one. ( )
* Now aggregate (sum) the sentiments per hour (**NOTE**: Since the stock market is only active from 09:30 to 15:59, include the sentiments from 16:00-09:29 and other days with no stock activity, like weekends and holidays in value of the next “09:00”-hour). The possible sentiments and their corresponding values are as follows:

|  |  |
| --- | --- |
| StrongPositiveSentiment | +3 |
| STOCK\_BUY | +2 |
| WeakPositiveSentiment | +1 |
| NeutralSentiment | ±0 |
| WeakNegativeSentiment | -1 |
| STOCK\_SELL | -2 |
| StrongNegativeSentiment | -3 |

Compile your results so far in a table as follows:  
(**NOTE:** If for any symbol no *stocktwits.com* data is available for a certain hour for which stock data is available skip that hour in your results instead of assigning a sentiment value of 0 to it)

|  |  |  |  |
| --- | --- | --- | --- |
| Symbol | Hour | AggregateSentimentValue | StockDeltaPercent |
| BBY | 2008-07-22 15:00 | +5 | -1.47 |
| BBY | 2008-07-23 09:00 | +4 | +3.03 |
| BBY | 2008-07-23 10:00 | -12 | -5.15 |
| BBY | 2008-07-23 11:00 | +8 | -7.22 |
| BBY | 2008-07-23 12:00 | +7 | +3.39 |

* The last step is to calculate the correlation coefficient for each symbol as follows:

To explain the calculation let’s use the values from the table above but round the StockDeltaPercent (denote B) to integers to simplify calculations (**NOTE**: Don’t do this in your calculations!). In the table below denote AggregateSentimentValues as A

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Hour | A | B | norm(A) | norm(B) | corrper\_hour(BBY) |
| 2008-07-22 15:00 | +5 | -1 |  |  | 0.75 |
| 2008-07-23 09:00 | +4 | +3 |  |  | 0.8 |
| 2008-07-23 10:00 | -12 | -5 |  |  | 0.8 |
| 2008-07-23 11:00 | +8 | -7 |  |  | 0 |
| 2008-07-23 12:00 | +7 | +3 |  |  | 0.95 |

Where and

Finally what remains is to calculate the correlation coefficient for BB. Given hours:

1. You can find all symbols in the SYMBOLS table. And you find all tweets talking about a symbol by finding all entries with the field TYPE = ‘SOCIAL\_MEDIA/TOPIC\_TWITTER’ and the corresponding symbol in the #-notation in the TOKEN [↑](#footnote-ref-1)