BdiF Assignment C

Xiaobo He

This assignment gives us a great opportunity to employ big data tools for uncovering valuable finance information, which is buried in huge mount of Twitter data. Twitter data has a quick response for market sentiment and emergency events, which is one of the best indicators for the market trends. Profits can be achieved if the accurate and rapid analysis is running on the real-time or semi real-time Twitter data with efficient programs.

Step one: Data Exploration

Thanks to Andrew so we have a cleaned sample data containing twitter information. The sample data is comparably small so I can use Excel and Python to analyze it. Some words can have positive impacts for the market, and some others may indicate negative impacts. The basic idea will be to take the frequency and intensity of twitter sentences with positive words as a positive indicator for one specific company, and also take the negative words as a negative indicator, and then we can take the weighted impact for this specific company to predict the market trend in stock markets.

Step two: Data Programming

The raw twitter data is pretty dirty with a lot of unrelated information. Before we can do serious analysis, we have to filter out the noise, which will speed up our following data process.

Two predefined word lists for positive and negative are built up. The positive word list will be "BEST, ABLE, ACHIEVE, BENEFICIAL, GOOD, NEW, HIGHEST, IMPROVEMENT, INNOVATIVE, POPULAR, RESOLVE", and the negative one will be "INFRIGEMENT, ABANDON, ABNORMAL, ABSENCE, ACCIDENT, ADVERSE, AGAINST, BAD, DENY, DETERIORATE, DIFFICULT, DOWNTURN, ERROR", respectively.

Also predefine a name list for interested companies we are watching in the market. All sentences containing the keywords will be scrubed as valuable information for further analysis.

Step three: Data Analysis

In this part, statistic analysis for the frequencies of the positive and negative keywords is performed. The program will be done in Apache Spark on AWS system. I plan to write a simple version in my local machine before to test with AWS.

Step four: Data Insights and possible improvement

As stated previously, the timing and accuracy is very important for the benefit from Twitter data analysis, so it will be better if a real-time version program is developed in future.