Around the World: A visualization of global news sentiment

Team Members

Name	Department Affiliation	Role
Parul Batra	Course 15: Management Science	Project Manager
Christian Landeros	Course 20: Biological Engineering	Designer
Bethany LaPenta	Course 6: Computer Science	Developer

Concept Overview

Thousands of news articles are published around the world every day, that reflect the most important events and opinions in the world at any given time. The effect of news media on society is direct and ever increasing. In today's digital world, we absorb media content every minute of every day from a number of sources and devices. This constant flurry of news articles and opinions has a profound impact on how we feel, how we think, and how we act.

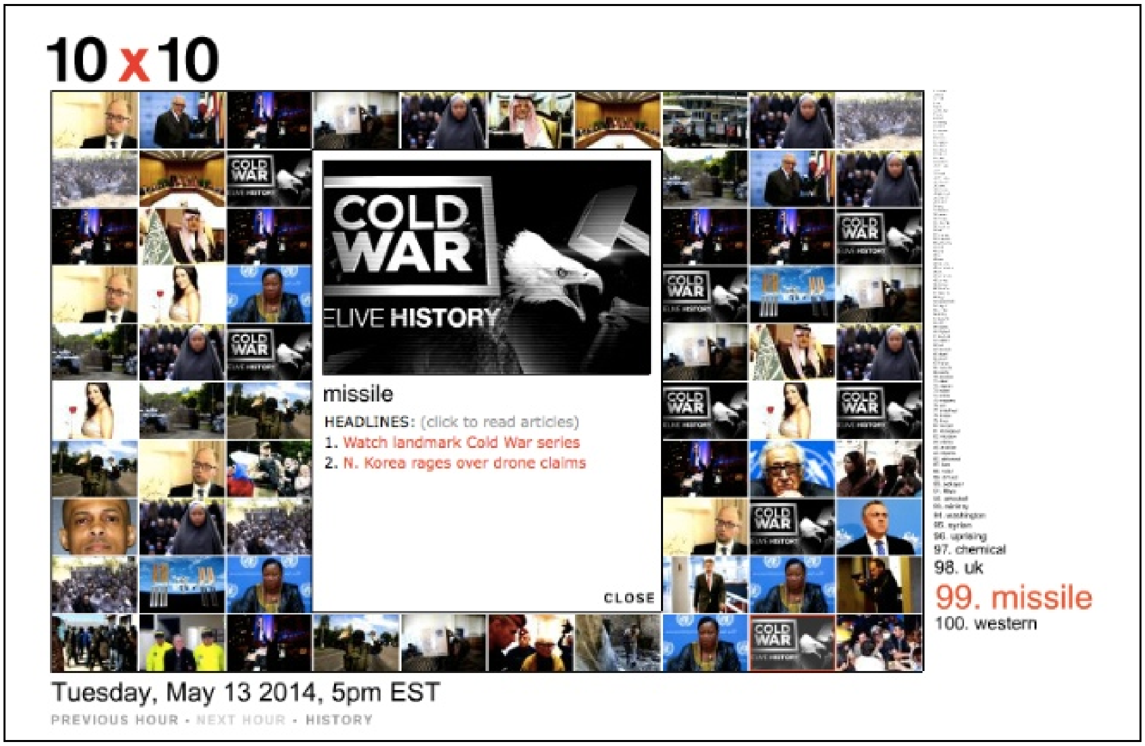
Given the large amount of important content in news media, computer scientists and mathematicians have developed several tools to analyze text data in order to extract important insights. In the past few years, news analytics has emerged as a powerful and widely used tool in the financial sector, where banks analyze news data to extract relevant insights about the stock market to make smarter decisions about trading stocks and managing portfolio risk. However, the application of text analytics on news data to understand how it impacts society and global sentiment has been limited.

Around the World is a visualization of the sentiment (positive or negative) of the most important news articles published every day. Through this project, we hope to capture and understand the sentiments that surround us everyday in the form of news articles, and affect our views about the world, specific events, and people. The sentiment in a news article is informed by the type of event the article is covering, but also by the personal opinions and style of writing used by the journalist. We believe that this piece of work will be helpful for media studios and journalists in evaluating the sentiment of the content they share with society. We believe this project will also be of interest to government organizations who will very quickly and easily be able to see the sentiment generated by a specific policy decision. We have used data from the 10X10 API and NYTimes API for this project.

Background Research

We were inspired by 10X10 (http://tenbyten.org/10x10.html), a project by Jonathan Harris, a digital artist from Vermont, that collects the 100 most important words and images from six global news sources every hour. The project visually presents the data as a single image that encapsulates that moment in time. Over the course of days, months and years, this project leaves behind a trail of images that describe the state of the world at any given time in history.

We looked into how the data was collected, analyzed, and visualized in the 10X10 project. We learned that the input data is collected from six global news sources(CNN, ABC, Reuters, BBC, MSNBC, The Guardian), and text analytics is used to determine the top 100 most frequently occurring words (excluding common words like ‘but’, ‘at’, ‘the’, etc.). The projects then pulls images related to the articles featuring each of the the 100 most important words, and displays them together as one single interactive structure.



