CS-GY 6313 : Information Visualization

GrapeVine

A tool for assessing the value of news spread

Aditya Rajmane - anr331

Monil Suthar - ms8624

Mark Weisenborn - mw1556

1. Introduction

The aim of this development is to create a visualization that helps the social media researcher to understand how nodes of information relate and how the information is tied to changes in the prices of stocks.

The external collaborator, our mentor, for this project is a doctoral student at NYU who is a co-founder of Accern, a news analytics company in New York City.

Anshul Vikram Pandey

500+ connections

Co-Founder at Accern

Brooklyn, New York | Research

Current	Accern, New York University
Previous	Deloitte, Human Computer Interaction Laboratory, NUS, Keio-NUS CUTE Center
Education	New York University
Recommendations	**1** person has recommended **Anshul Vikram**
Websites	Personal Website

ACCERN

For the past several years, Anshul has designed Accern programs to analyze social media announcements and regular news announcements. He and his team have developed a method to scan a very wide range of online sources in very nearly real-time. The objective of the company i[1]s to find information that may end up at major news venues. Company clients are mostly hedge funds and other investors engaged in trading on public

[1] Front page: http://accern.com

information before the information has fully saturated the market. The company tracks some 20,000,000 sources in real time.[2]

We are therefore working in collaboration with Accern (http://accern.com/), which analyzes million of news to provide information for traders to help them into their decision making. This visualization tools will help Accern better understand how news evolve and how their internal automated systems for new detection behave."

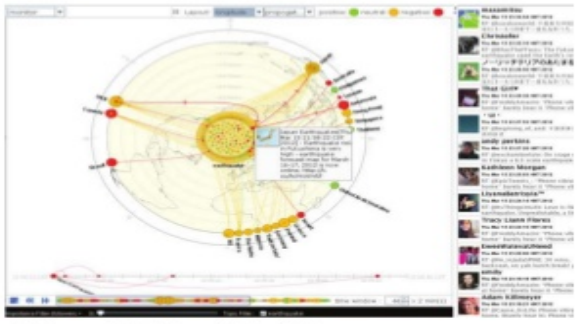
This problem statement drove us to have a cohesive visualization which would suffice the needs of people at Accern, our major focus in this area hence would be visualizing the trend and giving out the temporal analysis of the overall feeds and its direct comparison to the real fluctuation in that financial entity.

This would as defined in the problem statement would help people at Accern to understand the evolution of news and compare it directly with the behaviour of the news feed.

2. Related Work

Contemporary news studies typically depict relationships between sources with wheel charts, as shown in the examples below from Google Ripple and Whisper.

Google Ripples [F. Viégas et al. 11]



Whisper [N. Cao et al. 12]

[2] Google Ripples and Whisper:
http://www.slideshare.net/ycwu8/visual-analysis-of-topic-competition-on-social-media

The disadvantage to these tools is they don't depict the related price change. We therefore refocused efforts on academic works from economics. In terms of related work there is a fairly small body of knowledge published on the subject of how stocks change price after internet news is published about them. The most relevant study gained attention in the Wall Street Journal recently and is viewable here: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1807265.

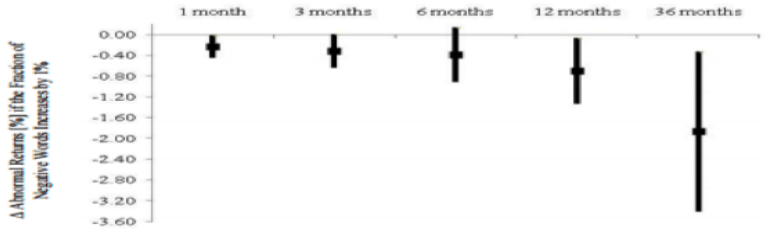
In the study professors identified that stocks do react to news published at a single source - Seeking Alpha. Our study is much broader as we are helping Accern examine countless more sources using a systematic method that shows price movement. Below is the example of how the original authors of the first study accomplished this.

