Streaming Data Monitoring

Description: This project built a visualization application to visualize the articles about entities(companies) using CloudLines techniques. The data is from Accern API who provides articles from 200 million websites. The visualization enables users to detect articles episodes about entities and analyze sentiment trend of articles about specific company.

Yash Balar	ydb219@nyu.edu
Fenil Tailor	fst216@nyu.edu
Tianhao Li	tl1924@nyu.edu
Haoran Zhu	hz1025@nyu.edu

Project page: https://github.com/NYU-CS6313-Fall16/Streaming-Data-Monitoring-1
Video: https://vimeo.com/196806769
Demo link: We have written a clear instruction in the README file. Readers can easily follow our instruction and play with the system locally. For now we are still not able to deploy our server on neigher Heroku or our university's server.

1. Problem

In the real world, it is hard to detect some major events happening on companies/sectors in the real-time base. However, it is important for data analyst to understand what is going on right now so that they can extract valuable information as soon as possible. Our visualization system allows users to monitor streaming data and find if there are some important events happening on or across companies.

2. Analytical Questions

2.1 Question 1:

Whether there is a major event or event episode for a particular company/sector?

We want to show people if there is any major events happening by mapping articles from our streaming data source to the dots in our chart. If the dots are getting dense, it means that there may be an event so users can take a closer look at the details by hovering and clicking the dots. The details include the article itself, the sentiment,the impact score,data source and some other properties that users are able to identify by going through it.

2.2 Question 2:
If any major event occurs then is it positive, negative or neutral in sentiment?

This question means that an event could be positive or negative towards a company, which is shown by sentiment attribute in our data. So if we detect an event, we want to understand its sentiment to perform further analysis.

3. Data Description

Our data comes from Accern API, which are basically articles captured from medias. Each datum represents an article, which is about certain companies on some events. For example, a lawsuit to Barclays. The events, like the lawsuit, have also been defined and categorized by the data.

Attribute	Type	Meaning	Range
Ticker	Categorical	The representation of companies in stock market	All companies in the market
Sector	Categorical	The group that a company belongs to.	All sectors defined by the data
Harvested_at	Quantitative	The time that this article be captured	From the earliest time that the API is able to track till now.
Sentiment	Quantitative	Determines if the article was written positively or negatively by the author.	[-1, 1]

4. Related Works

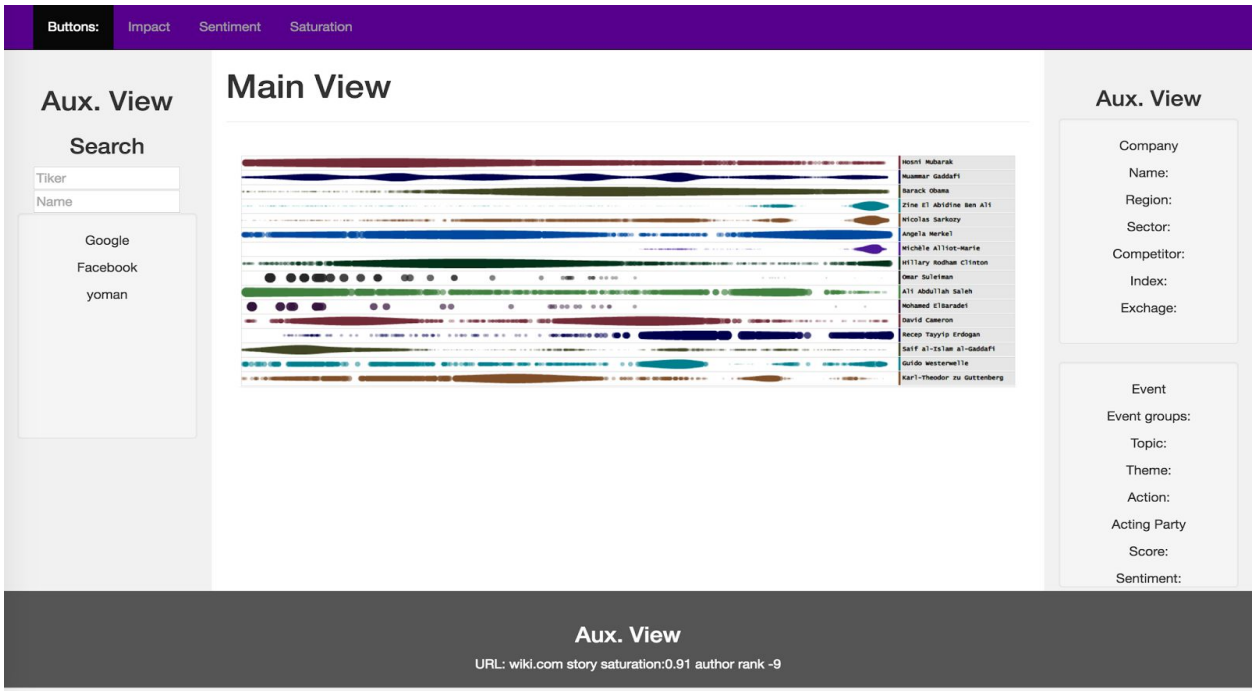
The exact solution for this problem based on the paper *CloudLines: Compact Display of Event Episodes in Multiple Time-Series* by *Milos Krstajic, Enrico Bertini, and Daniel A. Keim* was to divide the timeline in set of beans each of a specific size, then calculating the size of the bubble for that bean by a function which takes into account the number of dots/articles in that bean and giving a size for the bubble in the cloudline. Also, a group of our peers are doing the same project with us. In the previous presentation, they mapped the dot size to impact score, which also looks good.

