CART 451: Project Proposal

October 5th, 2022

Amanda Clément

ID 40136259

NAME

A Visual Representation of Collective Sentiments Related to COVID-19

THEME/TOPIC/GOAL/ISSUE TO BE TACKLED AND ITS IMPORTANCE

My project will be an interactive data visualization that illustrates emotional trends related to the COVID-19 pandemic. The challenges that accompanied this crisis such as social distancing and economic shutdown, as well as the controversies such as the effectiveness of vaccines and government public health mandates, make the public's response to the pandemic an interesting topic of study.

The data surrounding the COVID-19 pandemic is predominantly numerical, such as reported cases and death tolls. This quantitative data, although crucial to our understanding of the virus' spread, can be difficult to engage with and to relate to, and gives no information about the public's emotional response to this global crisis. As such, my goal is to present these numerical statistics alongside visual patterns created using the associated qualitative information about the public's collective sentiment to make the data more approachable and digestible.

By displaying these data sets side-by-side, we can analyze, compare and contrast the results that emerge. Some of the questions I wish to gain insight on, and to encourage users to think about, are: What is our collective emotional response to the COVID-19 pandemic? How have these emotions evolved? Do sentiments differ noticeably depending on the geolocations of the individuals the data was collected from?

FORM & AUDIENCE

My project will be an interactive visualization that uses the data collected from publicly available online resources regarding sentiments towards the COVID-19 pandemic. By categorizing the information by time and location, the form in which the findings are visually organized will loosely

resemble a map. By crowdsourcing the data from social media, my project is a sort of social experiment that explores trends amongst social media users. Ultimately, the form my project will take should invite users to try to extract information from the data involved, expand their research, and potentially challenge their existing beliefs.

Although this project could be of interest to a broad range of individuals due to the global impact of the COVID-19 crisis, my target audience is people who are looking to gain insight on the social effects of the pandemic. By condensing a large quantity of data into simple patterns, I hope to curate the gathered information into a form that helps to identify and highlight trends in a way that is easier to understand and digest. My project can also be of particular interest to those who do not use social media frequently or at all, as it gives them insight on the behavioral patterns that exist on these platforms.

HOW THE READINGS HAVE INSPIRED AND MOTIVATED MY CHOICES

The issue of figuring out how to extract information from the sea of data we have access to is one I wish to tackle through my project. My goal is not to gather data, as there is already so much of this available to us, but rather to figure out how to filter it to find patterns within this data.

Hito Steyerl's take on *dirty data* in their article "A Sea of Data: Apophenia and Pattern (Mis-)Recognition" inspired my desire to work with data collected from social media platforms. Considering the common understanding of dirty data as a scrambled, inconsistent and worthless set of information, social media posts can certainly be categorized as such. Riddled with false news, misinformation, bias, hate, spelling errors and bots, social media posts can easily fit into this definition of dirty data. However, as Hito Steyerl explains, dirty data is not necessarily worthless, but can instead be seen as real data that documents the struggles of real people. Hence, my challenge is to sift through the noise to find patterns within the myriad of social media posts, and to organize and present this data in a way that encourages the user to see it as more than just noise.

Moreover, in the article "The Point of Collection", Mimi OOnuoha highlights the significance of the means of collection of datasets. Although I do not plan on gathering my own data, I will be careful in deciding where to source my dataset from. I will research the intentions of the people or parties

responsible for gathering the data as well as the methods they use, and will include these details in my project to be as transparent as possible with my audience. I will also acknowledge how my own bias influences the datasets I chose to use, my analysis of the data and my design choices. Although I cannot avoid or eliminate these biases, being mindful of their existence should help me use the data more responsibly.

MEDIUM

As I am aiming to make information more accessible and easy to understand, the chosen medium should help achieve these goals. As such, my data visualization will be presented on a website as the majority of my target audience is presumably comfortable with accessing and using websites. Also, since I am planning on crowdsourcing data from online platforms, a website is an appropriate medium on which to present the information so that those who contributed to the project can access it.

In terms of flexibility, this medium enables fast and efficient collection and manipulation of large datasets, which is a crucial component to my project. It is also an appropriate choice for working with live datasets should I choose to take that approach. There is also an abundance of excellent resources available, such as libraries and APIs, that can help me realize my project.

