News Visualizer

An application to visualize and analyse the swiftness of news reporting by news sources in different topics.

Arun Govindaiah (ag5305@nyu.edu, ag5305)
Rahul Khanna (rk2795@nyu.edu, rk2795)
Vasudeo Deshmukh (vsd236@nyu.edu, vsd236)

Github: https://github.com/NYU-CS6313-Fall16/Visualizing-News-Reporting-2
Video: https://drive.google.com/open?id=0B0jLvFTLXH1ZckFYYm9jMnlEREU
Working Demo: https://nyu-cs6313-fall16.github.io/Visualizing-News-Reporting-2/Staticpage.html Yet to be deployed. Working with Cristian to resolve.

What is the problem you want to solve and who has this problem?

With the exponential growth of the internet, the number of news websites has grown tremendously. In this plethora of news sources, researchers have been using traditional techniques to identify importance of a website, or a story, that included visualizing in/outbound hyperlinks in the form of networks, an approach like Google's PageRank. The underlying assumption is that more important websites are likely to receive more links from other websites.

Still, there is very little understanding of how stories are reported over time by news organizations. The problem with the aforementioned method of understanding news reporting is that a growing number of websites lack in/outbound hyperlinks making it less than useful to use such a method. Also, another trend is that the news organizations are increasingly focusing on a narrow set of topics thereby doing away with the conventional way of being generic in nature while reporting.

Given these trends, there is a need of a way to understand and comprehend online journalism over time. By concentrating on aspects such as website's swiftness in reporting, reputation of the source, the number of articles, and the entities in the stories, we can arrive at a better technique of making sense of the new-generation news-reporting.

What are the driving analytical questions you want to be able to answer with your visualization?

1) Do some websites report faster than others in some topics?
    - Is there a consistency among these websites (Do they tend to be quick throughout) in their swiftness?
    - If so, which websites tend to report stories faster?

2) How to compare several websites, based on the topics they report and speed, over time?

- Given a topic, how do the news sources contrast others or how similar are they when reporting news in terms of promptness and quickness.

What does your data look like? Where does it come from? What real-world phenomena does it capture?

Attribute Name	Attribute Type	Meaning	Values	Derived
entities_name_1	Categorical	Name of the company	8,000-plus U.S. public equities	No
entities_sector_1	Categorical	Sector of the company	12 sectors	No
story_id	Quantitative	Unique ID per story	object identifier (OID)	No
story_volume	Quantitative	Number of articles related to a story	1 or greater	No
harvested_at	Quantitative	Date Accern received the article	yyyy-MM-dd'T'HH:mm:ss.SSSZ Z (UTC TIME ZONE)	No
Source Category	Categorical	News sources are put in 4 categories based on the time they report.	1)Fastest reported 2) News Source In Top 6 fastest reporters 3) Just Reported the news but not fast 4) Never Reported the News. (These are color coded in the application)	Yes
Source Score	Ordinal	A score given to the news sources based on the aggregate of its rank (derived from source category) for each month.	0 to 8	Yes
Source Name	Categorical	Name of the source that has published the article	String	Yes

What have others done to solve this or related problems?

Visualizing the swiftness of the news sources is relatively less popular than other forms of visualizing news scenarios. There have been attempts to visualize news and analyze it for