5. Design Iterations

5.1 Initial Mockup

Our first mockup was a simple html page that describes functionalities we wanted to provide in our project initially.

Screenshot 1



Description 1

After understanding the attributes Accern API provide in response,we thought cloudlines based on companies would be a good idea to visualize the event and event episode.So for our main view we decide to put cloudlines.We also decided to put a interaction feaure that helps in searching companies in the comany list box.For our auxilliary view we decided to put two seperate information window to include the information about the event .The first window includes company information for which the event occures. The second window includes the event information.

At that time we were not sure what our dot size represent in our cloud lines. We boiled down to three features that we thought could be use to represent size of dots 1)Saturation 2)Impact Score and 3)Sentiment.

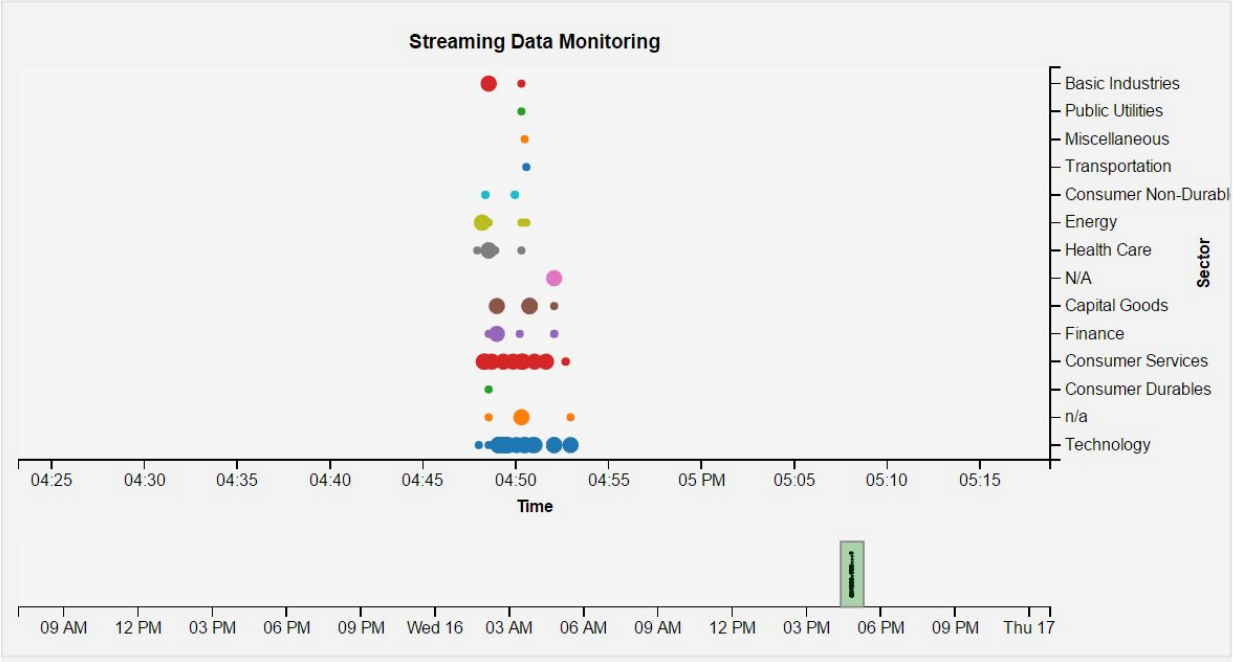
5.2 Second Mockup

We decided following Important points to consider for second mockup

1) Dot size represent story saturation value
2) Implement the zoom lens feature
3) Y-axis values