Furthermore, I plan on making a website as it aligns best with my knowledge and skill set. Due to my comfort working with this medium, I can focus more on manipulating the data in interesting and informative ways rather than concentrating on learning a new technology.

DATA

For the qualitative data (i.e. emotional responses to the pandemic), I plan on using databases available on websites like Kaggle, openICPSR, and IEEEDataPort that collect information from social media platforms and articles published online. Many of these datasets crowdsource content from Twitter users since Tweets are generally text-based, time-stamped, have a character limit, categorized by hashtags, and are often linked to a geolocation. More specifically, I plan on using a dataset that features a sentiment rating associated with each entry since that information will be used to study and organize the collected data in my visualization.

For the quantitative data (i.e. death tolls related to COVID-19), I will search on these same databases but will also expand my search to information released by governments, medical experts and news articles. For this data, I plan on cross-checking my findings across multiple reputable sources to ensure the information is as reliable as possible.

I am choosing to work with these datasets as they generally include a large number of entries sourced from people all over the globe. Many of them are also managed by multiple people, frequently updated, and corrected based on new findings.

ALGORITHMS

I plan on using Node JS to build the website, complemented by the Express JS framework for ease of use. Since I will be using a local environment for the development of this project, my first algorithm should implement a local http web server which enables clients within the same network to make requests and receive responses from that server.

Next, I will implement a Node module to allow for serving static pages. These pages will include my HTML page, CSS style sheet, and JavaScript files.

Since I plan on working with a dataset of social media posts amongst others, it may be necessary to make modifications to the sample. For instance, Twitter does not permit full datasets of Tweets to be distributed, although they do allow them to be shared. As such, Twitter datasets often only include the Tweet IDS and not the actual post content. Hence if I wish to use this content, I will have to hydrate these Tweets using the Twitter API and a third-party tool like Hydrator which will allow me to convert these Tweet IDs into a JSON or CSV file.

Once the file is ready, I plan on using MongoDB to map the chosen datasets into objects that can be worked on using code, alongside Mongoose for accessing the MongoDB data using queries. These objects will then be styled, sorted, and manipulated in particular ways using JavaScript and CSS to create the visualization.

INSPIRING PROJECTS

Yesterday, Today, Tomorrow by The National Film Board of Canada

https://yesterday.nfb.ca/

Yesterday, Today, Tomorrow is an interactive data visualization by The National Film Board of Canada. It illustrates the collective emotional waves of the COVID-19 pandemic by categorizing hundreds of thousands of Tweets as one of four emotions: joy, confidence, sadness or fear. The visualization is made up of a collection of dots, each dot representing a Tweet and assigned a color representing the associated emotion. The user explores the data in the form of a timeline of months, each having their own collection of dots, starting from January 2020 to the current date. As such, they gain a better understanding of the public's evolving feelings in relation to the pandemic.

Yesterday, Today, Tomorrow is inspiring to me as it is not only conceptually similar to the project I wish to create, but it also does an excellent job at visually depicting qualitative data (human emotions) in a quantified manner. The form in which the data is presented in *Yesterday, Today, Tomorrow* gives shape to the invisible virus in the form of a story told through a timeline of beautiful visual patterns. The visualization also includes certain controls (i.e. filtering by emotion) to help the user navigate the content, making the data more approachable and digestible.

Medusae by Cristina Tarquini

https://artsexperiments.withgoogle.com/medusae/

Medusae is an interactive data visualization by Cristina Tarquini. Told in the form of a story, the project highlights connections between the increase of jellyfish blooms and climate change. The artist's goal, as explained on the website, is to use data to generate discussion, awareness and interest in climate change.

The experience is divided into chapters, each introducing using a statistic (e.g. *WATERS 30% MORE ACIDIC*) followed by patterns of dots that move around the screen. When clicked, these patterns provide more insight on the topic explored in the current chapter. This method allows the user to gradually reveal information at their own discretion, without being overwhelmed by a plethora of statistics all at once.

