Big Data Application Development - Initial Project Proposal

Part 1. General Information

Submitted by: Lizi Chen

Project Title: Real-Time Financial News Sentiment Analysis

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

Use Flume or Spark Streaming to periodically ingest news (mostly titles) from various online news publishers such as CNN Money, Wall Street Journal, Bloomberg. Analyze each batch of news, eliminate similar ones; or the ones that has shown in the previous batches. (This step can leverage the program and approach that I had done in the previous class project in Real-time Big Data class.) Identify the news that has not appear in the history, or has extra information compare to its previous version. Analyze the content of the latest ingested news and assign unique IDs for each. Based on the sentiment analysis conclusion, provide concise investment suggestions based on the template such as [Investment Action: Buy/Sell] [Subject: Company Name] [Assertion Level: (0, 1)]. For example: [Buy] [Amazon Stock] [0.892], or [Sell] [Amazon Stock] [0.013]

Who is a typical user of your application:

Individual investors, Hedge Fund Portfolio Manager, Business Strategists, etc.

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 - Initial Project Proposal *(continued)*

Part 2. General Data Source Information

Name of Data Source 1:

Financial News Dataset from Bloomberg and Reuters：https://github.com/philipperemy/financial-news-dataset

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

450,341 news from Bloomberg and 109,110 news from Reuters.

Full-text news dataset from Bloomberg and Reuters, including title, date, and news content.

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

Data Collection Frequency

Is the data source a static, periodic, or realtime (i.e., near realtime) source?
This is static data source for training the sentiment analysis model.

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?
N/A

If not realtime data, will you collect a batch of data periodically pr just once (static)?
This is for training the model, we will have to label it then feed as it goes.

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?
N/A

Big Data Application Development - Initial Project Proposal *(continued)*

Part 2. General Data Source Information (continued)

Name of Data Source 2: (This will be many live data scrapped from financial news website)

http://money.cnn.com/news/

https://www.bloomberg.com/markets

https://www.wsj.com/news/us

We may have to add more or remove less in the future based on our time availability.

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

We will collect the real-time data from live financial news websites, such as CNN money, Bloomberg market, and WSJ. We will mostly focus on the news title and probably a short abstract in the beginning of the news. The approach to process and digest the content of the news is pending for further discussion.

We will scrape the front page of the website with featured news periodically in a batched fashion.

By using Sparking streaming(and probably some other data streaming and ingest frameworks), we will be able to process data rapidly.

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

MB

Data Collection Frequency

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

Near real-time / periodically.

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?

We will assume half an hour since the front page of financial news may not update drastically in a short peiord of time. Each interval we will process several websites and the overall data volume may range from 5 MB to 20 MB since it's mostly text.

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

N/A

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?

We will assume half an hour since the front page of financial news may not update drastically in a short peiord of time. Each interval we will process several websites and the overall data volume may range from 5 MB to 20 MB since it's mostly text.