10 x 10

missile

HEADLINES: (click to read articles)

1. Watch landmark Cold War series
2. N. Korea rages over drone claims

CLOSE

94. washington
95. syrian
96. uprising
97. chemical
98. uk
99. missile
100. western

Tuesday, May 13 2014, 5pm EST

PREVIOUS HOUR · NEXT HOUR · HISTORY

Image 1: 10X10 visualization of 100 most important words and images every hour

Upon analyzing this project, we were intrigued by how these words and images affect society. We wondered if it would be possible to capture the sentiment behind these words and images, to create a trail of global sentiments that would form a continuous tapestry of human life.

We were also inspired by the progress made in news analytics and sentiment analysis in sectors such as finance, risk management, branding, etc. For example, companies like Thomson Reuters and Bloomberg analyze news data to help their clients make smarter investment and trading decisions and identify new opportunities. A number of technology startups today analyze twitter data feeds and blog posts to understand the sentiment expressed by consumers towards particular brands. These companies have built sophisticated sentiment analysis models, but use very basic visualization techniques to display the data

.

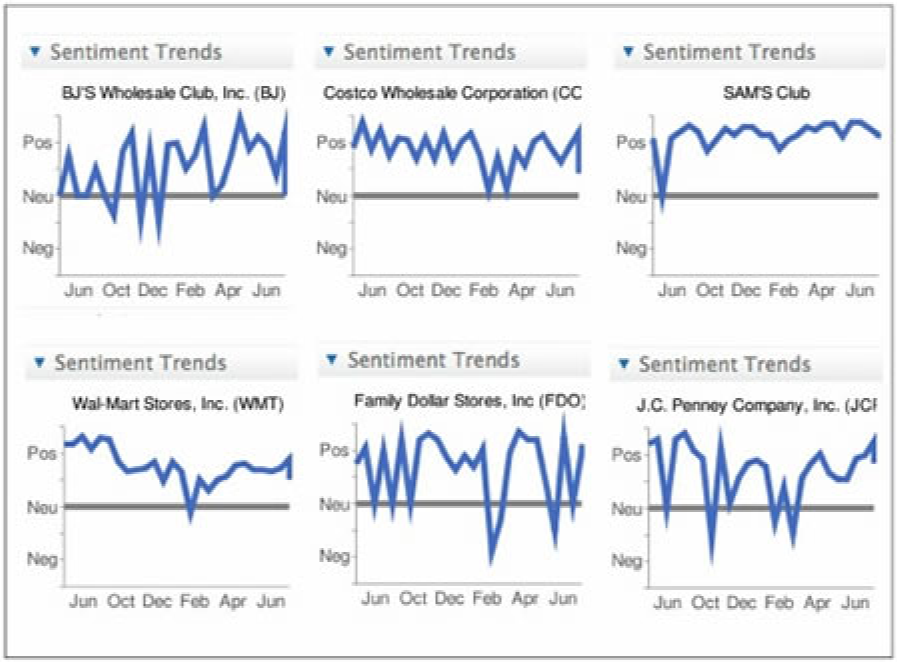


Image 2: Visualization of sentiment analysis on news data about corporations

Another area of research that we looked into was the Open Gender Tracker project developed at the MIT Center for Civic Media. Open Gender Tracker is an upcoming suite of open source tools and APIs that will make it easy for newrooms and media monitors to collect metrics and gain a better understanding of gender diversity in their publications [http://opengendertracking.org]. The

MIT team has developed analytical models to compute metrics from news articles that define gender balance, and display them in the form of insightful charts and bar graphs.