We also located additional work that shows intraday price movement following publication at a single news source, Seeking Alpha. This study shows the price movement during the day, as fully described in the pasted body of this sample visualization. Here is an example that one of our group members designed to show intraday movement of the price. We will adapt this model for this client.

Figure 3. Seeking Alpha and Abnormal Returns over Different Holding Periods

This figure reports coefficient estimates from regressions of abnormal returns on measures of the views reflected in *Seeking Alpha* (SA) articles and comments. The sample period is 2005-2012. Abnormal returns are the company's raw returns minus the return of a value-weighted portfolio with similar size/book-to-market/past return-characteristics. The horizons over which cumulative abnormal returns are computed are 1 month, 3 months, 6 months, 12 months and 36 months. The regression equation is identical to the one in column (3) of Table 4. Here, we plot the coefficient estimates on $NegSA_{i,t}$ and $NegSA\text{-}Comment_{i,t}$ along with their corresponding 95% confidence intervals. Standard errors are clustered by firm and year-month.

Panel A: Coefficient Estimate on $NegSA_{i,t}$



Panel B: Coefficient Estimate on $NegSA\text{-}Comment_{i,t}$

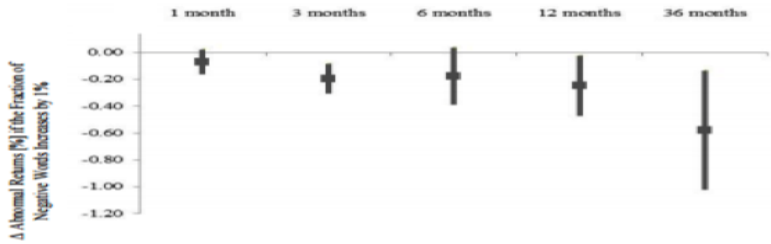
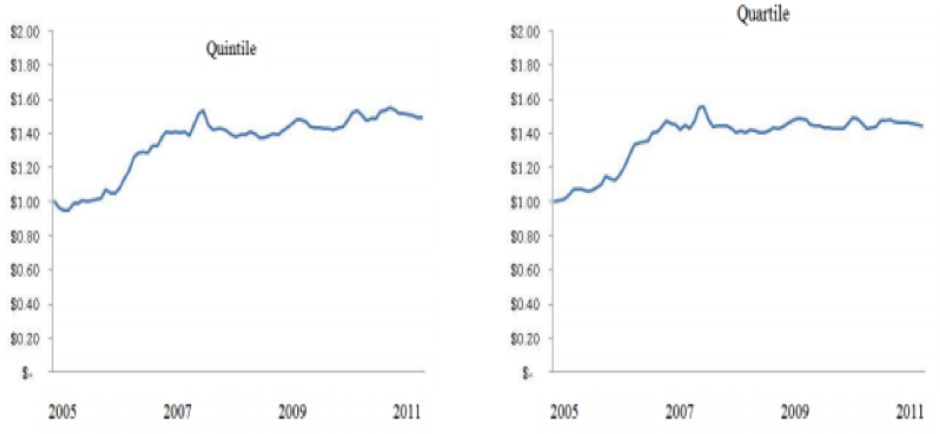


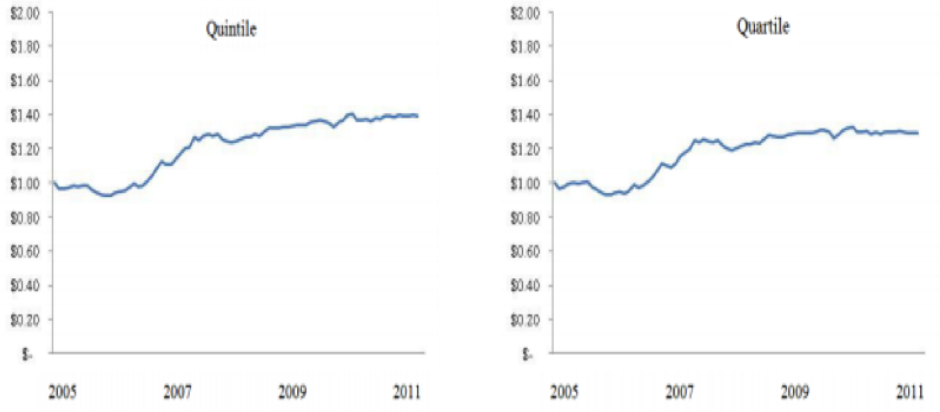
Figure 4. Seeking Alpha and Abnormal Returns over Different Holding Periods