After getting feedback for the first presentation,we modified some of the questions and we decided to take saturation value of story to represent dot size in our cloudlines.

Screenshot 2



Description 2

Our first idea is to plot the cloudlines for all the companies in which user interested.But as the size for the list could be huge we might not be able to accommodate all companies into y axis.So,then we decided to put all the sectors(11 sectors)into y-axis.Our x-axis depicts the timeline for which user wants to see visualization.We were showing three days of the data ending with the day when application was started. We also implemented the focus lens to zoom to specific time bandwidth in the main view. This lens feature helps to see the events clearly when we have clustered events in cloudlines.

Some of the dots still overlap with each other especially for the events happening simultaneously.In that case the dots representing event with high saturation value covers the dots representing event with low saturation value. Because of this issue,we couldn't able to get details of the event for which the saturation value was low

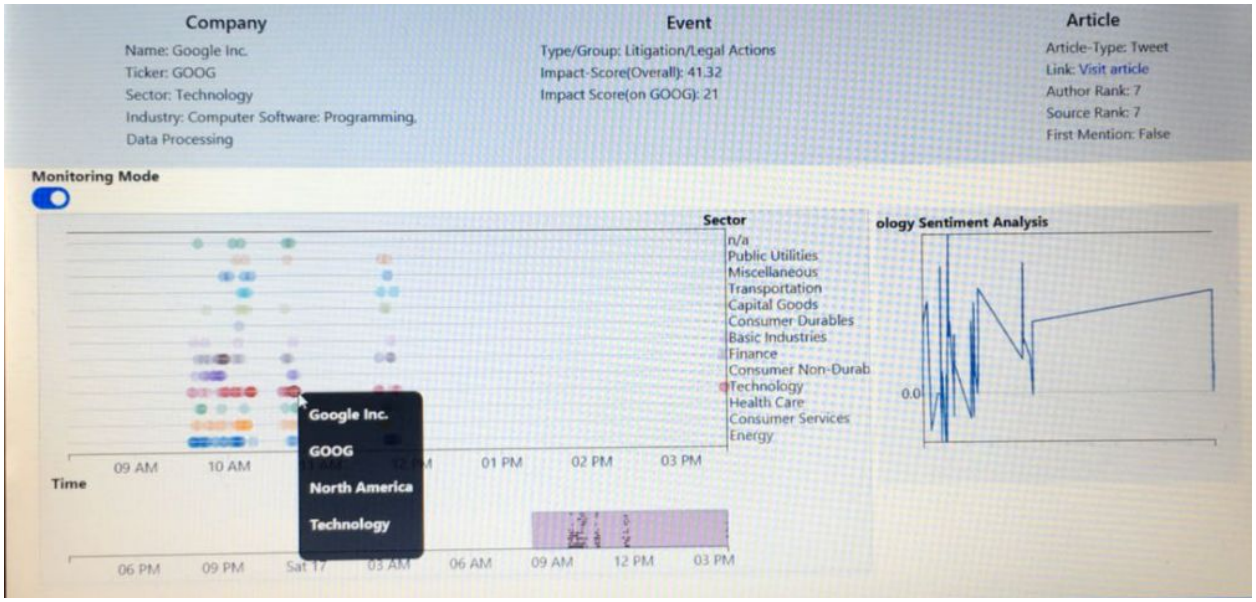
5.3 Third Mockup

We decided following Important points to consider for third mockup

1) Dot size should be same for all articles across all sectors.
2) Line chart for sentiment value for particular sector/company
3) Monitoring Mode Switch Button
4) Information Window for article

As, Saturation value was not good choice for our visualization,we read the cloudline paper where they described two complex function to determine size of the dots. We discussed the problem with professor and he told us that our project should be less interactive and more of the monitoring purpose. After getting guidance from mentor and professor we decided to take the size of the dots same for each article and for each sector/company.