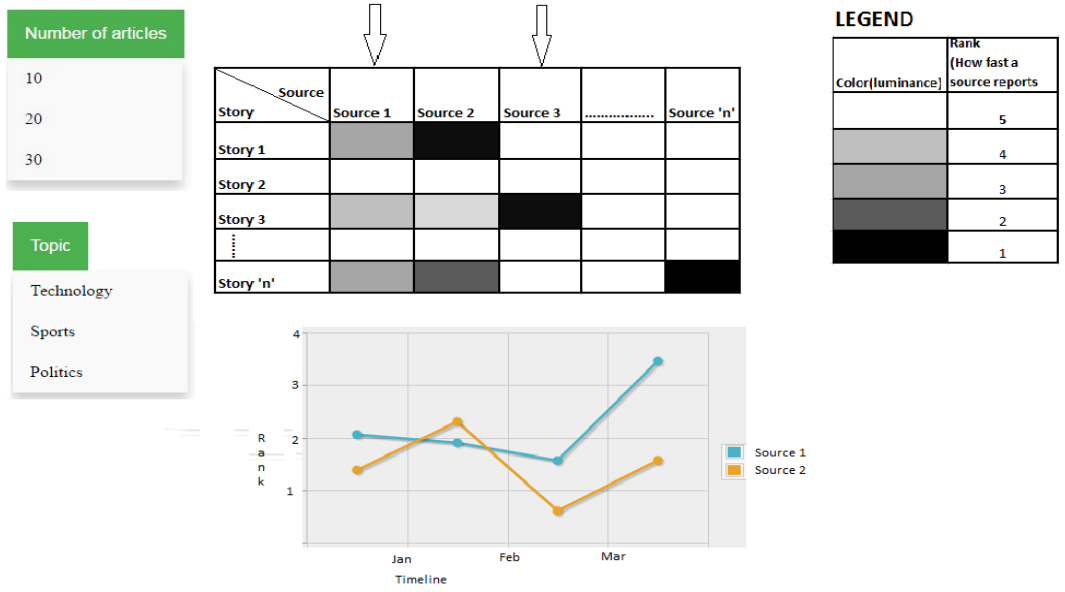
sentiments as soon as they are reported. Other projects include visualizing the current trends in news reporting. Below are some of the examples of the same.

NewsMap is an application to visualize the importance and popularity of news stories. This is an application that is using real time data. It displays importance of a news story in terms of its area.
http://newsmap.jp/
Another interesting application is trying to solve people's sentiments in different topics and how their outlook changes as the reports emerge. Here is a link to the project.
http://www.public.asu.edu/~ihsiao1/project/dataviz/index.html

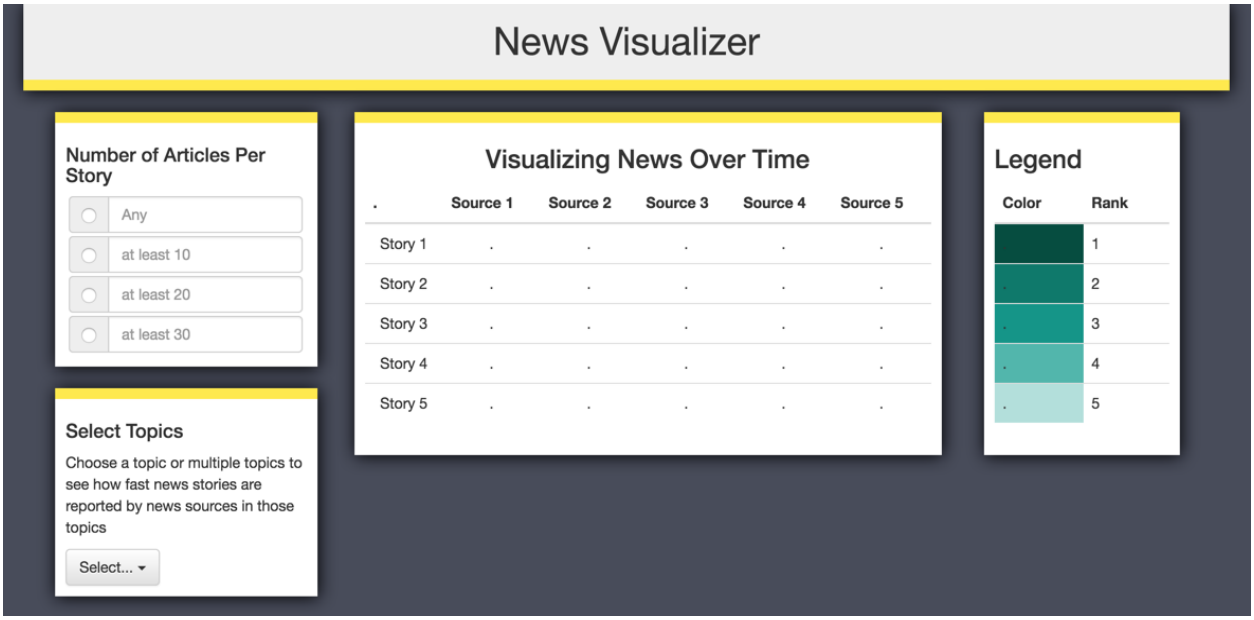
A Paper on Extracting News Stream for stock market trends -
http://link.springer.com/chapter/10.1007/978-81-322-3592-7_30
This paper suggests that the visualization of stock trends can be correlated with actual market prices. Their proposed analysis on current news stories helps to predict stock trends.