This figure depicts how $1 invested in a simple calendar-time trading strategy would have evolved. The trading strategy is as follows: At the end of each trading day t, we assign stocks into quintile (quartile) portfolios based on the average fraction of negative words across all articles published on SA about company i on day t ($NegSA_{i,t}$); we also form quintile (quartile) portfolios based on the average fraction of negative words across SA comments posted over days t to t+1 in response to the SA articles ($NegSA\text{-}Comment_{i,t}$). We skip two days and hold each stock in its respective portfolio for three months. Based on the daily returns of a long-short portfolio, where we go long stocks in the bottom quintile (quartile) and short stocks in the top quintile (quartile)), we plot how much $1 would have grown/shrunk through calendar time.

Panel A: $NegSA_{i,t}$ - Based



Panel B: $NegSA\text{-}Comment_{i,t}$ - Based

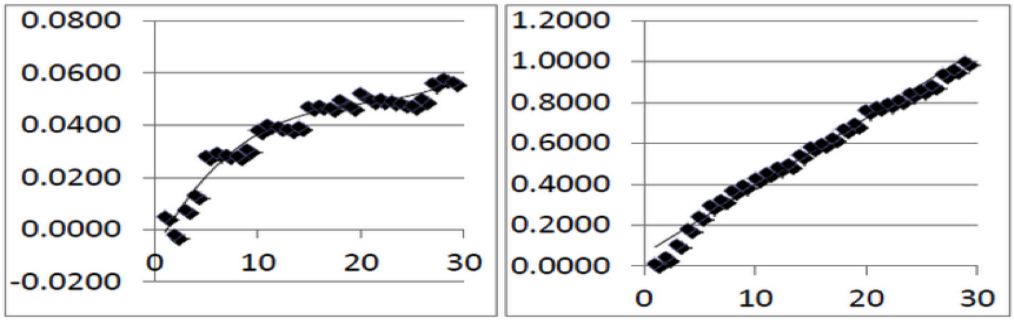


Panel 1. Average Price Path and Average Cumulative Average Volume

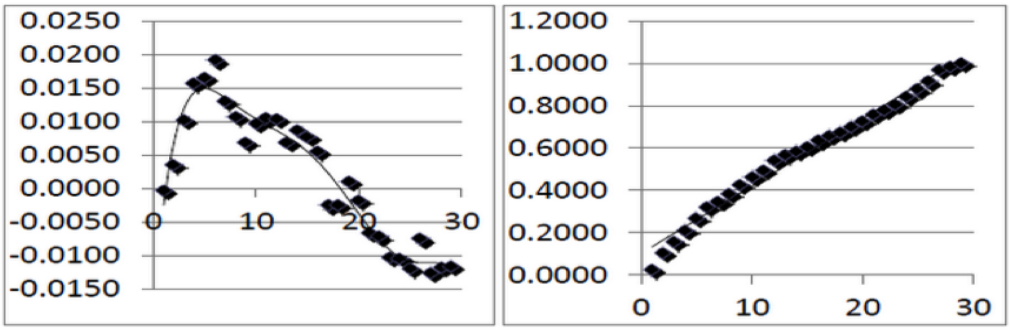
These figures present average trading price paths for the days of trading that were statistically significant. For each revision, price begins at time 9:30am and completes at 4:00pm on the x-axis and magnitude of change is presented in percentage format on the y-axis.