Aside from it being a striking visual experience, *Medusae* is inspiring for how immersive it feels to interact with the website. The data structures, through their colors and movements, resemble waves and jellyfish. The background music uses field recordings captured in the Mediterranean sea. These elements make the experience feel absorbing and intriguing, encouraging the user to keep exploring the data that is being presented. Furthermore, Tarquini's voice shines through in a way that isn't obtrusive or critical, but rather calm and instructive to effectively generate awareness and interest without overwhelming or discouraging the audience.

***The Divide* by Women Will**

https://dataexplorer.womenwill.com/intl/en/

The Divide is a visual experience illustrating the worsening gender gap in Asia Pacific based on a 2017 Google study. The content is presented in the form of a story, with each section presenting a statistic accompanied by a visual representation to help the user analyze the data. Although these visuals seemingly do not match the numerical data accurately, they help the user navigate the study in a more comprehensible and interesting way.

For instance, when shown the statistic "1 in 5 men don't believe in equality at work", the visual is composed of 5 floating dots, one of which is larger to represent the *1 in 5*. Then to demonstrate the gravity of this statistic, the number of dots increases to show that this *1 in 5* actually represents *over 215,000,000 people*. These dots are then organized into the geographical shape of Japan to show that this number also represents *almost double the population of Japan*. With the visual morphing according to new information being revealed, it helps the user put the numerical data into perspective while enticing them to dive deeper into the experience.

Unlike the other inspiring projects mentioned above, *The Divide* aims to prove a point to its users rather than help them draw their own conclusions. The creators' bias is prevalent throughout the experience and their intentions are stated before the user starts interacting with the visualization. This feels like a more honest approach to take when trying to inform, persuade, or convince the user of something as they are made aware that the data is being framed in particular ways to advocate for certain interpretations.

BIBLIOGRAPHY

Onuoha, Mimi. "The Point of Collection." *Medium*, 2016. https://points.datasociety.net/the-point-of-collection-8ee44ad7c2fa#.y0xtfxi2p. Accessed 1 October 2022.

Steyerl, Hito. "A Sea of Data: Apophenia and Pattern (Mis-)Recognition." *e-flux Journal*, 2016. https://www.e-flux.com/journal/72/60480/a-sea-of-data-apophenia-and-pattern-mis-recognition/. Accessed 1 October 2022.

Tarquini, Cristina. "Medusae." *Google Arts & Culture*, 2022. https://artsexperiments.withgoogle.com/medusae/. Accessed October 1 2022.

Women Will. "The Divide." *Google*, 2020. https://dataexplorer.womenwill.com/intl/en/. Accessed October 2 2022.

"Yesterday, Today and Tomorrow." *National Film Board of Canada*, 2020. https://yesterday.nfb.ca/. Accessed 30 September 2022.

SENTIMENT ANALYSIS OF COVID-19

USING DATA COLLECTED FROM TWITTER

ABOUT THE PROJECT

SADNESS

ANGER

FEAR

NEUTRAL

DATE: 2020 ↑

HOMEPAGE: What the user sees when they open the website. For this sketch, suppose I'm using a dataset of Twitter posts that are categorized based on the emotion they best represent. Each dot represents a Tweet from the dataset, colorized based on its assigned emotion and sized according to the popularity of the Tweet (e.g. number of retweets). A legend is displayed in the bottom right corner, and a button to learn more about the project appears in the top right corner.