Design Iterations

Our initial mock up began with a sketch as shown in the figure below



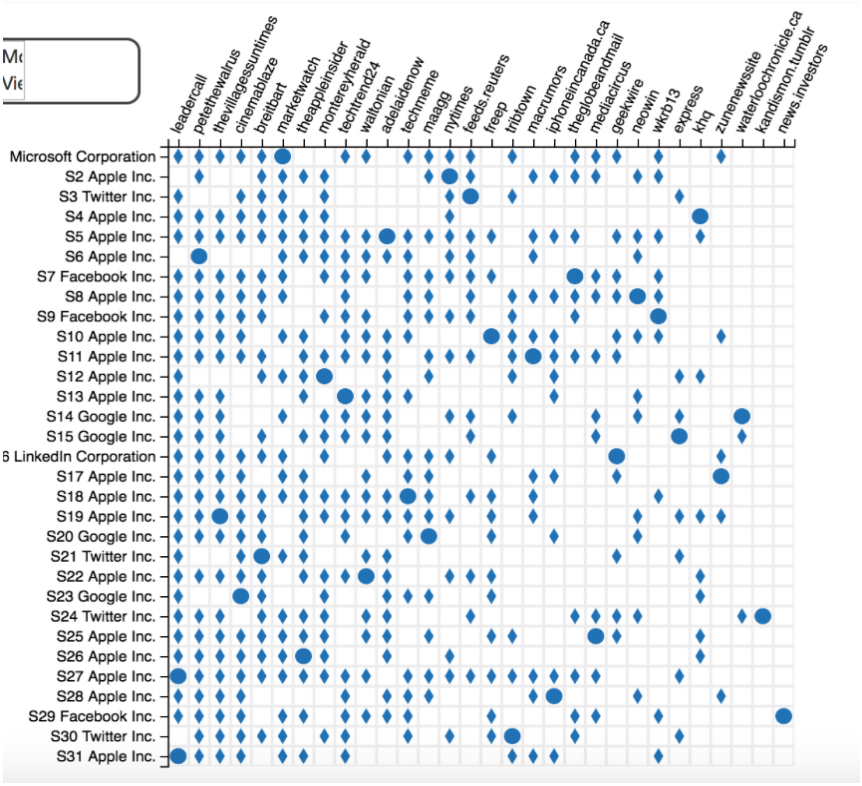
We proposed a main view with a heat map of sources vs stories. On the Y axis we have stories arranged chronologically and on the X axis we have sources. The cells are colored based on how fast they report news. The fastest news reporter will have the highest saturation and so on. We decided to take top five sources into consideration. In the auxiliary graph we sources compared over time. X axis is time and Y axis is the same source rank as it had in the main view. Based on that we had our skeletal UI that looked like below.



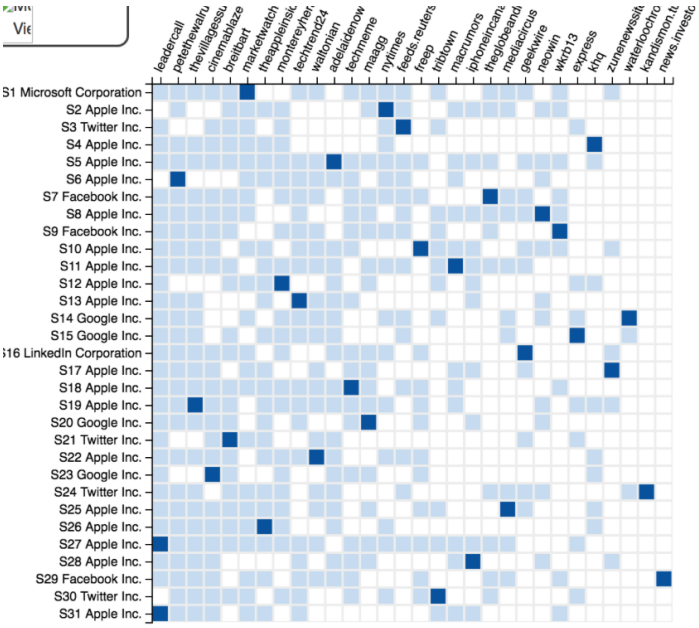
First Major Change - We realized that we cannot just take the first five sources and put them in our main view, because there are simply too many sources that report stories. There needed a way to filter out news source data. Else we would end up with a salt and pepper noise or a step-like graph from which we could have made little sense. Therefore, we decided to take news sources that have been fastest at least once for a given topic.

