

Electronic Theses and Dissertations, 2004-2019

2016

The Weight of Words: Collecting and Visualizing Data from Twitter

Daniel McSwain
University of Central Florida

 Part of the [Film and Media Studies Commons](#)
Find similar works at: <https://stars.library.ucf.edu/etd>
University of Central Florida Libraries <http://library.ucf.edu>

This Masters Thesis (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Electronic Theses and Dissertations, 2004-2019 by an authorized administrator of STARS. For more information, please contact STARS@ucf.edu.

STARS Citation

McSwain, Daniel, "The Weight of Words: Collecting and Visualizing Data from Twitter" (2016). *Electronic Theses and Dissertations, 2004-2019*. 5253.
<https://stars.library.ucf.edu/etd/5253>

THE WEIGHT OF WORDS: COLLECTING AND VISUALIZING DATA FROM TWITTER

by

DANIEL RICHARD MCSWAIN

B.F.A., The University of West Florida, 2012

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Fine Arts
in the School of Visual Arts and Design
in the College of Arts and Humanities
at the University of Central Florida
Orlando, Florida

Fall Term
2016

© 2016 Daniel R. McSwain

ABSTRACT

The Weight of Words is a web-based artwork designed to capture snapshots of Twitter discussions concerning the most popular topics of the day. The growth of social media in recent years has led to a sharp increase in thought and opinion sharing among the vocal population on the Internet. Twitter's use of trending topics allows users to be aware, and be a part of fun or silly stories as well as important news headlines and social movements. *The Weight of Words* is an exploration into using Twitter's always changing landscape of conversation to generate graphic visualizations based on the most frequently used words at the time. This thesis includes a discussion regarding design considerations, application architecture, and data mining, as well as an examination of data visualization, social media, and human behavior. Through the construction of these visualizations I aim to provide a unique opportunity to discover patterns and trends from the popular topics of that current day. By providing viewers of this work with a unique perspective, I hope to encourage reflection and discussion of the current state of our culture's behavior and values.

To my parents,
Richard and Wanda McSwain,
and all of my family,
without whom none of my success would be possible.

ACKNOWLEDGMENTS

I owe my deepest gratitude to my committee chair, JoAnne Adams, whose outstanding guidance, encouragement and support from the beginning to the end of this thesis was invaluable. I would also like to sincerely thank the other member of my committee, Keith Kovach, Dr. Anastasia Salter, and Dr. Peter Smith for their advice, criticism, and understanding.

I am indebted to many others at UCF who have been influential in the development of my work. In no particular order, thank you to Jason Burrell, Natalie Underberg-Goode, Phil Peters, Stella Sung, Mark Price, Dan Novatnak, Mark Gerstein, Joseph Fanfarelli, and Richard Grula.

I must also thank my cohort in the Digital Media and Studio Art programs at UCF. To Reina Castellanos, Taylor Battle, and Alesha Hassard, I truly appreciate your friendship, humor, and encouragement during our time together.

To my closest friends Rashaad Rosalle and Nathaniel Johnston, thank you so much for being there for me over the years. Your honesty, dependability, and thoughtfulness made me want to be a better person and friend.

Finally, to my family, you are all the best people I know. Thank you to my parents, Richard and Wanda McSwain, for your unending generosity and selflessness. Thank you to Rachel and John Tanner, for your advice, insights, and confidence. Thank you to my brother, John McSwain, for your mentorship and inspiration. Your love means more to me than anything in this world.