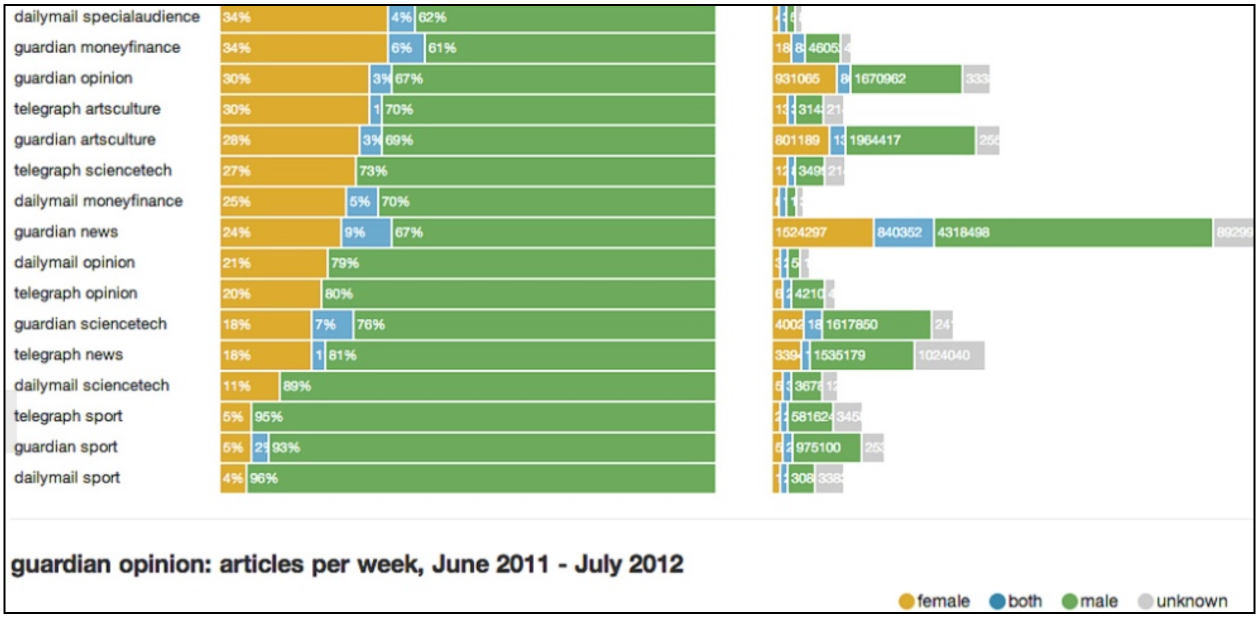


Image 3: Visualization from the Open Gender project at MIT Centre for Civic Media

We were inspired to use the recent developments in text analytics to answer important questions about the sentiment that surrounds us everyday in the form of news media:

- On an average, what is the sentiment in the news articles that we read everyday?
- To what extent is the sentiment in news articles driven by the event/topic they are covering, versus the opinions and writing style of the journalist?
- Can we create a continuous timeline of global news sentiment that leaves behind a trail of emotions that would represent a continuous tapestry of human life?

We believe this analysis will be helpful for a number of audience including media houses and journalists, who will quickly be able to evaluate the sentiment that their work adds to the world. It will also be helpful to government organizations in understanding the sentiment that particular laws or policy decisions are generating in news media. And finally, we believe this will be of interest to social scientists and all others who are interested in understanding the impact of news media on society.

Project Development Process

Roles

Given our diverse backgrounds and interests, we assigned roles to leverage each individual's members strengths and interests. Bethany, who is studying Computer Science and is proficient in backend programming, was responsible for all the back end programming required for this project, including pulling data from the 10X10 and NYTimes API, running sentiment analysis on the news articles, putting together the final dataset that was used to generate the visualization, and making a few intermediate prototype visualizations. Christian, who is studying Biological engineering and has a strong interest in design and visualization, came up with multiple visualization ideas for this project and was responsible for creating the final visualization. Parul, who is currently pursuing her MBA and was previously a management consultant, brought project management and strategy skills to the team. She was responsible for creating the written deliverables, identifying the data sources to be used, coming up with visualization ideas, and shaping the overall direction of the visualization project.

Process

Our team started by researching existing news sentiment analysis projects, and clearly defining the questions that we are trying to answer through our project. We decided to build a timeline showing global news sentiments and their drivers, for every hour in 2013. Upon deciding the goal of the project, we looked into various data sources that could provide us with the news data we were looking for.

Data source identification

Inspired by the real-time visualization in the 10X10 project, we decided to use the top 100 words captured by the 10X10 project every hour as our basic data source. To be able to conduct sentiment analysis, we needed text i.e. news articles related to these words. We looked into several news websites such as CNN, as well as news aggregators such as Google News, but found that most data sources either required payment, did not have APIs, were difficult to use, or had very strict rate-limits. Finally, we decided to use the New York Times Article Search API to find articles about the top 100 words from 10X10, because the NYTimes API was free, easy to use, and had a reasonable rate limit (10,000 API calls per day).

Data collection and cleaning

Upon finalizing the data sources, we wrote a shell script to pull the top 100 words from the 10X10 API. For each of these 100 words, we pulled news articles from the NYTimes API that included that word. Due to restrictions imposed by the NYTimes API, we were able to pull at most ten articles for each of the hundred words, as well as analyze only the summary paragraph since there was no way to access the site's article in plaintext.

Our initial plan involved running sentiment analysis on NYT articles featuring the top 100 words generated by the 10X10 data source for every hour in 2013. Due to restrictions imposed by the NYTimes API, as well as how long the shell script took to execute per day, we were unable to pull data from their API for every hour in 2013. Instead, we were able to pull articles once for every week in 2013. As a result, we decided to pull the top 100 words from 10X10, and up to ten NYT news articles for each of the hundred words, for the 1st, 7th, 14th, 21st, and 28th of every month in 2013. At the end of this step, our outputs were:

- 100 most important words for the 1st, 7th, 14th, 21st and 28th of each month in 2013 from the 10X10 API
- For each of the 100 words, up to 10 NYT news articles from the same day containing the word fetched from 10x10.

Sentiment Analysis

Next, we ran the news articles for each day through a sentiment analysis tool. After looking through a number of available sentiment analysis tools on the web such as Luminosity and ML Analyzer, we decided to use the Free Natural Language Processing Service (https://www.mashape.com/loudelement/free-natural-language-processing-service#!) for this project because it is free, easy to use, and effective according to user reviews.

The sentiment analysis tool processed each article and generated a score between -1 and 1 for each article, where -1 means a 100% negative sentiment, 1 means a 100% positive sentiment, and 0 means a neutral tone. We then averaged the sentiment of all articles for a given word, to calculate the sentiment for each of the 100 words. Finally, we averaged the sentiment of each word for a given day, to calculate the overall sentiment for the day. At the end of this step, we had a sentiment score for each of the news articles, each word, and each day in our dataset.