Second Major Change - The number of news articles per story was way higher than we anticipated, therefore we increased the filter to 100, 500 and 1000 articles per story. We then began seeing a pattern where certain news sources were actually faster generally.

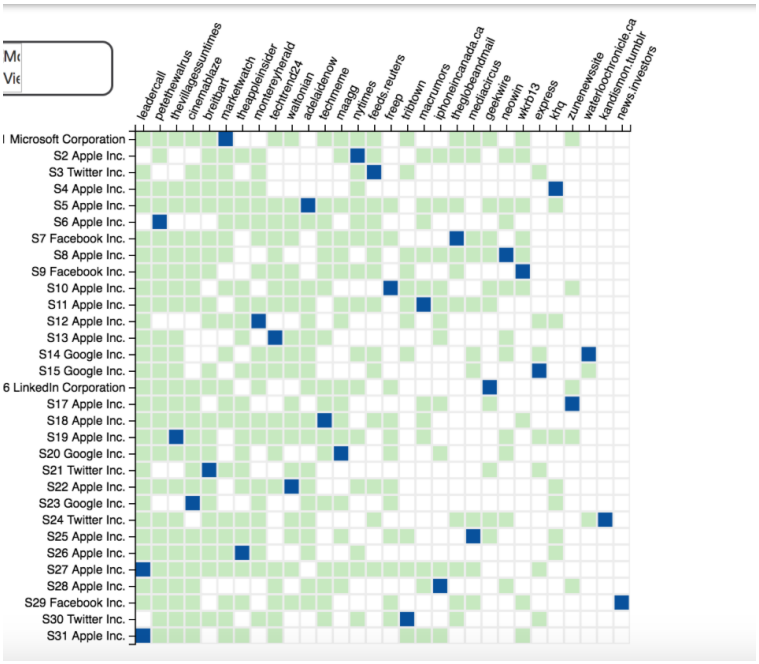
Third Major Change- We realized that our way of ranking news sources was not ideal. It did not take websites that were slightly less faster, but consistent, into consideration. Therefore we came up with a three bin approach. Here, we had the fastest, and just reported (but not fastest) and never reported as the three bins or categories. Since now we had only three bins, we thought shapes would give us a good visualization. This approach is seen below.



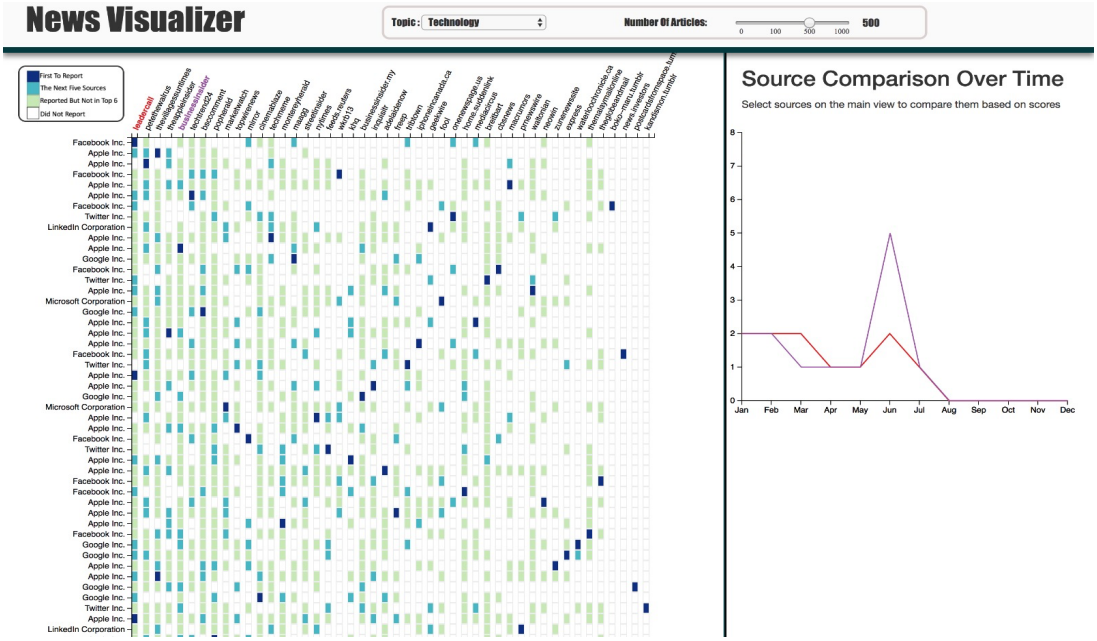
However, after discussions with mentor, we realized that having shapes is not the best idea since although our data is categorical, more emphasis should be on the fastest news sources. Therefore we came up with the below approach.