Screenshot 3



Description 3

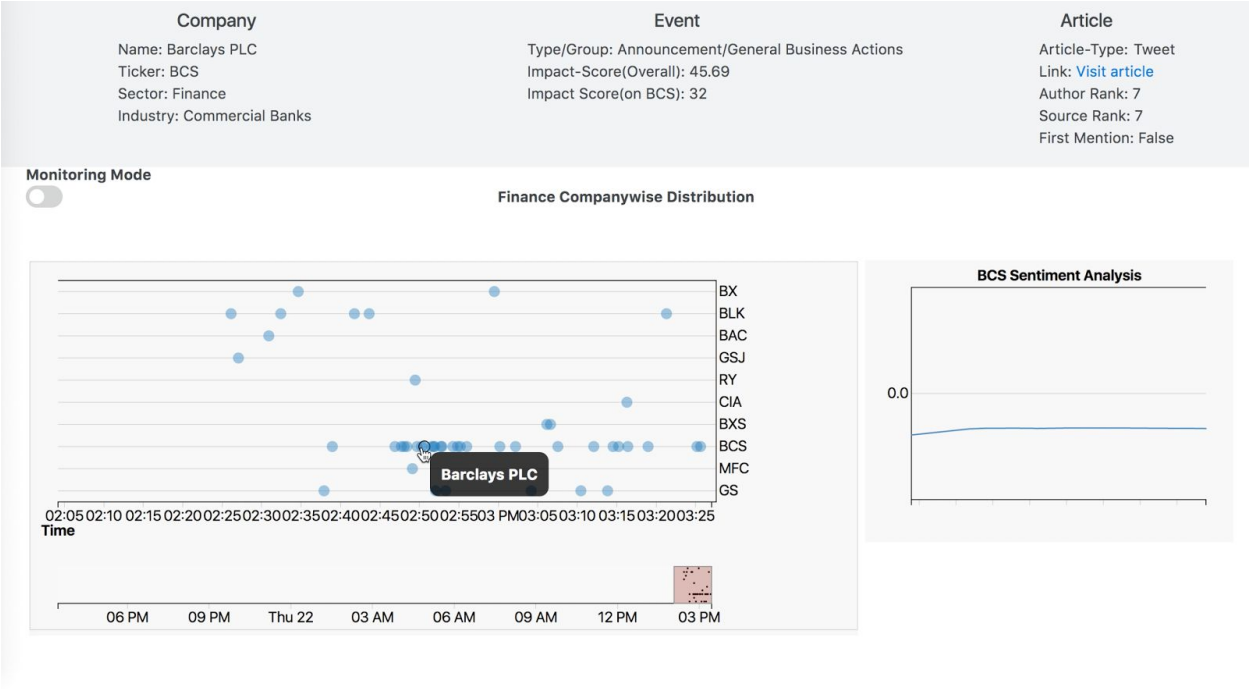
In our third mockup we decided to take size of each dot same for the purpose of easy monitoring. We also put the monitoring switch button.If user doesn't want the timeline to be

moved and interested into that specific time bandwidth, he/she can turned off the monitoring mode.Also to show the information for article we put information window as our auxiliary view on top of the main view. We also plot the line graph to show the sentiment value for the specific sector/company onto the right side of the main view.

On the line chart, the x axis represents a time line and y axis represents the sentiment, from -1 to 1. When we click a company/sector, the line represent the sentiment articles talking about this company/sector. This way we are able to find the trend and most talked about sector/company.

We took different color of dots to differentiate the companies/sector from one another. This was not required as we already mentioned the name of company/sector explicitly. This was came to know us by getting feedback from professor during final presentation. After then we decided to put the dots of same color in our visualization.

6. Final Visualization



In our visualization, the main view is the right bottom part. In the main view, there is a chart containing some dots aligned in horizontal lines. The dots represent the articles in our data. Dot size and color are fixed and mean nothing here. The x axis represent a time line, where each dot (article) has a coming time shown in x axis. The y axis represents companies/sectors, which the articles are talking about. There is also a scroll bar in the bottom, which represents a time range. Users can change the range from a month to a day, which will then show on the x axis.