The left chart is price, the right chart is cumulative volume.

"Top" Micro Long Day t



Micro Short Day t



Price and volume charts were constructed using tick data provided by Wharton Research Data Services. The trading day is constructed as a series of 15-minute bins beginning at 9:30am. Each point represents an average of prices or volumes closest to but not after each bin-time. For each category a small (<20) sample of securities that had tick data available from September to December 2013 was used to develop average values. These sorts of charts may be helpful for a trader trying to minimize implementation shortfall and to help understand the different price and volume behaviors of longs versus shorts. Because this data set is from the group that was announced outside of market hours the initial rise in price on the short chart may be due to closing a downward gap in the initial opening price print. In order to fully understand the price formation process every security should be included in the values used for depiction, a task beyond the scope of this academic study but worthy of future research.

3. Data:

The format of the data will be in JSON/CSV/XML it would be given to us by FTP. The data would contain following attributes

1. Format
2. Categories
3. Coverage
4. Company (x) 1-Year
5. Analysis

Format explains the field, which is present in the dataset. The category would explain the current event category. The coverage field contains a list of all the entities we cover in the data. Company (x) contains x-year historical data on Company. The analysis tab contains a very short analysis on Company. Fields that are highlighted in the database have some special meaning like: orange symbolized values that updates/moves in real-time.

Our example data is shown here and is an intensive data set. We will measure the price impact of each source_url per for each keyword. We will also replicate this for author_name. At the end of the analysis the company will know which source_urls are providing the greatest price impact and will understand which keywords matter. Those keywords that are not statistically significant in terms of resulting price impact will be ignored but those that are associated with significant movement will be fully explored to understand how each source_url is influencing other source_urls in that universe of relevant material. We feel that limiting the scope to only the most statistically significant keywords will help define the study so that the client gains an understanding of the most relevant parts of the information flow and then uses those relevant parts in their discussions with their own clients. Examples of keywords, stock symbols, source_urls, and authors is below:

keyword	sym	source_url	author_name
"Speculation",	DDD	http://www.katv.com	Dan Levine
"Intellectual..	MMM	http://english.capital.gr	Susan C. Schena
"Partnership",	ABT	http://www.stltoday.com	Auto Shop

"Allegation",	ABBV	http://news.findlaw.com	Jim Brewer
"Fraud",	ANF	http://www2.counton2.com	Scott Davis
"Layoff",	GCH	http://www.orlandosentinel.com	Kristian Gore
"Accomplishment	JEQ	http://www.sify.com	Seth Barnet
"Payments",	SGF	http://www.wwmt.com	Nolan Pearson
"Budget",	ABM	http://ap.savannahnow.com	Zach Kirkland
"Financial ...	AKR	http://www.wcpo.com	admin
"Financial ...	ACN	http://www.hawaiinewsnow.com	Pimentel
"Financial Investments",	ACCO	http://www.mromagazine.com	Eva Dou