The shell script and created API were designed to be dynamic and self-updating. The plan was for the visualization to be able to run daily on a cronjob in the future and be able to update itself without any human interaction, and effort was put into producing a robust, fail-safe, and error-handling backend implementation.

Final outputs from this step are:

- Produced API Here, in a format similar to 10x10's
- A JSON representation created in PHP that will dynamically parse the results in the folder and put it in a JSON representation, again completely self-updating and generated on call. (Located Here)

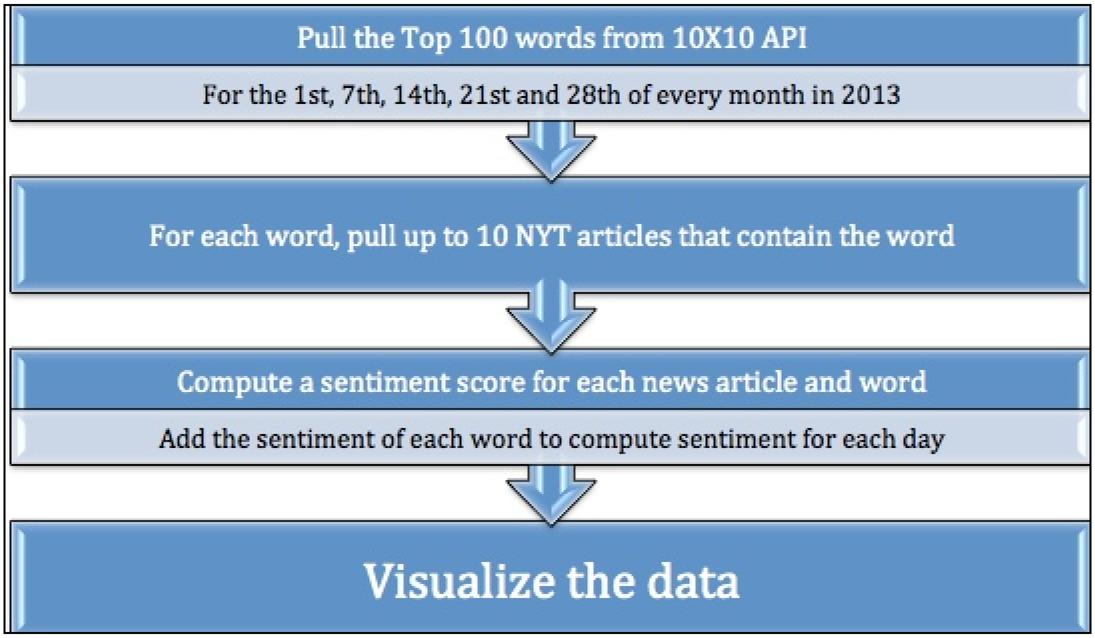

Pull the Top 100 words from 10X10 API
For the 1st, 7th, 14th, 21st and 28th of every month in 2013
For each word, pull up to 10 NYT articles that contain the word
Compute a sentiment score for each news article and word
Add the sentiment of each word to compute sentiment for each day
Visualize the data

Image 4: Summary of steps followed to create visualization

Developing the visualization

The final step in our project was to visually represent the data we had collected, in a way that produced meaningful insights for our audience. We came up with several ideas for visualizations, and debated which ones would be the most intuitive and provide the most insights to our audience. While we came up with a number of ideas, we were constrained by our skills and the timeframe of the class, since none of our team members were familiar with Javascript, which is the programming language needed to build interactive data visualizations in D3.js, our tool of choice.

Ultimately, we decided to build a timeline visualization showing the overall sentiment for the day with an interactive callout box giving important information about the words that were driving the overall sentiment for the day.

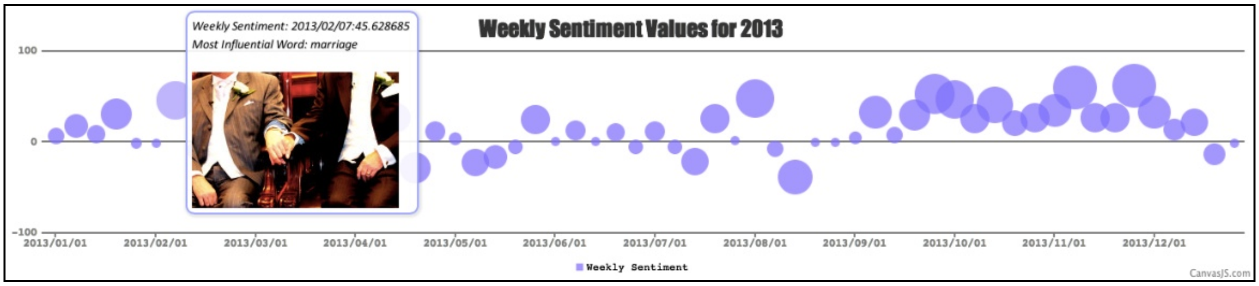


Image 5: Timeline of news sentiment by week for 2013

Enabling Technologies

A number of technologies have enabled us to build this visualization. The 10X10 project API and the NYTimes API provided us with the dataset we needed for the sentiment analysis. The Free Natural Language Processing Service enabled us to perform sentiment analysis on the news articles. We used a shell script to pull and manipulate the data. Finally, we used D3.js and canvasjs.com to build the final visualization. During the course of the project, we used a virtual server created through xvm.mit.edu.

