Ruida Xie

Big Data in Finance

Assignment C

03/19/2016

Stock Price Prediction by Tweets

To put the idea in the simplest way, I load the data into Apache Spark and then transform them from the .json format to .txt format. The reason for making this choice is simply that text files are easier to work with. The Python scripts contain multiple functions, whose purpose includes counting the instances of company names of my choice and computing the ratio of good-sentiment words and bad-sentiment words.

To predict the stock returns, I filter those tweets that contain the target companies and study them. There are many ways to analyze the tweets, but the simplest way is to search the tweets for words with positive or negative sentiments. Next, I define a variable called the good/bad ratio as an indicator for whether the stock prices will move up or down. If the ratio is above a certain level, 1.2 for instance, I predict the stock price will increase. I have written a merge function to combine tweets at a certain time with its related stock returns at roughly the same time. It is a concern that most of the tweets were posted outside of the 390 minutes of trading time per trading day and many more were posted during the weekend. Therefore, I count tweets after 4:30 pm as affecting the next trading day.

The most important step for predicting stock returns is to construct a regression relationship between the good/bad ratio and the stock returns. I choose to use the simple regression as my model for predicting stock prices. For example, the tweets and the Google stock returns in March 2013 gave me the regression formula y=0.5x-0.5. If the good/bad ratio for April 2013 was 1, I expect the stock price to stay at the original level. If 1.1, then the price is expected to move up by 5% within days. To make the model more accurate, I need to train the model by using the historical data to fit its predictions and adjust the constant terms in proportion.