TABLE OF CONTENTS

LIST OF FIGURES.....	vii
LIST OF ACRONYMS/TERMINOLOGY.....	viii
INTRODUCTION.....	1
HISTORY OF INTEREST.....	3
SOCIAL MEDIA AND DATA.....	6
DATA VISUALIZATION AND ART.....	12
RELATED WORKS.....	14
We Feel Fine	14
Listening Post	16
MIMMI	18
APPLICATION ARCHITECTURE.....	20
DESIGN CONSIDERATIONS.....	24
DISCUSSION.....	34
CONCLUSION.....	36
APPENDIX: COPYRIGHT PERMISSION LETTERS.....	37
REFERENCE(S).....	40

LIST OF FIGURES

Figure 1: 016, Image by Author.....	4
Figure 2: We Feel Fine, by Jonathan Harris and Sep Kamvar. (Source: http://number27.org/wefeelfine).....	15
Figure 3: The Listening Post, by Mark Hansen and Ben Rubin. (Source: https://www.flickr.com/photos/arselectronica/5453987194). CC BY-NC-ND 2.0.....	17
Figure 4: MIMMI, by INVIVIA and Urbain DRC. (Source: http://www.gsd.harvard.edu/project/mimmi).....	19
Figure 5: Word cloud comparison of two State of the Union speeches by two U.S. presidents, by Pyrsmis. (https://en.wikipedia.org/wiki/Tag_cloud#/media/File:State_of_the_union_word_clouds). CC BY-SA 3.0	25
Figure 6: 2015 State of the Union Address tag cloud, by Jason Griffey. (http://jasongriffey.net/wp/2015/01/20/state-of-the-union-2015-tag-cloud). CC BY-NC 4.0	25
Figure 7: D3 Radial Graphic Prototype, Image by Author.	27
Figure 8: #Debate from October 9 th , 2016 from 7:55 to 8:15pm. Image by author.	29
Figure 9: Display of #Debate and #DebateNight series. Image by Author.....	31
Figure 10: #DebateNight from October 19th, 2016 from 9:23 to 9:43pm. Image by author.	32
Figure 11: #DebateNight from October 19th, 2016 from 11:08 to 11:28pm. Image by author.	33

LIST OF ACRONYMS/TERMINOLOGY

API – Application Programming Interface

ASCII – American Standard Code for Information Interchange

CSS – Cascading Style Sheets

D3 – Data Driven Documents

DIY – Do It Yourself

HTML – Hypertext Markup Language

JSON – JavaScript Object Notation

LED – Light Emitting Diode

MEAN – MongoDB, ExpressJS, AngularJS, Node.js

NPM – Node Package Manager

SVG – Scalable Vector Graphic

REST – Representational State Transfer

INTRODUCTION

My practice as a maker is driven by my desire to engage viewers and users with my work, both physically and mentally, by presenting opportunities to participate and interact with the artwork. Past and current projects have been influenced by modes of historical and experimental do-it-yourself (DIY) methodologies used in photography and videography, as well as the manipulation of analog and digital objects. With my work, I aim to educate myself, and others, about the processes and capabilities of Digital Media.

During my candidacy for the M.F.A in Emerging Media at the University of Central Florida, I have explored and experimented with various methods of image making, both still and moving. My work prior to my enrollment at UCF involved the digital deconstruction and reuse of visual information through data manipulation, for the purpose of creating colorful images and understanding how visual information conveys meaning. While studying how these still and moving images can be dissected into their basic components of code or data, my curiosity began to shift. Rather than wanting to understand how digital content breaks down, I became interested in understanding how data can be used to construct new information.

My research expanded into observing how we, as consumers of information, use visuals to informally learn about a wide variety of content and its related context, especially in social and educational settings. As I observed my own online social networks, I became more conscious of the power of social learning and the ability for people to share visual information within their own networks. While studying various gateways of social information, such as Twitter, I began to focus specifically on popular social conversations and trending topics which usually involve a massive volume of discussion created by a considerably large number of participants.

Using a variety of web tools and libraries, I constructed a digital application that allows me to collect data directly from Twitter and repurpose it in a creative and educational way. This data is searched for and collected using Twitter's Streaming API, which provides developers with near real-time access to Twitter's global streaming data. Other useful tools implemented in this project include the D3.js visualization library, AngularJS html extensions, Node.js networking tools, and other basic web development standards (HTML, CSS, and JavaScript). These development tools used in conjunction gave me the ability to collect dynamic content from the web, use the new content to create informative graphics, and redistribute the visualizations back to the web where they live.