SENTIMENT ANALYSIS OF COVID-19

USING DATA COLLECTED FROM TWITTER

ABOUT THE PROJECT

INDIA

CHINA

USA

CANADA

PHILIPPINES

SOUTH AFRICA

NIGERIA

NETHERLANDS

UK

BRAZIL

NORWAY

AUSTRALIA

24% SADNESS

27% ANGER

12% FEAR

37% NEUTRAL

KEYWORD SEARCH: _____

KENYA

MALAYSIA

DATE: 2020 ↑

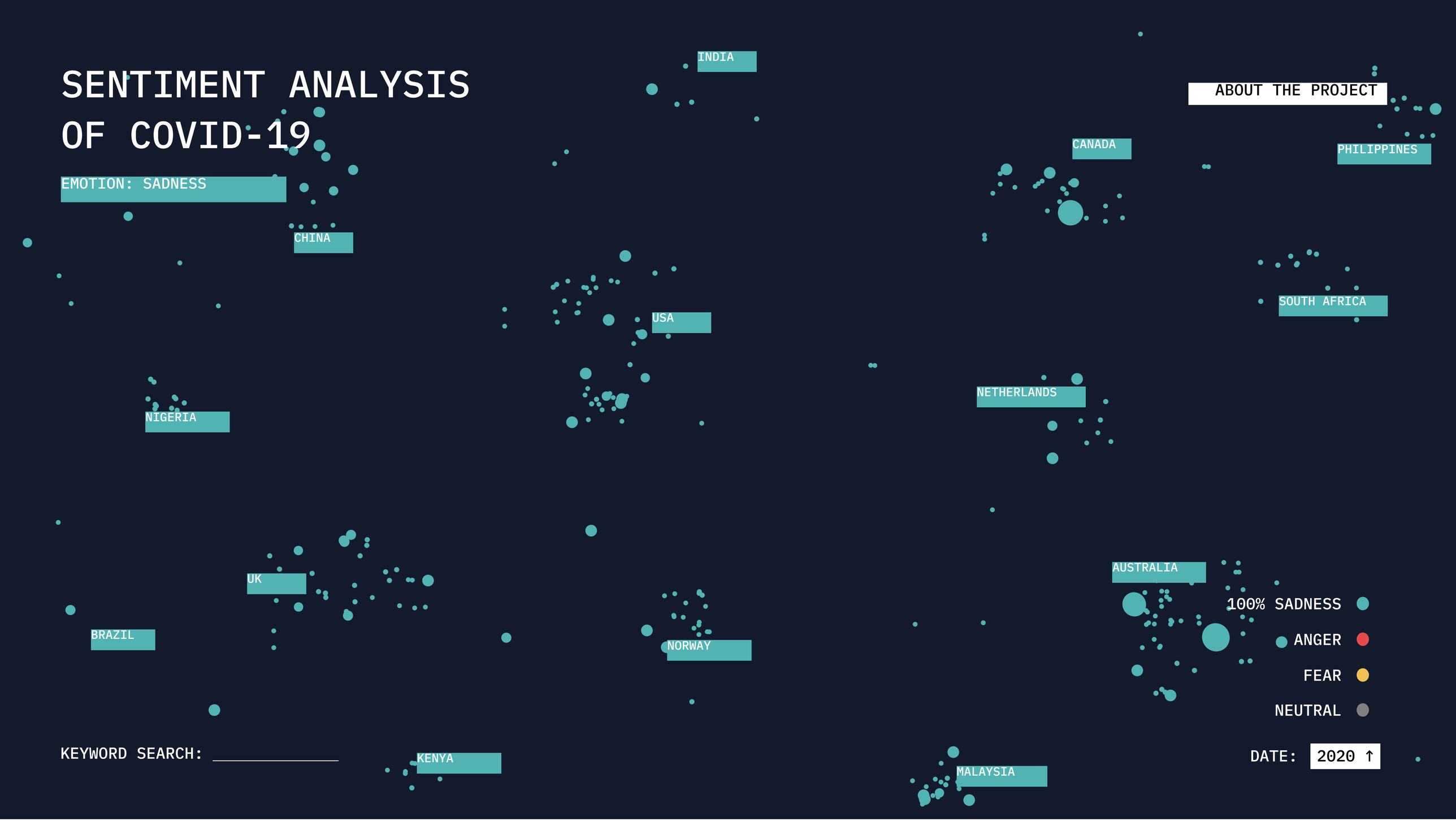
After a few seconds, the collection of dots separates into smaller clusters. These are grouped based on the Tweets' geolocations. The data entries that don't include a geolocation simply float all around the screen. The legend (bottom right corner) is populated with percentages, and the user may click on one of these emotions to remove all of the dots from the visualization that are not categorized by that emotion. The user is also given a drop-down menu (bottom right corner) to change the date to a specific month, which alters the visualization accordingly. They may also enter a keyword (bottom left corner) to search within the Tweets, which only displays the dots of data entries that include that keyword. The user may also click on a country name to zoom in on that cluster.

SENTIMENT ANALYSIS OF COVID-19

USING DATA COLLECTED FROM TWITTER

ABOUT THE PROJECT

A demonstration of what happens when the user clicks on a dot. The Tweet content (if available) is displayed.



