CS 5344: Big Data Analytics Technology

Project Status Report

By

Li Chenfang A0055805E

Li Xijun A0148517Y

Yu Jianmin A0137787L

12 Feb 2016

The Idea

Although the academic generally accept the Efficient Market Hypothesis which explains that the stock market price follows a random walk pattern, tremendous researches have been conducted in the field of stock market price prediction. We want to test a hypothesis from the perspective of behavioral finance, that the social medias have a noticeable impact on individuals' decisions, thus, building a direct correlation between "public sentiment" and "market sentiment". Specifically, we hope to discover certain pattern between the stock price of Regal Entertainment Group (REG) and related Twitter discussion data of new movies. REG operates the largest movie theater chain in USA. Whenever there is a new movie release, Twitter will have the corresponding discussions. We assume that the more tweets about new movies on Twitter, the more people will go to cinemas to watch, hence, bringing more revenues to REG.

The Methodology

We study the problem of co-relating the Twitter activity with the change in stock price of REG for selected new movies on each available month in 2015. For example, there was a hot movie Cinderella release in Mar 2015, which had a large number of tweets discussing the movie. For a period one week before and after the movie release date, we will select tweets and analyze them based on certain characteristics of the tweets including the count of users and tweets with the movie (#Cinderella). Besides, we will try to study the stock price performance of REG in this period. And we will calculate the holding period to profit most if any. Also, we will try to analyze the tweets trend one week before the move release date to predict the popularity of the movie, and then we will use it as a guide to swing trade the REG stock. Hopefully, we can study all the available months with tweets data for this kind of analysis and see whether there is a pattern.

Twitter Data

We will download the raw tweets research data from:
http://archive.org/search.php?query=collection%3Atwitterstream&sort=-publicdate

for all available months in 2015.

REG Stock Price Data

We will download the data from Yahoo Finance! with symbol RGC.

Project Status

We have downloaded the Mar 2015 tweets data for analyzing. The data size is about 16G. We don't know how to split the data for running Hadoop program in our own laptop. We find that the total tweets data from archive.org may exceed 150G. Also, we are reading papers related to Twitter sentiment analysis on stock price prediction and learning R programming.