The purpose of this thesis is to capture content within the discussions taking place on the Internet, in order to better understand the nature of the conversation. Furthermore, through extended observation of social media sites like Twitter, I have discovered that many trending topics rise in popularity very quickly to only fade away when another topic becomes more popular. The information and discussions relating to a buried topic likely still exists on the Internet, however as time moves on the connections and specifics are tangled with other closely related topics. These ideas and concerns all helped shape the central question to my exploration:

Using visualized data, how can I provide my audience with an opportunity to explore what the public discourse is focusing upon in order to understand what our culture values?

The challenges I encountered throughout this project involved the learning of complex methods of application development, as well as how to be a fair and thorough curator of the information presented through digital imagery. Much of this thesis focuses on this process of development and the decisions that were made to reach the current version of the application.

HISTORY OF INTEREST

Prior to my enrollment at the University of Central Florida, I worked in a variety of mediums including photography, video, sculpture, sound and circuitry. The projects and artworks I created usually involved methods of deconstruction and reconstruction of materials in order to create something that could only exist if something else was broken. These projects led me to several variations of object and data manipulation, including the “bending” of physical and digital components, and ultimately a fascination for creating “glitch art.” The word “glitch” is a term frequently used in computing, usually when referring to an error that changes the normal operating behavior of a component or computer function.

My interest in glitch art evolved from early experiences in my life when I was taught to tinker and take objects apart to understand how they work. Raised in a family of engineers and artists, I was constantly presented with ideas about how and why things work and break, as well as why it was important to investigate those ideas. These lessons remained with me as I grew up and began exploring more advanced methods of object and image manipulation. One particularly important skill I developed prior to college was the manual creation of circuits and circuit-based artworks. I started with simple kits that taught me how to solder and build small devices, that in turn provided me with the confidence to work with more complex circuits. My interest in creating small working objects began to shift to wanting to take existing circuits and change them to work in a completely different way. I began to research and experiment with different methods of short-circuiting and breaking various components from objects like electronic toys, video mixing boards, and even cell phones. This method of circuit manipulation was given the term “circuit bending” in the 1960s by Reed Ghazala, when he accidentally short circuited a toy against his metal desk and it made strange

sounds. (Ghazala, 2000) The concept of circuit bending was what initially led me to begin working on the process of data bending. Both of these ideas involve taking something that exists and changing it to become a totally different representation of itself. That conceptual process is a primary driving force behind much of my work.