Journey Map:

This section walks through the different ideas for visualizations that our team came up with at the beginning, and then discusses the one we ultimately decided to build.

At the beginning of our project, our team ideated and came up with three compelling visualization ideas. These were the calendar view, timeline view, and the sunburst view.

The calendar view allows the user to scroll through time and view news sentiments over time, where the overall sentiment for a day is represented by its color e.g., light blue for negative sentiment, dark blue for a positive sentiment. If the user hovers their mouse over a specific day,they would be able to see the most negative and positive words for the day, and the associated articles that drive the sentiment for the day. The use would also be able to zoom in or zoom out of the calendar, depending on whether they want to view sentiments for a week or for a year.

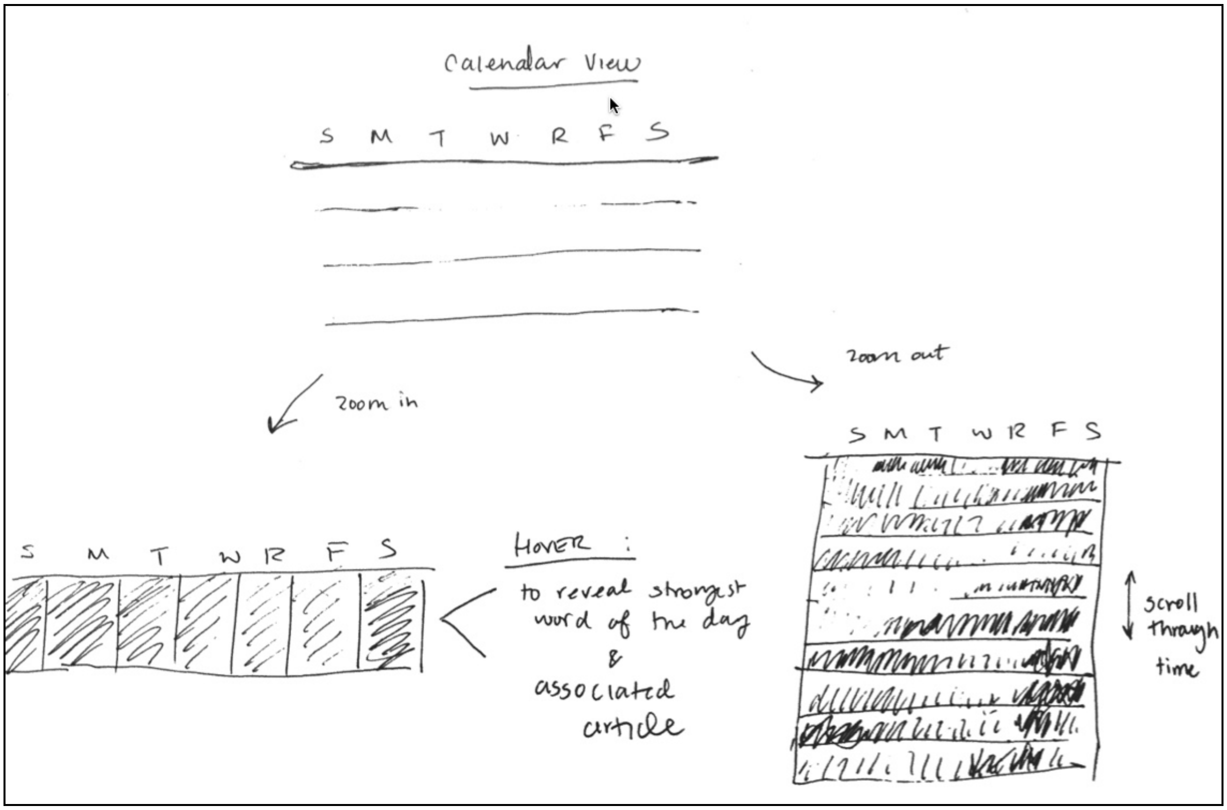


Image 6: Calendar View

The timeline view would show a continuous timeline of sentiments, and display the most negative and positive words for each week in the form of a bubble near the timeline, where the size of each bubble would depend on the intensity of sentiment associated with that word. If the user hovers over a given point on the timeline, an interactive callout box would appear and show them the most negative and positive words for the day, along with links to the related news articles.i