We used color saturation to signify our three bins, so that the highest saturation will be for the fastest reporters. However, when there is saturation one might think that we are representing some quantity and the darkest color signifies the largest quantity which is not what we wanted to display. We came up with different hues for different bins as below.



Fourth Major Change- The above mentioned approach was very convincing for us until we encountered some more problems. We had only 3 bins and news sources which slightly less faster than the fastest were put together with just reported group. This was detrimental for the overall aggregation where we try to find the fast news sources. Therefore we ended up with the visualization that was much better than all our previous iterations and more closely answered our analytical questions. Based on feedback, we also shrunk the size of each cell in width so as to streamline the visualization to focus on the columns separately and to reduce interference. The below screenshot depicts this approach.



Fifth Major Change- The squares were then changed to rectangular cells by shrinking the cells widthwise. The feedback we received for this change was critical since we are changing the cell structure that might not be the best way to visualize a heat map. Our final visualization eliminates this issue by making the cell squares again. This is explained in detail under Final Visualization.

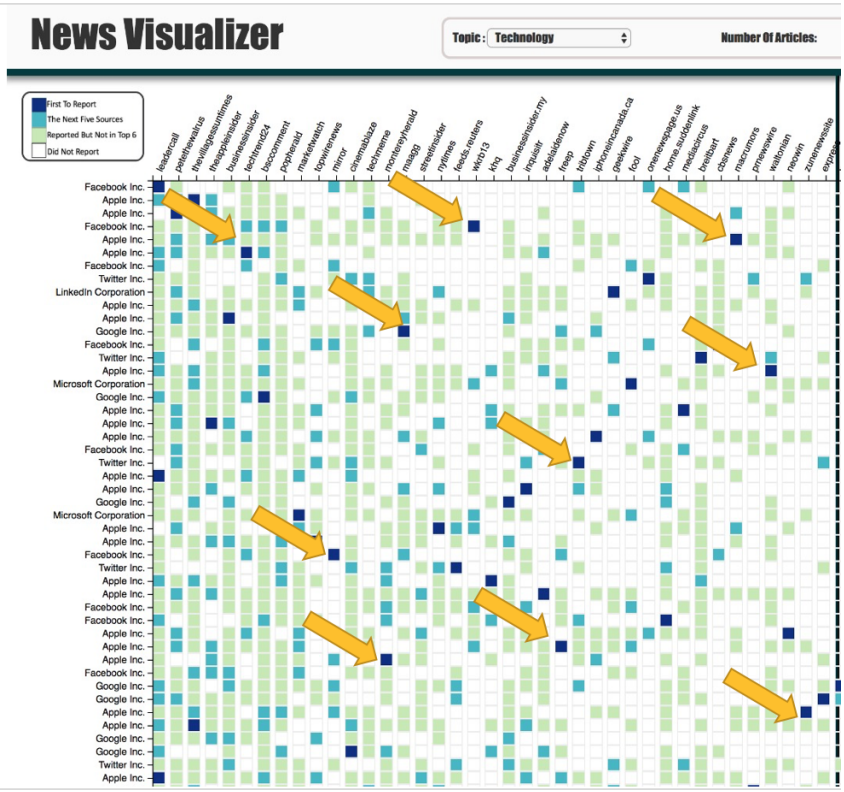
Final Visualization

The above image is our final visualization screenshot that shows the various options, Main View and the Auxiliary View. The main view is a heat map like visualization of sources and news stories. On the X axis we have news sources sorted according to their aggregate scores (calculated by summing up their ranks in individual stories). On the Y axis we have upto top fifty news stories on certain topic chosen by drop down menu. The user can also limit the news stories by the number of articles that reported it. There are four bins in our final visualization. The fastest news reporter, the next top five reporters, sources that reported the news but not in top 6 and those that never reported at all. We have a multi-hue color saturation encoding for this purpose.
Our auxiliary view is a line graph that compares the scores of news sources over a time period. These scores are derived by aggregating the individual ranks of news sources for stories for each month. We can click on the news source on the main view to add the source to the auxiliary graph and click again to deselect the same.