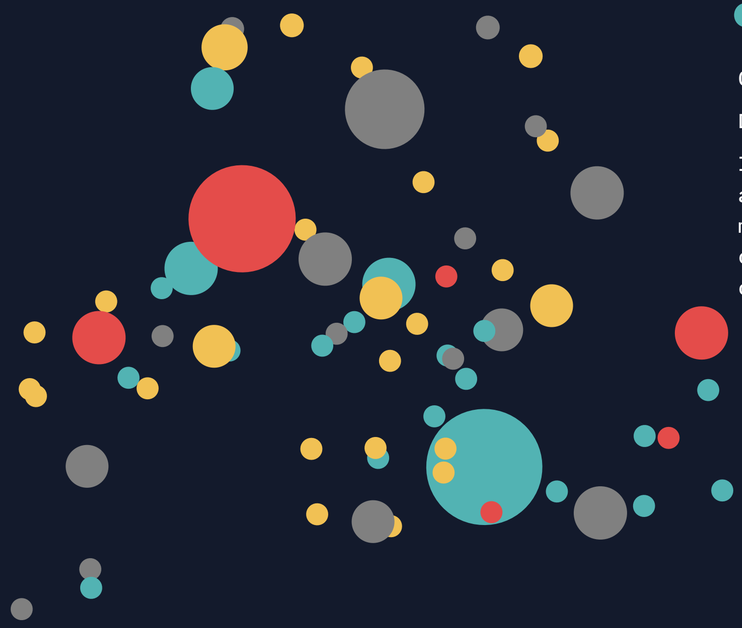
A demonstration of what happens when the user chooses to filter the content to only show data related to *sadness*.

A demonstration of what happens when the user choose to filter the content to only show data related to the keyword *PFIZER VACCINE*. The legend's percentages are updated accordingly. The background color of the subheading (*KEYWORD: PFIZER VACCINE* in this case) is set to the color of the predominant emotion.

SENTIMENT ANALYSIS OF COVID-19

COUNTRY: CANADA

ABOUT THE PROJECT



CANADA

DATE: JANUARY 2020 ↑

36% 18% 12% 34%

Cases: XXXXXXXX

Deaths: XXXXXXXX

Important News: Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut mauris diam, congue vel malesuada sed, cursus non ex. Integer euismod tempor congue. Donec aliquet ut arcu sit.

36% SADNESS

18% ANGER

12% FEAR

34% NEUTRAL

DATE: JAN 2020 ↑

A demonstration of what happens when the user clicks on a country button. It zooms in on the cluster and a popup appears with more details including the emotion percentages, number of COVID-19 cases, number of COVID-19 deaths, and important news from the selected timeframe (in this example, we show data from January 2020). Additional data may be added (e.g. number of recoveries, percentage the cases have increased by since the previous month, etc) based on the information that's available. The background color of the subheading (*COUNTRY: CANADA* in this case) is set to the color of the predominant emotion.

RETURN TO THE VISUALIZATION

ABOUT THE PROJECT

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut at dignissim massa. Aenean vitae gravida eros. Donec venenatis porttitor odio a cursus. Phasellus sit amet metus vel ante interdum pharetra. Phasellus vitae tempus metus, eu porta eros. Integer mollis lobortis metus non accumsan.

Ut dictum sem ac risus viverra interdum. Vestibulum sit amet lectus rhoncus, eleifend mi non, imperdiet mauris. Sed a felis varius, efficitur eros sed, convallis elit. Curabitur facilisis, justo eu lacinia vestibulum, magna leo vestibulum metus, in ultrices orci nisi eget dui. Proin blandit commodo ligula, a dictum neque luctus sit amet. Cras mi urna, ultrices quis dapibus quis, dignissim non nibh. Donec vel mauris a nulla suscipit commodo.

In euismod interdum tempor. Aliquam erat volutpat. Integer facilisis arcu nec ante condimentum tincidunt. In feugiat elit nec volutpat ullamcorper.

ABOUT THE PROJECT PAGE: Includes a description of my goals/intentions, design choices, where the data was collected from, why the data was collected, credits, and any other information that is pertinent to the project.