High Level Documentation

Module 1 : Data collection

Subsection 1 : Web Scrapper

- Creating a time triggered web scrapper that will collect news from various sources
- first a set of News sites will have to be identified

Subsection 2 : Database

- Given that the store of files in MongoDB is already in the json format it will make things simpler as the file format will not have to be converted for processing

Note: A proper retrieval system will also be needed for frequent recovery and update of values in the DB. Some standard formats will also need to be created to minimise the processing mistakes.

Subsection 3 : Validation Bot or Model

- here i will have to implement an NLP to determine weather or not a certain article or input is worthy or being used in the input. the quality of data here will directly affect the output of the sentiment analysis model and that is a very critical function as it is the edge we have against simple time series analysis and make our prediction more adapt to the real world fluctuations of the world, mostly depicted through the article in information about the.
- I will need some good analysis model to do this with effectively.

Model 2 : Data Processing (Part 1)

Subsection 4 : The Sentiment analysis model

- here i will take the input as each article and grade on a board of 10 points to from -5 to +5 depicting the impact.
  - based on the performance and compute time i can increase the grading parameter and get more effective use from the model

    - but this will come at a cost of training time, i am not sure if we have this time or not hence for now i will stick to the basics and line this feature in the improvement section for later use.
- the quality of the data that is drawn is also very important in the way content bias can influence the decision making sequence. No out wright way to prevent or correct this but minimisation can be done by limiting data to certain creditable sources or greatly increasing the input size ensuring that both side of the bias are covered.

Subsection 5 : The Time Series analysis

- Taking the past data from an API a purely number based analysis will be performed. This will form the heart of the prediction as it will the closest value to true and this will ensure that all the predictions are as close to logarithmic values as possible.
- if we manage to get the values good, the Sentiment analysis model can really act as a buffer zone indicating that the values are likely to sway in one direction and which that will be. Bull or Bear.

Module 3 : Data Processing (Part 2)

Subsection 6 : Combination network

- Here the Output of the Time Sires analysis will be combined with the output of the sentiment analysis model.
- This will be a double or at most triple layer network that will look at the two values and determine the best correlation giving a direct output value that will be the prediction.
- The data required to train this model will be manually generated by me, this is the most performance effective way to ensure correct coupling as artifacts at this level will reduce the efficiency of the final output regardless of the working of the previous models.

Module 4 : Data Output

Subsection 7 : Data Delivery

- Here the Final output in a predefined format will either be updated to the DB or be updated in the file system as a json.

- My personal preference is adding the File to the MongoDB server as the files there are stored in Json format offering no change in the retrieval mechanism for the front end but will make things smoother on the back ensuring data integrity as a part of the DB

This will conclude the rough upper level functioning of the Product and the different modules constituting. Details are intentionally left vague as to give the implementer the freedom to act in a manner that most suits his capabilities. Any specifics will soon be updated to the low level document that will be dedicated to each of the Modules.

If further explanation are desired that are not clarified through the combination of low, high and technical documentation. Direct contact is recommended.