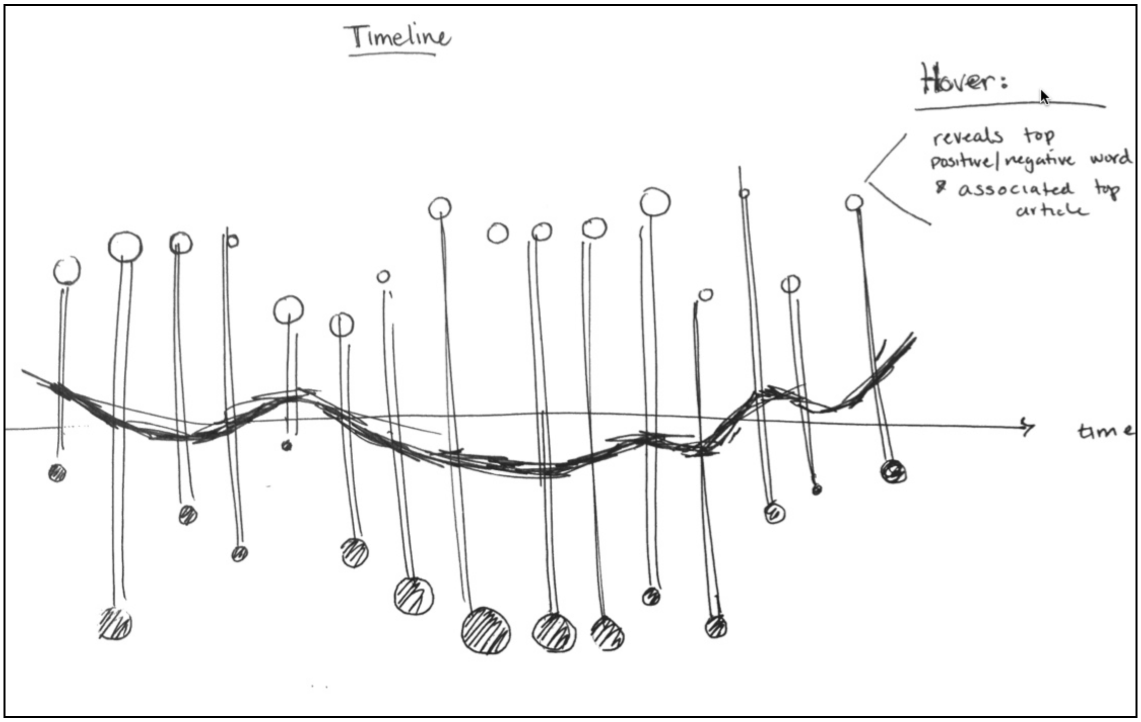


Image 7: Timeline View

The sunburst view is a unique way of showing the timeline in a circular manner, where the color of the sunburst would represent the sentiment score for the week, and the height of the sunburst would depend on the sentiment score of the most negative/positive word of the week. If the user hovers over a given week in sunburst, a callout box would reveal the most negative and positive words for the day along with the related news articles that drive the overall sentiment for the day.