DATA :

```
>db.accern.findOne()
{
        "_id" : ObjectId("54150c0d7798f25d47000004"),
        "event_groups" : [
                {
                        "type" : "Payment",
                        "group" : "Rumors"
                },
                {
                        "type" : "Speculation",
                        "group" : "Rumors"
                }
        ],
        "story_saturation" : "high",
        "correlations" : [
                {
                        "of_entity" : "AAPL",
                        "with_entity" : [
                                {
                                        "ticker" : "QQQ",
                                        "type" : "max_positive",
                                        "value" : 0.722533873054272
                                },
                                {
                                        "ticker" : "MGT",
                                        "type" : "max_negative",
                                        "value" : -0.0964395853499532
                                }
                        ]
```

```
                }
        ],
        "first_mention" : false,
        "article_type" : "news",
        "story_volume" : 85,
        "article_url" :
"http://www.myfoxdfw.com/story/19372866/news-summary-samsung-ordered-to-pay-apple-105b",
        "story_id" : ObjectId("543dd75a7798f259d3b7d6a6"),
        "entities" : [
                {
                        "sector" : "Technology",
                        "index" : "S&P 500, Russell 1000, Russell 3000, Wilshire 5000, BARRON'S 400, NASDAQ 100",
                        "entity_id" : [
                                "EQ0010169500001000"
                        ],
                        "name" : "Apple Inc.",
                        "exchange" : "NASDAQ",
                        "type" : "Public",
                        "industry" : "Computer Manufacturing",
                        "global_id" : [
                                "BBG000B9XRY4"
                        ],
                        "competitors" : "GOOG, HPQ",
                        "country" : "United States",
                        "ticker" : "AAPL",
                        "region" : "North America"
                }
        ],
        "version" : 1,
        "event_impact_score" : {
                "overall" : 74.1084760649978,
                "on_entities" : [
                        {
                                "on_entity" : 70,
                                "entity" : "AAPL"
                        }
                ]
        },
        "story_shares" : 0,
        "overall_source_rank" : 10,
        "article_id" : ObjectId("54150c0d7798f25d47000004"),
        "event_author_rank" : [
                {
                        "event" : "Wall Street Whispers",
                        "author_rank" : 10
                },
                {
                        "event" : "Payments",
                        "author_rank" : 10
                }
        ],
        "event_source_rank" : [
                {
                        "source_rank" : 10,
                        "event" : "Wall Street Whispers"
                },
                {
```

```
                        "source_rank" : 6,
                        "event" : "Payments"
                }
        ],
        "article_sentiment" : 0.11,
        "overall_author_rank" : 10,
        "harvested_at" : ISODate("2012-08-25T11:08:34Z"),
        "source_id" : 398,
        "author_id" : 331,
        "article_traffic" : 29984315,
        "story_sentiment" : 0.062,
        "story_traffic" : 29472109,
        "avg_day_sentiment" : [
                {
                        "entity_of" : "AAPL",
                        "value" : 0.088
                }
        ]
}
>
```


4. Tasks and Questions

Does the visualization show connections between venues?

The visualization shows the generalized links between news venues. We may add some single number above each arrow to show to frequency (as a percentage) of times a small news source preceded a larger news source publishing.

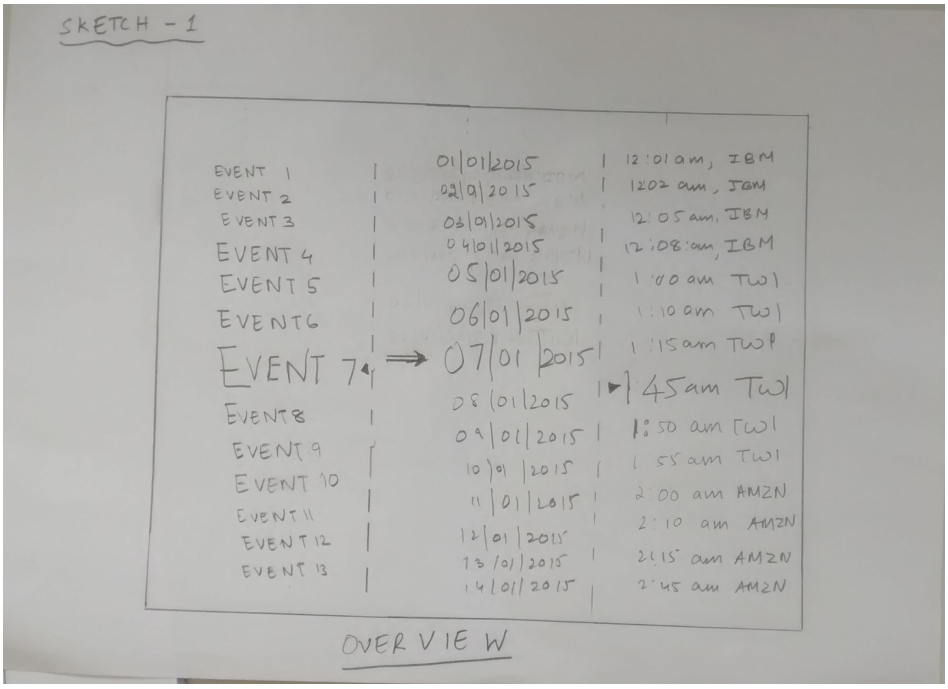
Does the visualization demonstrate a connection between information and price?