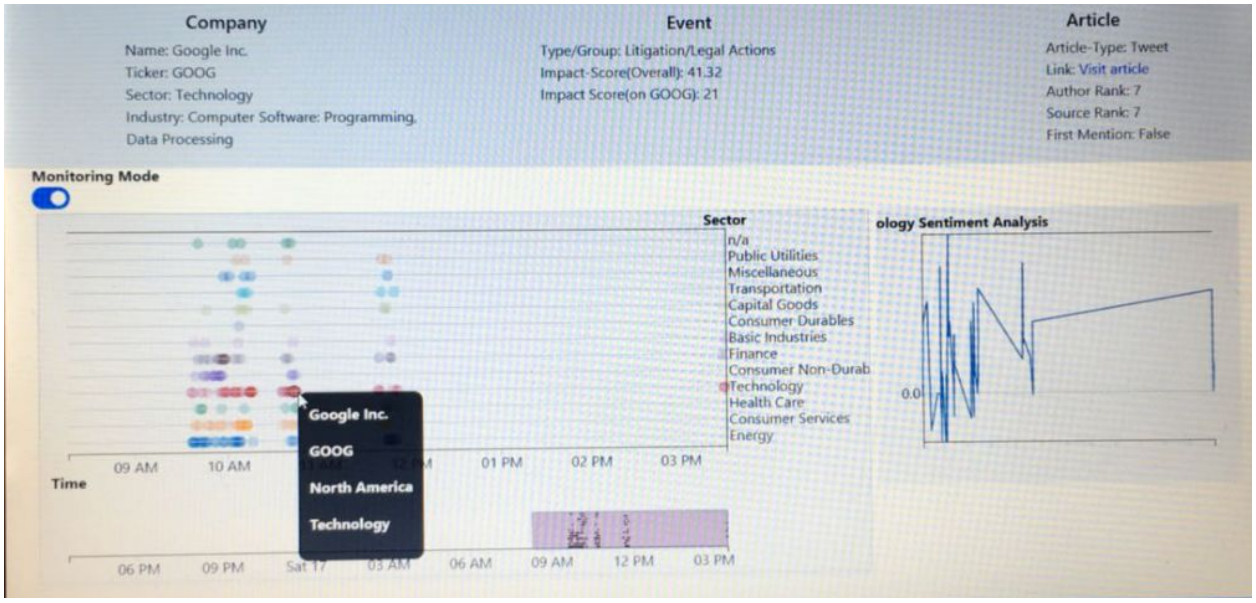
The auxiliary view on left is a line chart. The x axis represents a time line and y axis represents the sentiment, from -1 to 1. When we click a company, the line represent the sentiment articles talking about this company. This way we are able to find the trend.

The auxiliary view on the top is used to show the detail of each article. When we click a dot, it will show the detail of the article which the dot represents. This view allows users to take a closer look at each article. It contains lots of attributes of our data, among those there is a link to the origin article.

7. Findings

7.1 Question 1

Whether there is any major event or event episode for a particular sector?