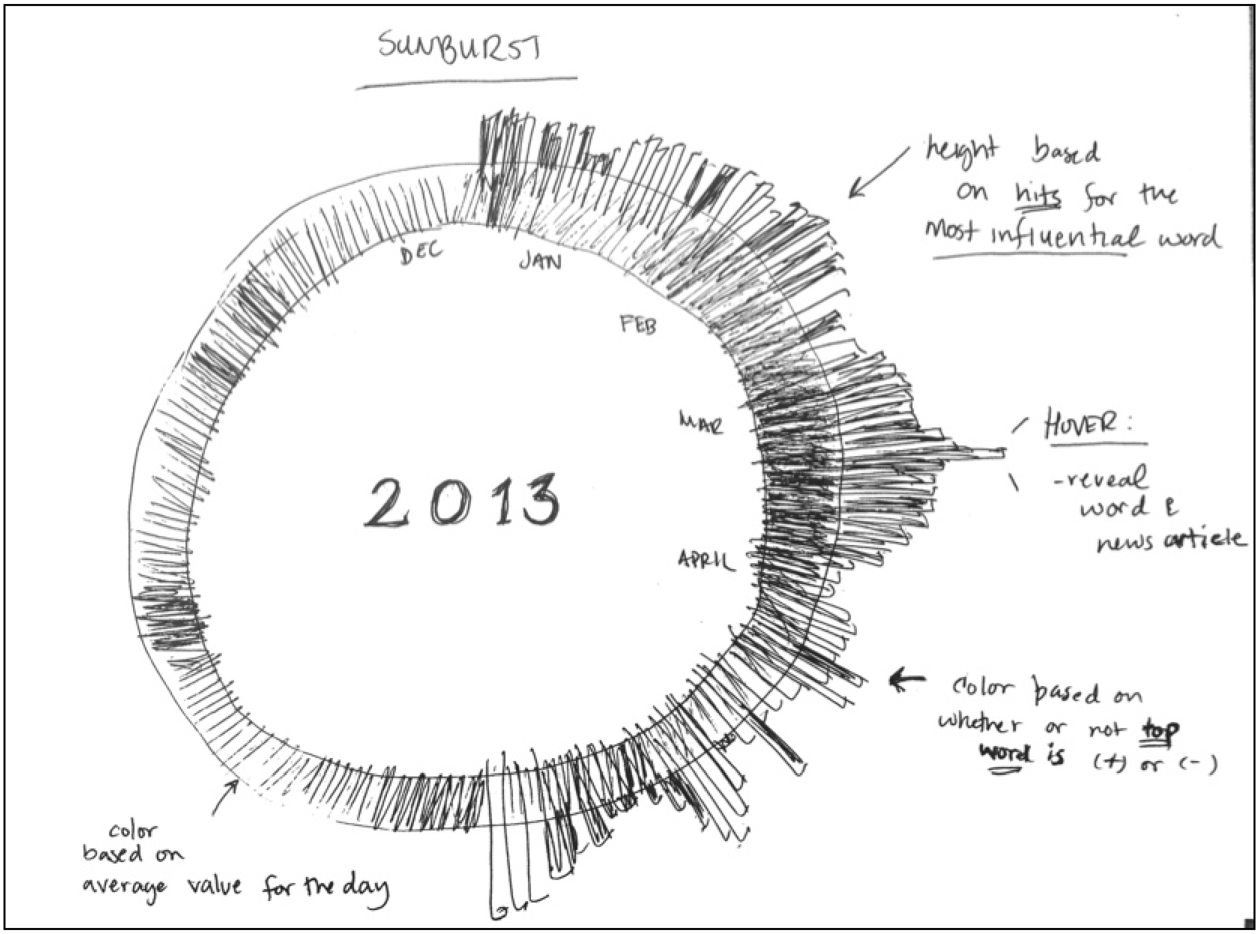

SUNBURST
height based on hits for the most influential word
DEC
JAN
FEB
MAR
APRIL
2013
HOVER:
-reveal word & news article
Color based on whether or not top word is (+) or (-)
color based on average value for the day

Image 8: Sunburst view

Ultimately, we decided to move ahead with the timeline view because it seemed the most intuitive for our audience, and also the most technically feasible for our team. Our final visualization is a timeline of news sentiments for calendar year 2013, where each point in the timeline represents the average sentiment for the given week (starting on the 1st, 7th, 14th, 21st and 28th of each month). When the user places their mouse over a given point in the timeline, an interactive callout box appears that displays the most important pieces of information about that week's sentiment such as, overall sentiment score, most positive words that week, most negative words that week, and an image depicting the sentiment that week.

Although in the long term, we would provide the user with the option to zoom in or zoom out of the visualization, for now we have created a few different versions of the visualization showing different levels of detail.

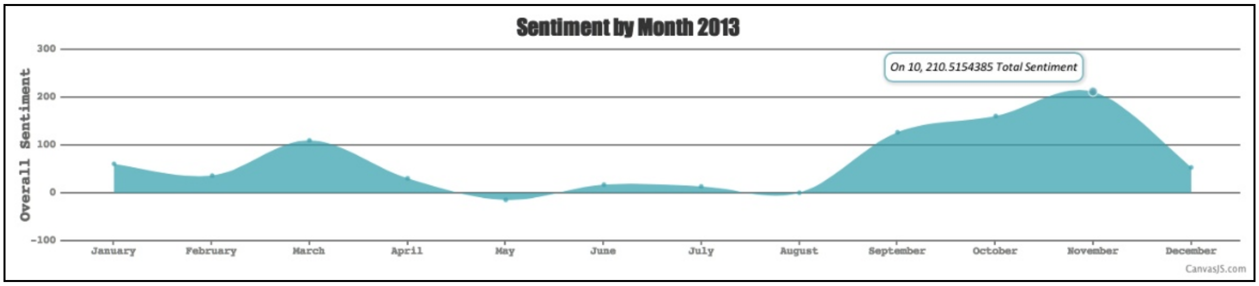


Image 9: News sentiment in 2013 by month

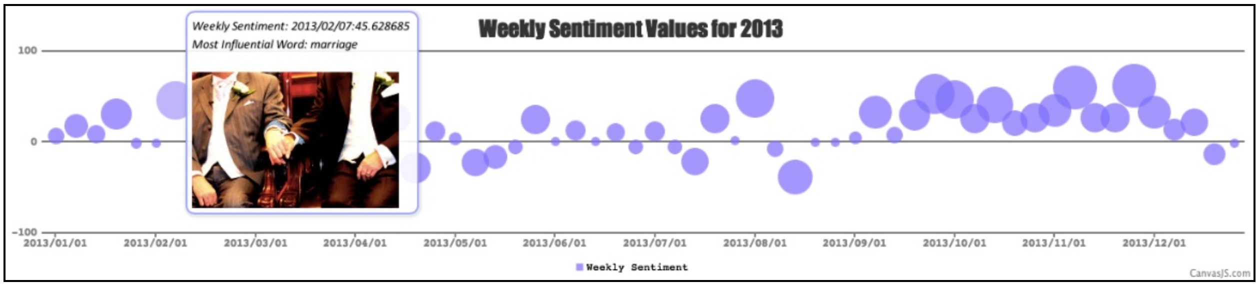


Image 10: News sentiment in 2013 by week

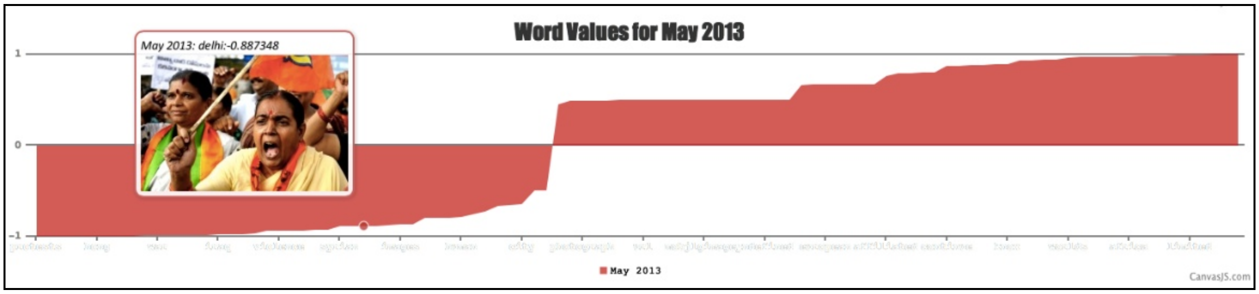


Image 11: News sentiment on 5/1/2013, with the top 100 words and images for the day