Findings

This application has thrown light on several interesting discoveries. Some of which are described below.

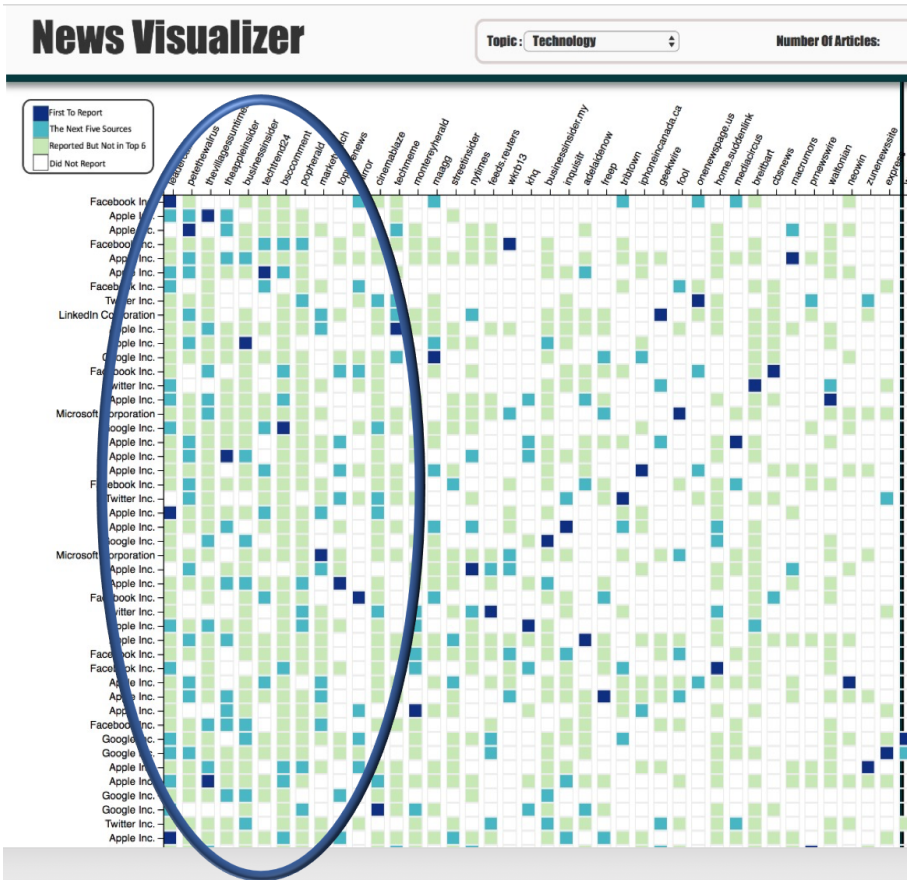
So, which are actually the fastest news sources?



If we see this carefully, the darkest cells are scattered all across (shown by yellow arrows) There is no concentration of the fastest news sources at any point. This has been the same for different topics.
This shows that only speaking in terms of absolute swiftness in reporting news all the time, there are no clear winners.

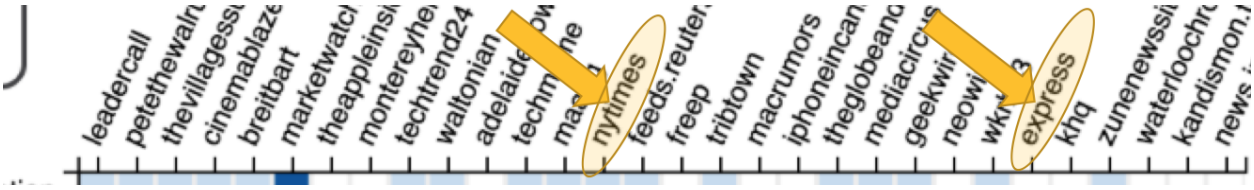
So, there are no consistent fast reporters at all?

