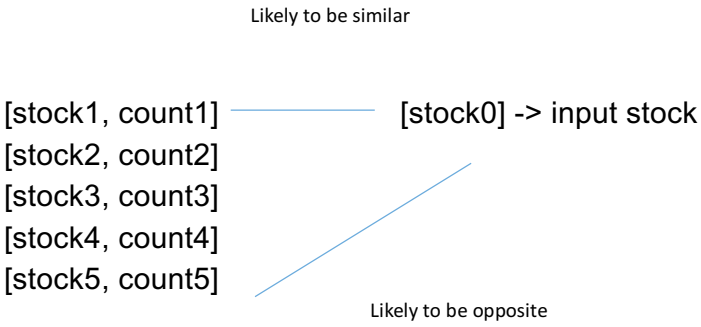

Method 1:

Firstly, we will have a set of stocks, with their everyday price within a fixed time period. Here is one input stock, our goal is to find that whether there exists some stocks that have strong correlation with the input stock, either similar movements of price or opposite movements of price. In the way, we can mine the hidden relationship between stocks, and help financial analysts to make wiser decisions.

The fundamental method is to use the weekly price of two stocks to identify their correlation. We don't choose to use the daily price so as to avoid overfitting. Because it is common for stocks price to have a fluctuation range within a few days. In order to fully use the advantage of MapRedue, we use the counting method to calculate correlation between two stocks. We need to find out whether a stock is increasing or decreasing its price this week compared with its price last week. For two stocks, we have a variable 'count' to record how many times do the two stocks have same price increase action or price decrease action, each time two stocks have the same future trend of increasing of decreasing the count variable will increment by one. Also, each time two stocks have the opposite future trend of increasing of decreasing the count variable will decreased by one.

[stock1, (price_day1, price_day2, price_day3, price_day4…., price_dayN)]
[stock2, (price_day1, price_day2, price_day3, price_day4…., price_dayN)]
[stock3, (price_day1, price_day2, price_day3, price_day4…., price_dayN)]
[stock4, (price_day1, price_day2, price_day3, price_day4…., price_dayN)]

For the input stock, we will compare it with all other stocks in our set, and calculate a 'count' variable for each time we compare. So there will be a list of 'count' variable with different value, we will sort these 'count' variable and find the top k maximum count variables and the bottom k minimum count variables. These stocks that have the highest count value is most similar to the input stock, and those with the lowest count value is likely to be the stocks that have mostly opposite behaviors from the input stock.

count variables are sorted here:

Method 2:
Besides of stocks' weekly price data, we also have raw data in the form of plain text. We will search for article from financial journals and try to abstract information from plain text articles. The basic idea is that if two companies name appear in the same article for many times, it may reveal that these two companies could have relationship in business. So, for the input stock's company name, we will try to count other companies which appear in the same article with it, and sort these count numbers to find k companies that appear with the input stock's company in the same article most frequently.

Method 3:
Since our data includes plain text, we can try to do more refinement besides counting method. We try to applied natural language processing's method to this project. We will have two sets of adjective words, one set is for positive words, and another set is for negative words. The two word sets will help us to do opinion mining for an input stock. Basically, we will use the NLP method of bag of words and calculate a percentage number for each input stock's company's positive rate.