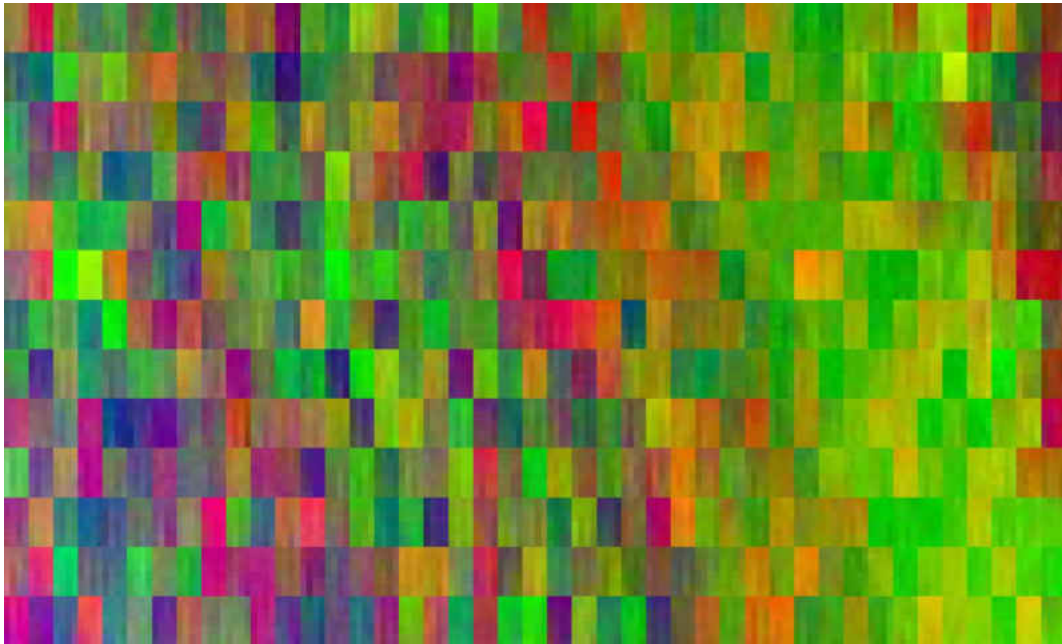


Figure 1: 016, Image by Author

The transition from analog to digital manipulation overlapped with a few projects that involved the analog manipulation of circuit boards in a video mixer. The images produced from these experiments contained drastic pixel and color distortions depending on the functions activated on the board and the locations of the wires I manually attached to the circuit. After experiencing the effects of image manipulation, I produced a series of images using data bending methods on many of my personal photographs. An example of this process can be seen in Fig. 1. The mechanism of manipulation involved the careful editing of image file code to be read and changed by incompatible

software like a word processor. My resource for this technique came from a blog post by a glitch artist called Stallio!, who presented the method as “The WordPad effect.” (Stallio!, 2005) This process is what led me to research more advanced methods of image code editing, such as pixel sorting, which requires programming knowledge to accomplish. This requirement meant that I needed to begin learning some code languages, in order to push my image editing abilities forward. I decided to start by learning some of the Java and JavaScript languages, introducing me to software like Processing and other development environments.

Throughout my second semester at UCF, I was involved in exploring techniques used for image making by coding small interactive projects in Java. In those few months, I created a few drawing applications that relied heavily on user participation. As I worked, I began to question what other methods I could use to create images with code, instead of relying on a user to physically draw on the screen. This was the beginning of the process of building larger applications to use raw data for dynamic image making. These explorations led to the start of my interest in using social media as a primary resource for fresh data. By studying my relationship to the Internet and how social discourse has shifted over time, I began to form the basis of this thesis project.

SOCIAL MEDIA AND DATA

Since the early 2000s, I have used various forms of social media including LiveJournal, Myspace, Tumblr, Facebook and Twitter. Each of these websites has provided users with the ability to share their thoughts and connect with like-minded people across the world. Some social media sites cater to more specific users or content, such as musicians or blog post writers, while other sites provide users with direct access to communicate with brands, celebrities or even politicians. For this project, I decided to focus on the use of Twitter and the massive amount of content produced by millions of users each day through tweets.

Twitter was launched 10 years ago, in 2006, and has since become one of the largest online social networks in existence. As of June 2016, Twitter has an average of 313 million monthly active users creating content to share with other people (Twitter, 2016). This content is usually text based posts limited to 140 characters, shared as original content or retweeted from a user's follower. A retweet is the act of sharing someone else's content with your own group of followers and expanding its reach to more networks. Many people retweet content as an act of agreement or affirmation, however, many other people may retweet and append their own opinion or disagreement to it. This addition of new content may start a conversation between users or groups of users, as they tweet and retweet back and forth. If a topic gains a high volume of discussion, then it may be eligible to be listed on Twitter as a trending topic. The topics that start trending on Twitter, in many cases, are grouped by using a hashtag word or a word prefixed with a “#” symbol. By using this hashtag, users participating in the conversation can make sure their content or opinions are included in the aggregate conversation. When querying Twitter for the current trending topics, the content that is returned is a varying list of hashtags, names, events or other popular

terms, as well each topic's "tweet volume." (Twitter Developers, 2016) The tweet volume of a topic is a measurement of how many posts contain the hashtag or trending word, with the measurement being reset every 24 hours. Throughout each day the list of trending topics evolves and changes as breaking news or popular events shifts the focus of the vocal Twitter population.

