Dataset Structure and Schema

The data used for this project will be based on the coloration of stock price change. This data will be collected from websites like Reddit that store a vast amount of conversation that stock is receiving on the wall street bets as well as stock data from financial websites like yahoo finance.

The data gathered from these websites will be stored in a JSON document, where each document will be a collection of nested subdocuments based on the name of the stock and the date. within each date sub object, there will be another sub object that differentiates on where the data is being extracted from i.e., yahoo finance- stock\_data and Reddit data. Due to the expansive amount of data present on Reddit, we will limit our stock count to only 26 well-known stocks as the interactions for a few stocks are way more than others, and for a few stocks there might not be much data available. We will also limit the period when we want to analyze the stocks due to the same reasons. The period that we are considering for the creation of our dataset will be from Jan 1, 2020, till Dec 31, 2020.

While extracting the data from these websites we will also try to strip off any occurrences of commas, tabs, newlines, pipes, quotes, or any non-ASCII values that occur so that the data is consistent and easy to store and analyze.

Following is a sample of the JSON format of how the data will be stored along with the data types of the name/value pairs.

**SCHEMA**:

Here STOCK\_NAME and DATE are of type string.

{

STOCK\_NAME : {

DATE: {

STOCK\_DATA : {

"fields": [

{

"name": "Open",

"type": "number"

},

{

"name": "High",

"type": "number"

},

{

"name": "Low",

"type": "number"

},

{

"name": "Close",

"type": "number"

},

{

"name": "Adj Close",

"type": "number"

},

{

"name": "Volume",

"type": "integer"

},

{

"name": "Change",

"type": "number"

},

{

"name": "Change Rate",

"type": "number"

}

]

},

REDDIT\_DATA: {

"fields": [

{

"name": "submissions",

"type": "object"

}

]

}

},

NEXT\_DATE: {

...

}

},

NEXT\_STOCK\_NAME : {

...

}

}

**SAMPLE DATA:**

{

"AMC" : {

"2020-01-02": {

"stock\_data": {

"Open": 1,

"High": 1,

"Low": 1,

"Close": 1,

"Adj Close": 1,

"Volume": 1,

"Change": 1,

"Change Rate": 0.1

},

"reddit\_data": {

"submissions": "subs"

}

},

"2020-01-03": {

…

},

...

},

"LC" : {

...

},

...

}