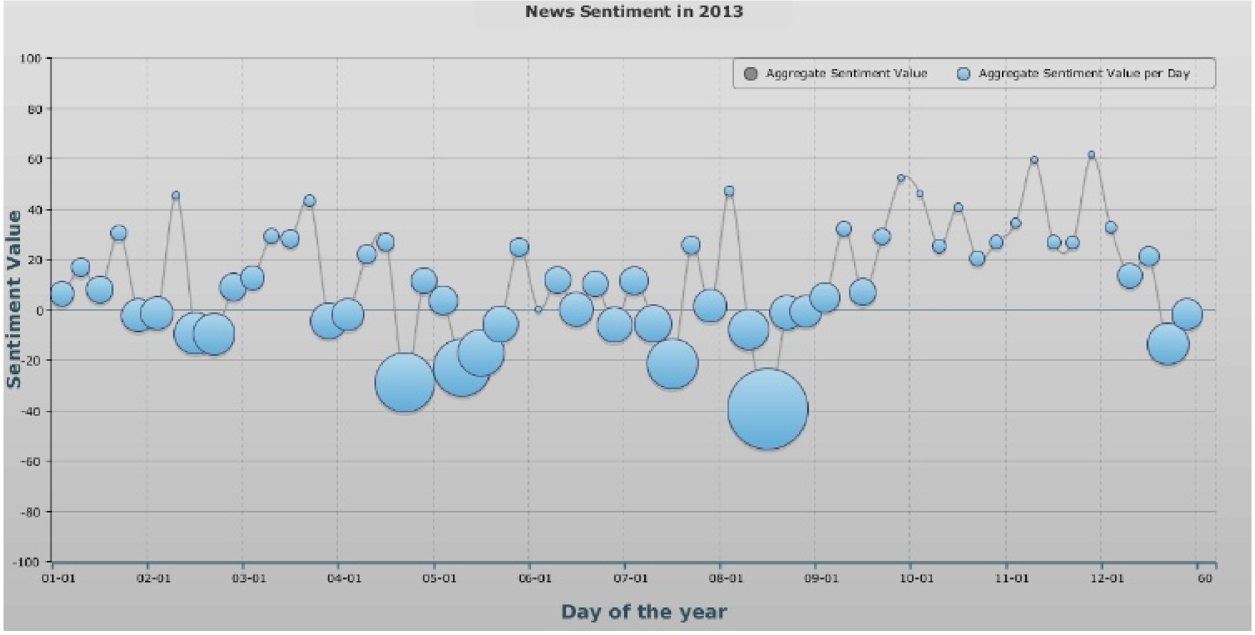


Image 12: Sentiment timeline by week, size of each bubble represents ratio of positive to negative words

Underlying Assumptions

There were a few underlying assumptions in the tools that we used to develop this visualization whose accuracy we were unable to verify. This includes the algorithm used by 10X10 to pick the 100 most important words, as well as the algorithm used by the New York Times to select the ten most relevant articles for a given word. Most importantly, this includes the algorithm behind the sentiment analysis tool that we used (Free Natural Language Processing Service) which is not shared publicly.

Future Directions

The current version of ‘Around the World’ is only the beginning of what can be done with the large amount of rich news data and text analytics technology available today. Given the short timeframe of the class, we believe that we have barely scratched the surface of the insights that can be produced from such analysis and visualization.

As an immediate next step, we would add the headline and a brief excerpt of the news article driving the sentiment in the most influential word of the day. This will provide the user a window into the context that is driving that day’s sentiment.

We think there is room to incorporate more global news data in this project, outside of the NYTimes API that our team used for this class. Ideally, one would use a global news aggregator API such as Google News or NewsCred to generate news articles from all over the world for this project. There is also a need to use a better sentiment analysis tool in the future, preferable one that is trained on news data and takes into account the context of the sentence, in addition to just the individual words.

An important improvement would be to conduct this analysis and visualization in real time on an hourly basis i.e. pull data from several news sources every hour and update the visualization to display updated global news sentiments and their drivers, every hour. However, this would require extremely high computational power, and was beyond the scope of what we could achieve in this class.

We believe there is room to enhance the final visualization and make it more interactive for the user. Given more time and resources, we would show the most positive and negative words on the timeline every hour, and provide the user with links to the news articles that are driving the hourly sentiment. This would allow the user to understand why the sentiment on a given day is what it is.

Finally, as a next step, we would like to share this visualization with our primary audience: media houses and journalists, and ask them if it provides them with any useful insights about their content and it's effect on society. We would like to better understand the questions and concerns on their minds, and tailor the visualization to their needs.

Works Cited:

10X10 (http://tenbyten.org)
MIT Open Gender Tracker: http://opengendertracking.org
Thomson Reuter Financial Products:
http://thomsonreuters.com/products/financial-risk/01_255/News_Analytics_-_Product_Brochure-_Oct_2010_1_.pdf