The use of data from social media to make art is a concept that has grown significantly in the past decade. A great example of this rise in social data use for creative purposes is the Twitter bot. If you are a Twitter user, you have most likely encountered a Twitter bot at some point during your browsing. However, most bots on Twitter are not there to create art or use data to generate interesting content. The majority of bots are built to promote specific content, add huge numbers of retweets to a post, or even advertise services for businesses. Yet, a unique contrast to these spam bots are a small population of creative bots that collect content, analyze and change it, and then send the new content back to Twitter. A couple notable examples include Adam Parrish's *everyword*, which tweeted every single word in the English language alphabetically. Created in 2007, the bot tweeted a word every 30 minutes for 7 years, completing the language in 2014. Another is Darius Kazemi's *Two Headlines*, which searches for two news headlines from different topics and combines them into a new headline. While most combined headlines don't make any sense, every once in a while a headline is created that is unexpectedly interesting or funny. The ability to create and deploy these small bots on Twitter provides developers with a way to contribute to the unique landscape of social Internet art.

As an observer of social media, I find that the way people speak and act is closely aligned with their true feelings when they are communicating through their computers or mobile devices, rather than when they share their thoughts and opinions in person. In many cases, users on Twitter speak behind an alias, or through some sort of anonymous guise that allows them to feel the

freedom to share their perspective without repercussion. While the use of anonymity isn't universal to all users of social media, it is common enough to be acknowledged as a popular practice. In a book titled *The Presentation of Self in Everyday Life*, Erving Goffman explains that:

...when an individual appears before others his actions will influence the definition of the situation which they come to have. Sometimes the individual will act in a thoroughly calculating manner, expressing himself in a given way solely in order to give the kind of impression to others that is likely to evoke from them a specific response he is concerned to obtain. (Goffman, 1959)

Goffman's ideas concerning an individual's ability to manage impression are still relevant today, even in an online setting. Users on the Internet have the ability to only disclose specific personal details to others in order to craft an idealized version of themselves. With the complexity of today's Internet, there is no singular reason for a person to separate one identity from another. Although, with increasing access to user profiles through the Internet, businesses have the ability to investigate a person's behavior on social media before making any future hiring decisions. This type of insight may cause individuals to use their public presence to present a positive impression to others, while their other identity may demonstrate a completely different behavior. I personally find it interesting that many individuals feel the need continually adjust their self-presentation in order to share their personal truths.

Using an alias on the Internet is not a new idea. When I was growing up, one of my first experiences with the Internet included the creation of my first email address. I remember spending more than a week thinking of a cool name that would make me stand

out from my friends. I settled on a name that related to something I enjoyed at the time which the tinkering of objects to learn how they worked. Thus the email and username “takeitapart89” was created. Since that time I have only created and used two or three other Internet aliases, typically as a response to new technology or a new community. Personally, my use of an Internet alias isn't due to me wanting to share thoughts without being judged or identified. I instead make use of an alias as a way to separate my real life from my virtual life. This is not a strict separation with a need for secrecy, but more of way to organize the people I interact with on a daily basis. For websites that I use to connect with people I see in real life, I always use my real name. There are also other sites where I only use my alias and many people only know me by that name. However, a few websites exist, such as Twitter, which list both my real name and my alias together. The concept of hiding behind a fake account is still a widely used tactic, so much so that the term “throwaway account” is now common. On websites like Facebook, Twitter, and Reddit, users can start a new account using a new email and have the ability to say whatever they want without anyone knowing who they are. After the user is finished saying what they want to say, they can abandon or completely erase the account and email without worrying about their personal identity. There is no way of knowing if what the people who make these accounts say is fact, or if it is even their true opinion, but I find that when people think they are unknown they tend to speak more openly. This is also a commonly used tactic for those who want to “troll” other people. Internet trolling is something that is done to elicit a reaction from other people. There are individuals and groups of people that seek out discussions and debates in order to interject a comment or piece of content that provokes others into responding emotionally. In many cases these statements are obvious, and can be ignored or not taken seriously.