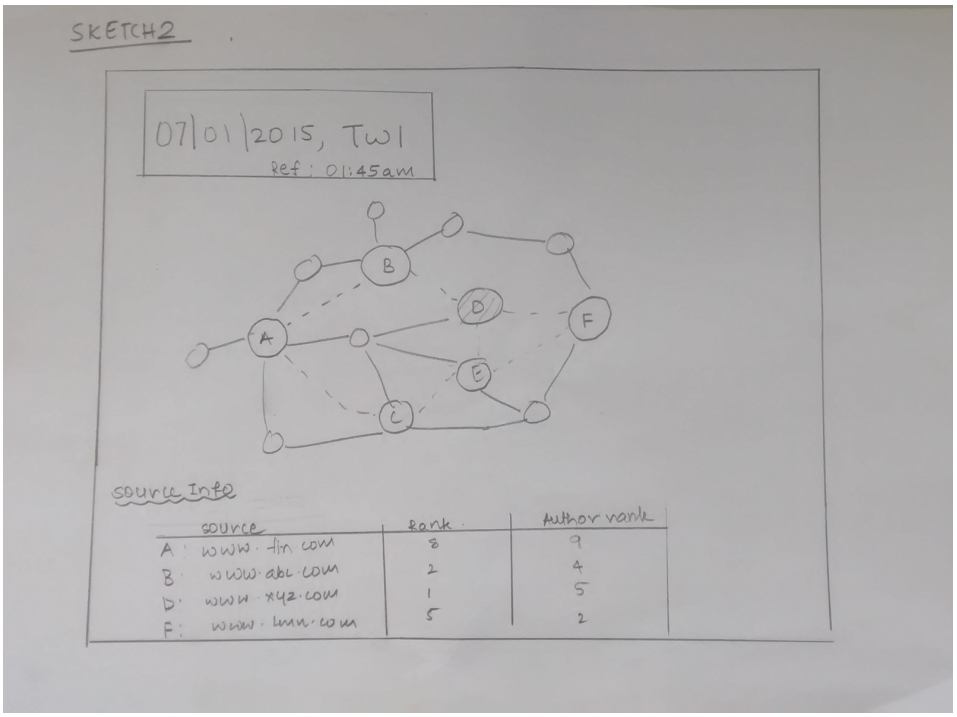
The visualization shows time, ordering, size and resulting price move over a generalized day. We are using a random sample of the full data set per headline word to make generalized displays for how each word tends to propagate through the news sources, and the associated price response.

We also currently working on mock-ups that incorporate “source traffic” and “story traffic” fields. We realize that the spread of news also relies on the fact that how much attention is a story getting on the web.

5. Sketches

Conceptual depiction in Sketch 1 (below) then visual depiction in Sketch 2 (bottom)