Certainly, that is not the case, as we can we see from this picture below. We can see there is a concentration of darker cells on the left that indicates that there are news sources, despite not being absolute fastest, but are generally faster than other sources. Take a look at the concentration of darker cells on the leftmost columns. They are sparse in the right side.



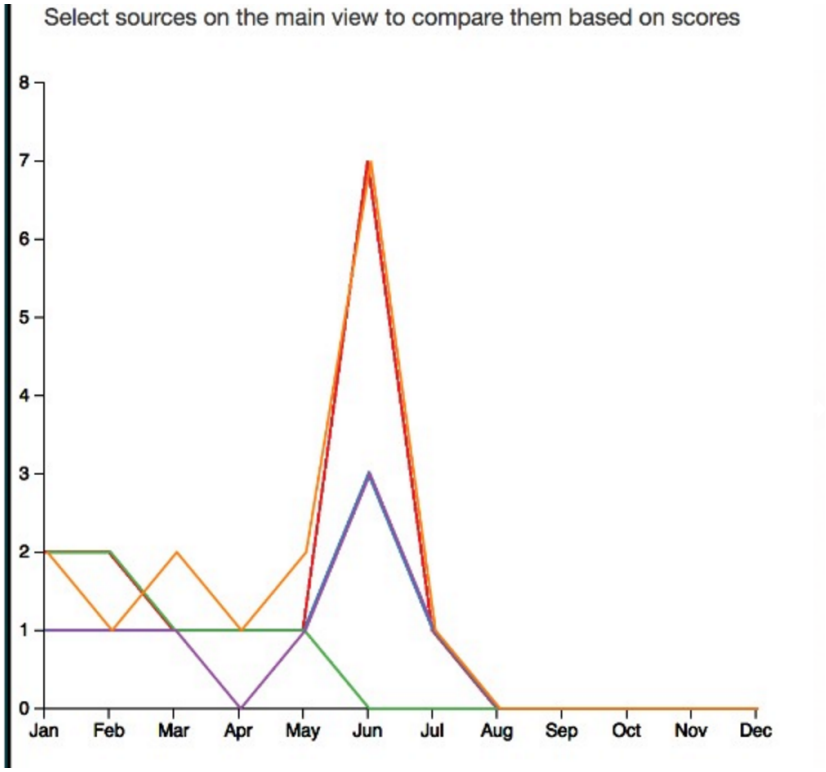
Are these sources that report news fastest very reputed in reality?

This is an interesting observation we made. The generally faster news are not always popular. And very popular news sources are not usually very fast in reporting stories. This shows that these sources take time in analyzing and checking their stories before reporting them. As we see in this image below, NYTIMES and EXPRESS are not in the left which are occupied by lesser known news sources.



How about the fast sources? Do they rank well throughout the time period?

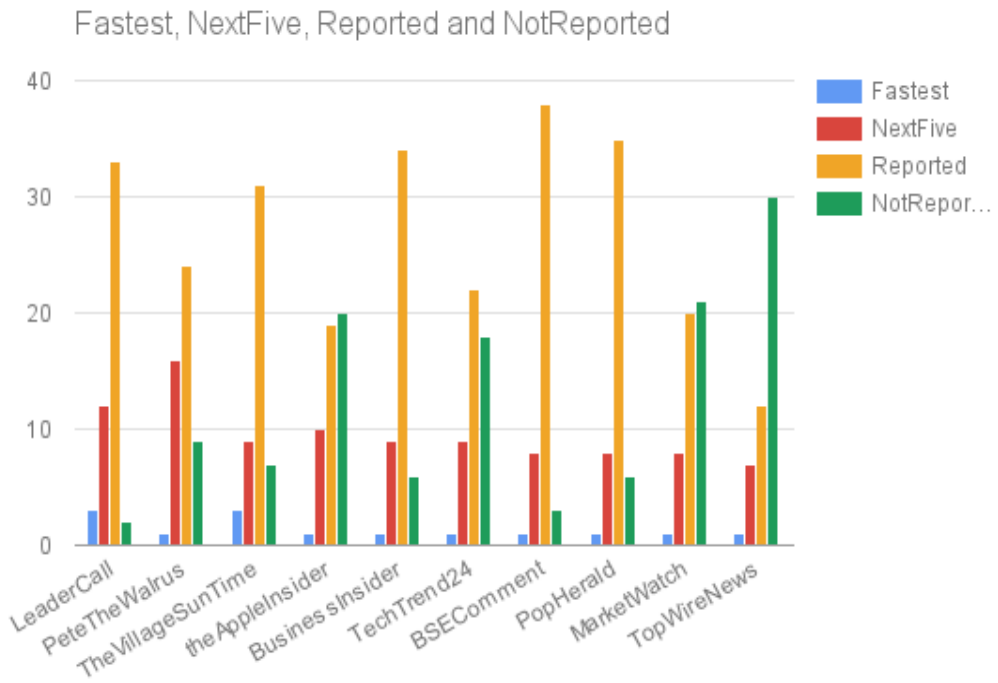
The auxiliary graph tells us another story. As we see this graph below obtained by selecting multiple top sources, no source consistently occupies the top spot. They do not fair the same all throughout the time period. There are occasional spikes and lows as we can see below.