Sometimes however, someone trolling a chatroom, blog, forum or other social media site can spark a reaction that disrupts the entire discussion into arguments and off-topic debate. Many of these inflammatory statements are carefully crafted to cause grief or offense to some people but not others. In a paper about identity and deception in the virtual community, Judith Donath equates trolling to a type of game:

Trolling is a game about identity deception, albeit one that is played without the consent of most of the players. The troll attempts to pass as a legitimate participant, sharing the group's common interests and concerns; the newsgroup members, if they are cognizant of trolls and other identity deceptions, attempt to both distinguish real from trolling postings and, upon judging a poster to be a troll, make the offending poster leave the group. Their success at the former depends on how well they – and the troll – understand identity cues; their success at the latter depends on whether the troll's enjoyment is sufficiently diminished or outweighed by the costs imposed by the group. (Donath, 1999)

The concept of anonymity on the Internet is important to this project because some people will only say what they truly feel if they know they can't be identified. Even though what is said behind an anonymous guise isn't always positive, correct, or even related to the topic, it is still what someone feels they need to say. Through the process of retweeting content, sharing similar ideas, and communicating with others, a series of words and ideas can rise to the top of the conversation to show pattern and commonality. Using these words as an echo of the vocal community, I can work to inform others about how our culture is speaking and behaving. Depending on the day and the topic, insightful reflections of public discourse may change how people feel about our culture.

Through visual exploration of the most common words, I can judge a shift in behavior and tone between the positive and the negative.

A practice I have observed in the social media data community is the analysis of user sentiment. The concept of “what other people think” can be an important part of how many people make decisions. As humans we tend to seek suggestion and affirmation for our choices, which means that we listen to the opinions of others and sometimes even share our own opinions. A problem with seeking other’s opinions on the Internet, however, is the fact that the English language is full of slang, double-meaning and sarcasm. A common method of discovering sentiment from data is to only pay attention to words that have been clearly defined and grouped into categories of feelings. By disregarding words that don’t match the list of possibly several thousand defined words, the analysis of sentiment can be more accurate. Without this type of foresight, the analysis of the many random words and phrases used on social media may lead to an incorrect determination of sentiment and feeling. This outcome is common with many sentiment analysis tools on the Internet that essentially tell the viewer whether something that was said is positive or negative. The majority of the time the result is wrong or inconclusive, due to incorrect understanding of a sarcastic phrase or word. By limiting the language that can affect the analysis of sentiment, a more accurate result is achievable. Due to the complexity of this issue, I decided that I did not want to limit the language or convince a viewer that something is positive or negative. Instead I opted to present the data in a way that allows a viewer to come to that decision themselves.

DATA VISUALIZATION AND ART

As a process-oriented artist working to learn how to develop a web application, much of my focus centered on finding a balance between the creation of a pragmatic and artistic visualization. In a paper discussing visualization criticism, Robert Kosara differentiates varying types of data-based visualizations by their ability to be read and recognized. (Kosara, 2007) He classifies these visualization methods using a scale that separates both pragmatic and the artistic approaches to either end of the spectrum. A pragmatic data visualization can be described as more technical, and is directed towards the analysis of the data. Whereas an artistic visualization is focused less on the analysis of data itself and more on the presentation of a concept. Both types of visualizations can deliver valuable content, as long as it is understandable. On Kosara's scale the pragmatic visualizations tend to have more utility due to the presentation of data in a recognizable and readable way. However, artistic visualizations deal more with sublimity, and are not usually recognizable or readable. Kosara writes that "Data collection is often an integral part of a visualization art piece. The fact that the data exists at all can be used to create awareness, and data flowing in in real time can make the piece "live"." In many cases a pragmatic visualization can use varying sets of data, leading to the use of a single visualization technique to compare and contrast different collections of information. On the opposite side of the spectrum, artistic visualizations are often created as unique works that aren't meant for analyzing information, but rather to communicate an idea or concern. Based upon Kosara's classifications, I cannot see this work being defined as solely pragmatic or artistic. The visualizations generated from the application have the capability of presenting different datasets over time, as well as the ability to communicate public ideas and concerns.

