Big Data Individual Project Proposal

Submitted by: Zhiao Zhou

Project Title: Stock Price Prediction by Real Time News Analytics

Project Description: (Write one paragraph to describe what this analytic will do.)

This project is aimed to build a real time Chinese stock price prediction model by detecting the relationship between a company's stock price and its corresponding news sentiment, using Machine Learning. First, history social media data about a company will be retained through either API or crawler from Sina Weibo, popular news websites, WeChat and so on, as well as corresponding stock prices in China. Then a sequence neural network model will be trained to detect their inner relationships. Finally, the model can be used to do real time sentiment analysis and stock price prediction based on real time streaming news and social media data.

Describe who will benefit from your analytic:

This project will benefit a lot of people or agencies such as individual investors, institutional investors, equity researchers and so on, since they can use the prediction as a reference. In addition, if possible, it can also be integrated into a real time stock trading system.

Describe how you will check the goodness of the analytic, i.e., how will you prove the results are accurate and can be trusted:

In the machine learning model, a evaluation metric will be determined before training. For example, accuracy rate, precision rate, recall rate and F-k score will be used to make sure the model is doing great. The accuracy should be at least 80% and the F-k score should be at least 85%.

Big Data Individual Project Proposal

Name of Data Source 1: Sina Weibo

Data Source Description *(Provide a short description of the data source.)*

Sina Weibo is a Chinese microblogging website – a Chinese version "Twitter". It is one of the most renowned and popular social media platforms in China, which has about 0.3 billion active users. So this makes it a super powerful symbol for Chinese people's sentiment.

Data Size (Estimate size, e.g. MB, GB, TB?)

20 GB

Data Collection Frequency

Is the data source a static, periodic, or realtime (i.e., near realtime) source?

realtime

If realtime data, what is the frequency with which you will collect the data and what is the volume of data collected at each interval?

2000 times/hour. About 50000 "tweets" per interval.

If not realtime data, will you collect a batch of data periodically or just once (static)?

If the data will be collected periodically, how often will you collect it and what is the volume of data that will be collected at each interval?

Big Data Individual Project Proposal

Name of Data Source 2: Baidu News

Data Source Description *(Provide a short description of the data source.)*

Baidu News is the largest Chinese news search platform in the world produced by Baidu which is a Chinese Tech Company. The news sources contain over 500 authoritative social media.

Data Size **(**Estimate size, e.g. MB, GB, TB?)

20GB

Data Collection Frequency

Is the data source a static, periodic, or realtime (i.e., near realtime) source?

realtime

If realtime data, what is the frequency with which you will collect the data and what is the volume of data collected at each interval?

Per second. All of data possible will be crawled as possible as I can.

If not realtime data, will you collect a batch of data periodically or just once (static)?

If the data will be collected periodically, how often will you collect it and what is the volume of data that will be collected at each interval?