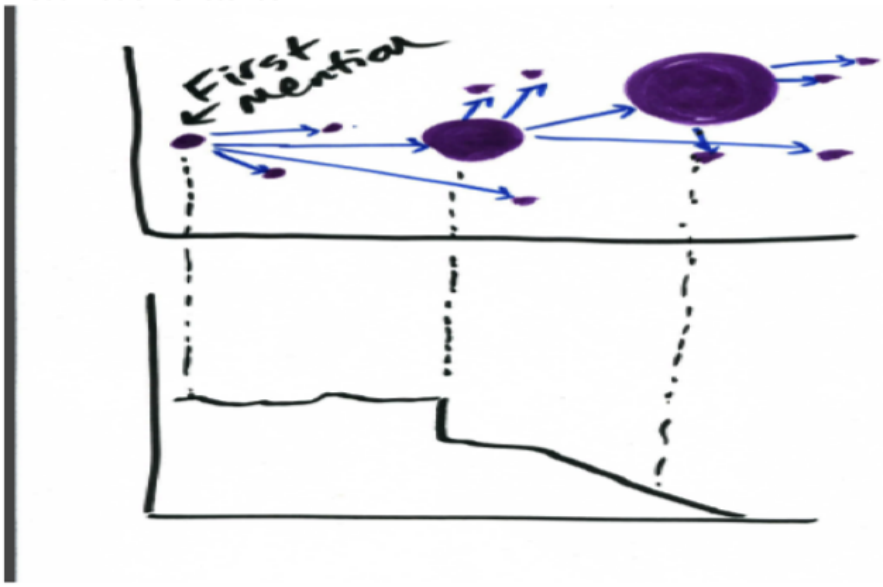




This sketch shows the culmination of **information spread** (encoded in bubbles "nodes") and **price formation** (encoded by the line at the bottom chart). Notice the

dashed lines tie the data together visually for the viewer to digest rapidly. Axis labels will be added but this is time on the x-axis and bottom chart has price on y-axis.

- **Visualization Sketches :**



6. Progress Report

15 May 2015

- We have fully downloaded the data (yay!) which involved client interaction and database manipulation. We will need for this study and are ready to begin analyzing the data in order to make reasonable prototypes.

- We have started the process of building Section 1 of our visualization sketch. Our approach to tackling this has been to start working on some preliminary sketches in order to get familiarized with the D3 platform as it related to our project.

- We revised our prior update in order to address the need for fully encoded sketches that correspond to the data and the information we are presenting in this report.

7. Work plan

The team has devised a following work plan:

There are three main tasks in the development phase:

1. Understanding the incoming data, after which the system will be programmed to use the data. (lead by Mark)
2. Developing a backend to assert our business logic. (lead by Aditya)
3. Develop the UI layer with optimum idioms and visual encodings. (lead by Monil)

Each of these tasks will be lead by a team member while, involving others actively.

Further break down by team members:

1. Mark: Data analysis, HTML UI
    a. Develop initial visual idioms.
    b. Analyse incoming data and provide insights.
    c. Work with Monil to develop the UI layer, and Aditya on Backend.
2. Aditya: Data analysis, Backend Development , UI
    a. Develop a sustainable Backend with respect to the insights provided by Mark
    b. Work with Monil to program the UI layer.
    c. Analyse incoming data and develop idioms with Mark.
3. Monil: Backend, HTML UI
    a. Work with Aditya to program backend
    b. Develop communication services between all system layers
    c. Develop the UI layer and visual encodings

8. Implementation

The visualization shows time, ordering, size and resulting price movement over a generalized day. We are using a random sample of the full data set per headline word to make generalized displays for how each word tends to propagate through the news sources, and the associated price response.

Also, currently we are just focusing on the close to open cycle of the market. So, we primarily are taking a day from 4.00 PM to 9.29 AM of the next day.

We currently are focusing on finalising the data that we will use to represent the news spread. So we are simply working with Mongodb in the initial phase, we will
lock down upon the data and generate a CSV file for the first prototype. We are expecting to put the first prototype along with the data in the next update.

For the next update we are also working on mock-ups that incorporate "source traffic" and "story traffic" fields. We realize that the spread of news also relies on the fact that how much attention is a story getting on the web.

Submission dates:

Project update I, March 15th: Team
Project update II, April 22nd, Team
Project update III, April 29th, Team
Project update IV, May 6th, Team
We understand these dates may need to adjust and will do so according to requirements and the needs of the group members and the course.

9. Release Notes

For this particular update:

a. Two new viz. sketches for highlighting our overview idea.
b. We have further polished the Task abstractions
c. The data section has been augmented by a snippet of the data and mongodb code.