During the process of development, I observed my application take the form of a fine art project. As a web-based artwork, what I created essentially behaves like a camera. However, instead of capturing light, the application captures data. Similar a photographic device, I am able to point the application at a subject, set an exposure time, and click a button to record the content. This concept further relates to this project in that what I am doing is capturing snapshots or moments in time separated from everything else. Social media today moves at such a fast pace that the conversations taking place now can easily fade away as new content is created and shared. The idea of presenting a fresh perspective to viewers is vital to the core of this project and its artistic attributes. The Oxford Dictionary defines fine art as “Creative art, especially visual art whose products are to be appreciated primarily or solely for their imaginative, aesthetic, or intellectual content.” (“fine art”, 2016) With this in mind, I sought to develop the ability to create visual images that allow viewers to see things that are usually unseen. This momentary capture of public discourse provides the new perspectives through which my audience can explore our cultural values.

RELATED WORKS

We Feel Fine

We Feel Fine is an artwork designed to collect and visualize emotional content from all over the world. Created by Jonathan Harris and Sepandar Kamvar in 2006, this work is a full web application built as a content collection and data visualization tool for the purpose of sentiment analysis and information sharing. Harris and Kamvar worked together to explore the possibilities of content categorization based upon the characteristics of the content and its creator. The main focus of their work is the heavy statistical analysis of the data, including categorization of each user's age, gender, location, language, etc. This analysis is presented in a series of visualizations showing relationships between emotion, context, and user content. Harris gives a brief explanation of the work's purpose:

At its core, *We Feel Fine* is an artwork authored by everyone. It will grow and change as we grow and change, reflecting what's on our blogs, what's in our hearts, what's in our minds. We hope it makes the world seem a little smaller, and we hope it helps people see beauty in the everyday ups and downs of life. (Harris, 2016)

We Feel Fine is a work that informed me of what is possible when mass amounts of data are collected and analyzed for similarities and patterns. In Fig. 6, the small colorful dots in the images are representations of individual words captured from sentences that include the phrase "I feel", or "I am feeling." The sentences are then searched, looking for words that match a list of 5,000 hand-picked terms describing possible feelings. Each word is separated from one another and presented in a series of interfaces that offer differing analytic perspectives. Users can interact with and sort the information to discover how other people share their emotions with the Internet.

In Harris' description of *We Feel Fine*, he explains that the artwork is authored by everyone. This idea is what first interested me in using data collected from Twitter. Because of this work, I understood that if I could put in place a series of rules and limitations for how data could be collected and displayed, I could capture any topic at any time and possibly discover something new from the results. *We Feel Fine* is an artwork that I still look to as an example of where my work could evolve.