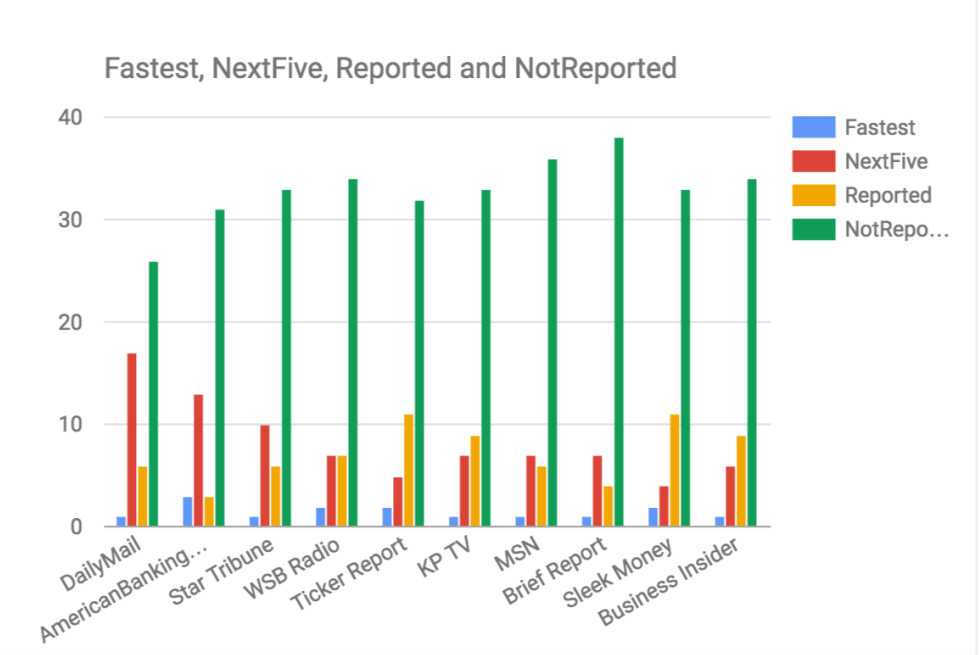
How about the same over different topics?

We see the trend in different topics when we visualize the aggregated ranks over topics and compare them with other topics. Here are two examples such visualizations. We have a bar graph that has the count of the ranks of the news sources against different ranks. The first graph is one such visualization for technology field for top ten sources. The second one is for finance sector. The interesting thing is that we see that the sources are usually faster but not always the fastest. We see a trend that is like a bell curve. The more left the maxima of the curve the better the news source is in terms of speediness of reporting news.

Comparison of News Sources in Technology sector.



Comparison of News Sources in the Finance Sector.



Limitations and Future Works

There was a lot of news data with us. We had to filter out a lot of data to focus on our main question. Our fear is that we might have lost some interesting patterns or observations while filtering out a lot of data. Also, at this point we cannot put in words the story name as it is limited by our data. Even, companies like google and facebook manually write story names for news stories. It would be better if we can extract story names from the data.

If there is more time, we would probably try and implement visualizations to view data across all topics at a time so we can see the trends at once and visually compare. In future, we would like to enhance the project by including the wide variety of more topics and find new ways to visualize much more data in the main view in a more meaningful and concise way.