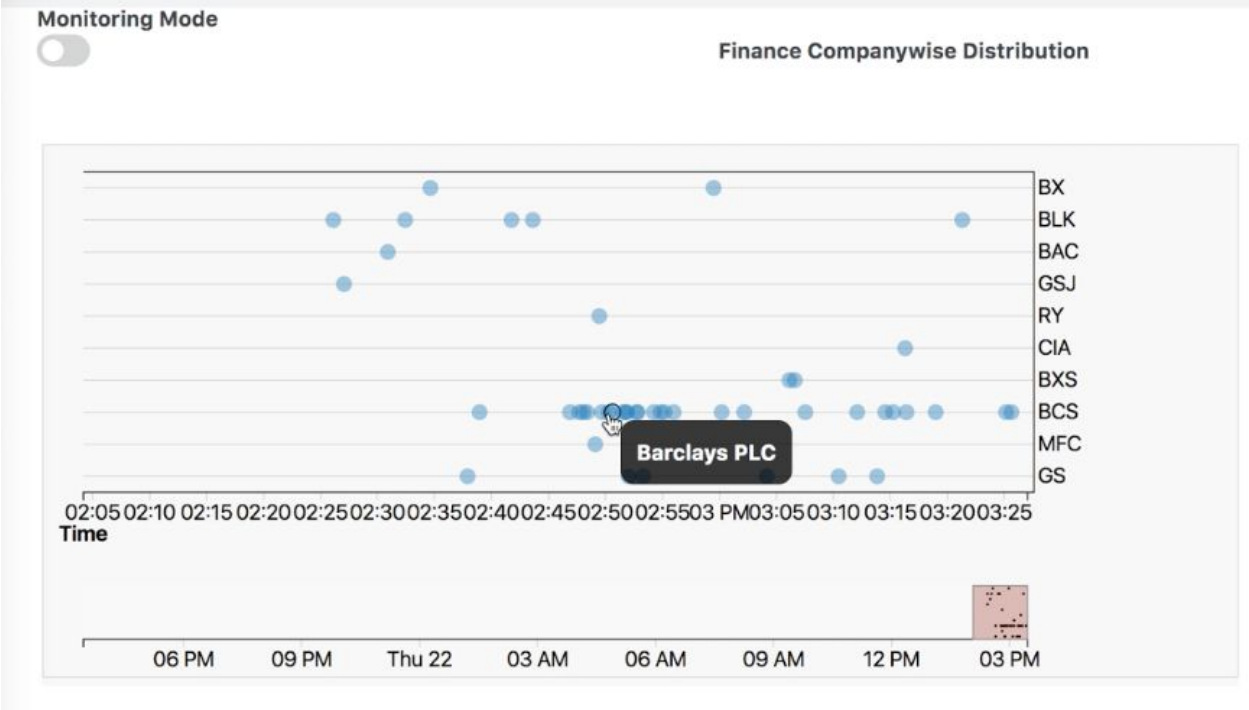


After start to run this application, it would display an initial view. It is a sector-wise cloudlines of 12 sectors in the dataset. The dots in the cloudlines represent an article about companies in this sector. When the dots become dense, they would overlap with each other so that the the dots around this point would be darker. User can detect the articles episode in this sector when they noticed the dark part in the cloudlines.

For example, in the picture above, the line for *Technology* sector has a group of dots which make there darker. User can hover the mouse on the dots, the window would display information about the dots there. Then user would also curious about what is happening inside each sector. This leads to question 2.

7.2 Question 2

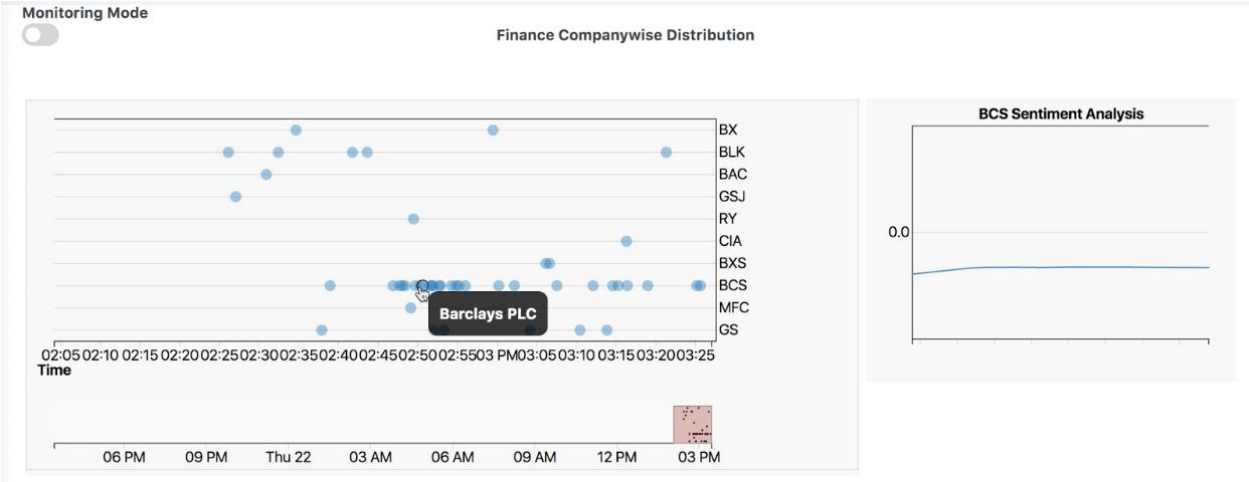
Whether there is a major event or event episode for top companies in each sector?



