Big Data Application Development - Team Project Proposal

Part 1. General Information

Team Members:
Lizi Chen, Chun-Yi Yang

Project Title:
Stock Price Prediction by Financial News Sentiment Analysis

Project Description:
In this project, we are going to build a stock price prediction model that takes in real-time financial news and produces investment suggestion such as buy or sell.

Who is a typical user of your application:

1. Individual Investor
2. Portfolio Manger
3. Stock Market Learner and Researcher
4. Business Strategist

What insight will you derive from the data?

1. By verify the ground truth and doing the sentiment analysis on financial news, we can understand sometimes a good news may not be an immediate boost for the stock price, and sometimes bad news does not add enough fluctuation to the market either.
2. By measure the time lag between a published news and how the stock price may change, we can tell the category of news that have a stronger assertion to provide more fluctuation in stock price or less.
3. By generally train the model to make investment suggestions according to news, we can see how blast of news and its language can affect the market.

What actuation decisions can be made based on the actionable insight?

1. buy a stock
2. sell a stock

Big Data Application Development - Team Project Proposal

Part 2. Data Source Information

Name of Data Source 1: Historical Intra-day Stock Price

Data Source Description:

We are going to retrieve historical intra-day stock price dataset according to the retrieved news. Each business entity will have its own CSV data file.
https://stooq.com/db/
https://www.google.com/finance/getprices?i=60&p=2d&q=IBM
https://www.quantshare.com/sa-426-6-ways-to-download-free-intraday-and-tick-data-for-the-us-stock-market

Data Collection Frequency	Data Size	Data Frequency
• Are you collecting the data in realtime? Or are you collecting it periodically? • Are you collecting static data? (e.g. historic data that you load once)	• *Estimate size of the data you will store, e.g. MB, GB, TB, PB*	• If realtime data, what is the frequency and volume of data (how often and how much data will you collect at a time)? • If batch data, how often will you collect it?
□ Realtime (ongoing near-realtime collection) □ Batch (multiple non near-realtimecollections) □ Static (one time collection)	□ MB □ 1-10 GB □ 10-100 GB □ 100-300 GB □ 300-500 GB □ > 500 GB	**If realtime data:** • How often will you collect data? □ Every second, or every few seconds □ Every minute, or every few minutes • What is the size of data you will collect at each interval? ________Very minimal KB size data____ **If *not* realtime data:** • Will you collect a batch of data periodically or just once (static)? □ Just once □ Every hour, or every few hours □ Every day, or every few days, or every week, or every month • How much data that will be collected at each interval? ____

Big Data Application Development - Team Project Proposal

Part 2. Data Source Information

Name of Data Source 2: **Reuters News Archive**

Data Source Description: We will retrieve Reuters News from Reuter's New Archive for the past several months. Also, we will keep tracking the upcoming new data for final product and for updating the new model.

Data Collection Frequency • Are you collecting the data in realtime? Or are you collecting it periodically? • Are you collecting static data? (e.g. historic data that you load once)	Data Size • *Estimate size of the data you will store, e.g. MB, GB, TB, PB*	Data Frequency • If realtime data, what is the frequency and volume of data (how often and how much data will you collect at a time)? • If batch data, how often will you collect it?
□ Realtime (ongoing near-realtime collection) □ Batch (multiple non near-realtimecollections) □ Static (one time collection)	□ MB □ 1-10 GB □ 10-100 GB □ 100-300 GB □ 300-500 GB □ > 500 GB	**If realtime data:** • How often will you collect data? □ Every second, or every few seconds □ Every minute, or every few minutes • What is the size of data you will collect at each interval? __________Around 20 MB for raw HTML files______________________________ **If *not* realtime data:** • Will you collect a batch of data periodically or just once (static)? □ Just once □ Every hour, or every few hours □ Every day, or every few days, or every week, or every month • How much data that will be collected at each interval? ____

Big Data Application Development - Team Project Proposal

Part 2. Data Source Information

Name of Data Source 3: WSJ News Archive

Data Source Description:

Similar to the Reuter News Archive, we will retrieve WSJ news for the past several months for model construction as well as keep tracking fresh news for further model refinement.

Data Collection Frequency

- Are you collecting the data in realtime? Or are you collecting it periodically?
- Are you collecting static data? (e.g. historic data that you load once)

□ [✓ Realtime (ongoing near-realtime collection)]

□ Batch (multiple non near-realtimecollections)

□ Static (one time collection)

Data Size

- *Estimate size of the data you will store, e.g. MB, GB, TB, PB*

□ MB

□ [✓ 1-10 GB]

□ 10-100 GB

□ 100-300 GB

□ 300-500 GB

□ > 500 GB

Data Frequency

- If realtime data, what is the frequency and volume of data (how often and how much data will you collect at a time)?
- If batch data, how often will you collect it?

If realtime data:

- How often will you collect data?
  □ Every second, or every few seconds
  □ [✓ Every minute, or every few minutes]
- What is the size of data you will collect at each interval?

______________Similar to Reuter, around 20 MB for Raw HTML files.____________

If *not* realtime data:

- Will you collect a batch of data periodically or just once (static)?
  □ Just once
  □ Every hour, or every few hours
  □ Every day, or every few days, or every week, or every month
- How much data that will be collected at each interval?

____

Big Data Application Development - Team Project Proposal
Pending Data Source (To Be Evaluated)
1. GDELT World News Real-time Archive
2. CityFalcon Rated News Archive