The application would automatically change the view. After display the cloudlines in sector-wise. The app would show top 10 companies within each sector. For example, we can detect that there are a lots of articles about company with ticker *BCS*. After hovering on the dots, the box would show the company name.

7.3 Question 3

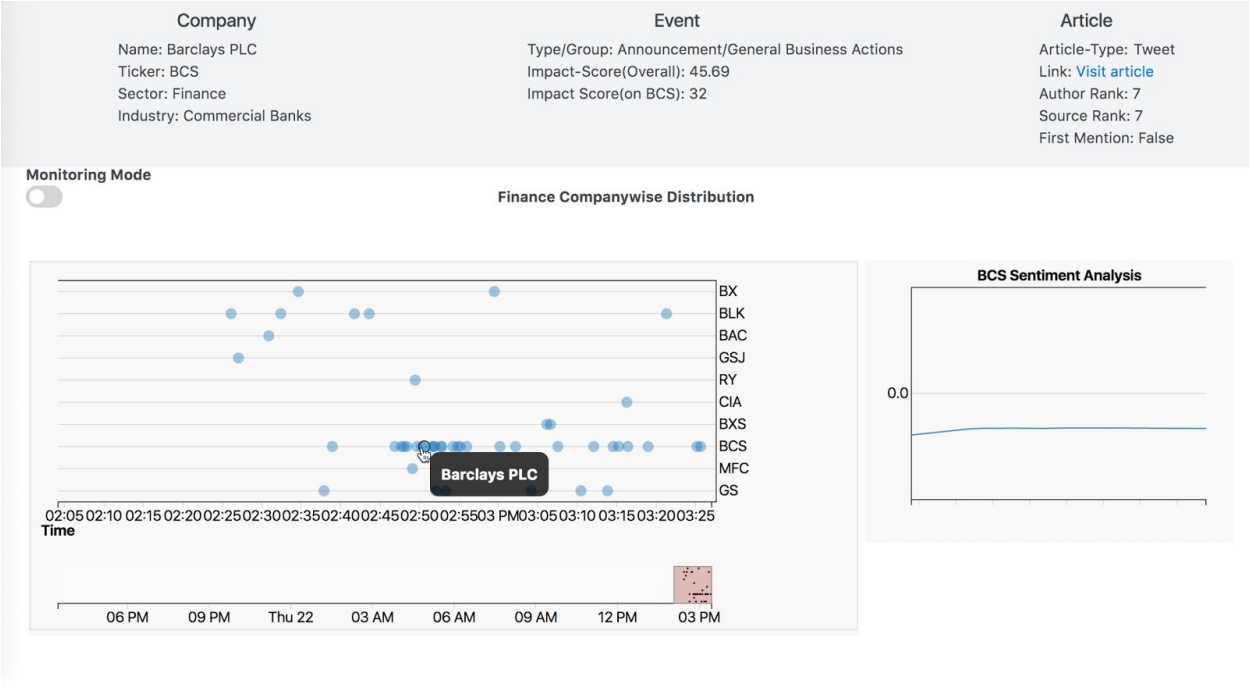
Are these articles' attitude positive or negative about this event and company?



After click the ticker of company in the Y-axis, the auxiliary view would display the sentiment line chart over time. Still take Barclays as example, the sentiment line of Barclays are all below the zero line, which means all articles are holding a negative attitude of the event happening to Barclays. And the trend shows that the sentiment was increased in the beginning, and then stay stable for a long time.

7.4 Question 4

What exactly has happened to this company?



By clicking the dot in the area of interest, the top bar would show the abstract information about the company, event and article, including the type or group of this event and its impact score. The top bar also provides link to the article so that users can check the article itself if they interested in details.

8. Limitations and Future Works

8.1 Encoding articles in more channels

According to the technique of CloudLines, the size and opacity of dots could also be used as additional channels to encode events. They are both based on the *Importance* function to calculate the value of size and opacity. The *Importance* function is related to the density of the events. The denser the events are, the larger and darker the dots are. In this way, the episode would be more obvious to be detected.

In our project, the data is streaming. It's not easy to implement the *Importance* value of the coming articles since it's hard to predict future density. But there might be some good solutions for this issue.

8.2 Visualizing more attributes from the data as auxiliary view

Accern API provided rich information about articles, events, companies and also many computed scores on different dimensions. We could have visualize more of them so that we can make fully use of the API data and help users to understand what's happening to sectors or companies better.