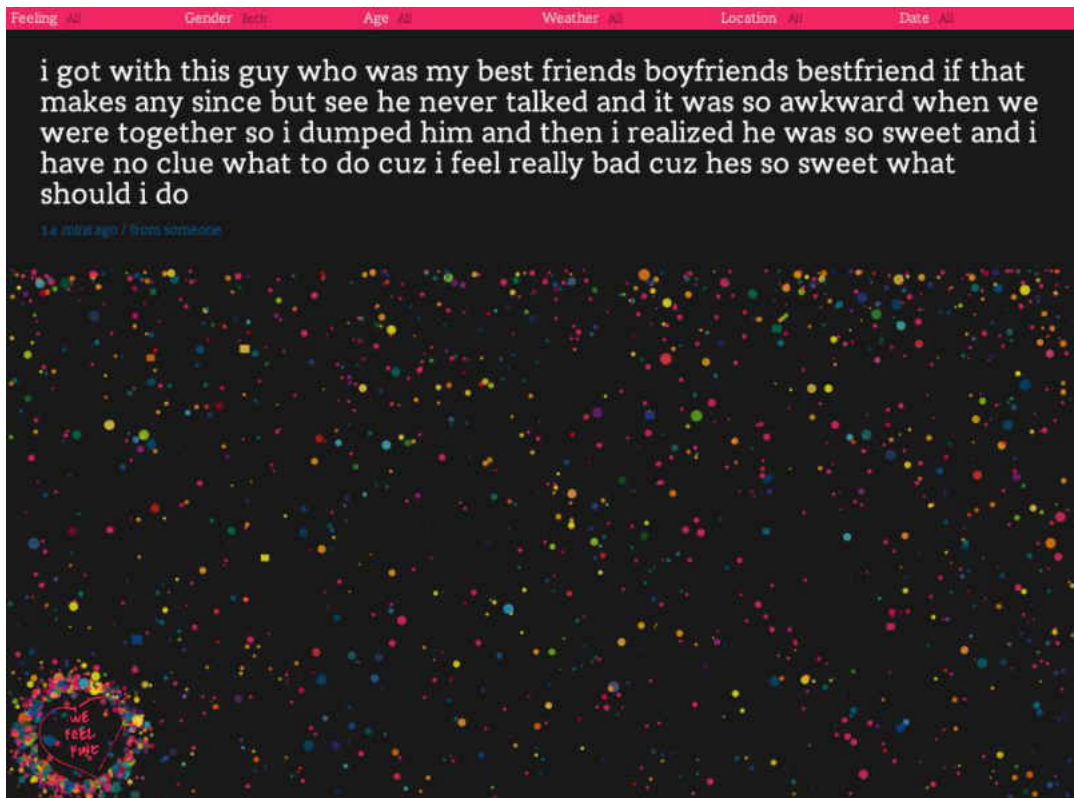


Figure 2: We Feel Fine, by Jonathan Harris and Sep Kamvar. (Source: <http://number27.org/wefeelfine>).

Listening Post

Listening Post is an installation artwork consisting of several hundred small digital screens displaying content from the Internet. The LED screens are suspended in a widely arcing grid, reminiscent of a science fiction computer display (Fig. 3), each displaying a phrase or a fragment of a sentence that begins with descriptive words like “I am” or “I like.” The project was developed in the early 2000s by Mark Hansen and Ben Rubin, and has since been updated with new data as the work is shown in different exhibitions. Alongside the visual content displayed on the screens is an audible landscape of digitized voices and clicking sounds as if the words were being typed and read aloud in the same room. Prior to the exhibition of this work, Hansen submitted a paper to IEEE examining the use and implications of crowdsources data. Hansen writes:

As we become more adept at observing, measuring, and recording both our physical and virtual movements, we can expect to see a proliferation of large, complex data sets. These new digital records of human activity often force us to consider difficult societal questions. From the current debate over privacy on the Internet to ethical concerns over the use of genetic information, we recognize that the now simple act of compiling data has serious political implications. To further complicate matters, the scale and structure of these data make them difficult to comprehend in the large, relegating activities like data analysis to a select few (Hansen & Rubin, 2000).

Hansen and Rubin present the idea of data collection as a catalyst to understanding how we behave as humans in the Internet age. With their work, they raise concerns about how our information and public opinions are used to analyze our intentions and emotions. Through the

presentation of thousands of self-defining statements, *Listening Post* demonstrates how willing people are to share their private thoughts with the world.

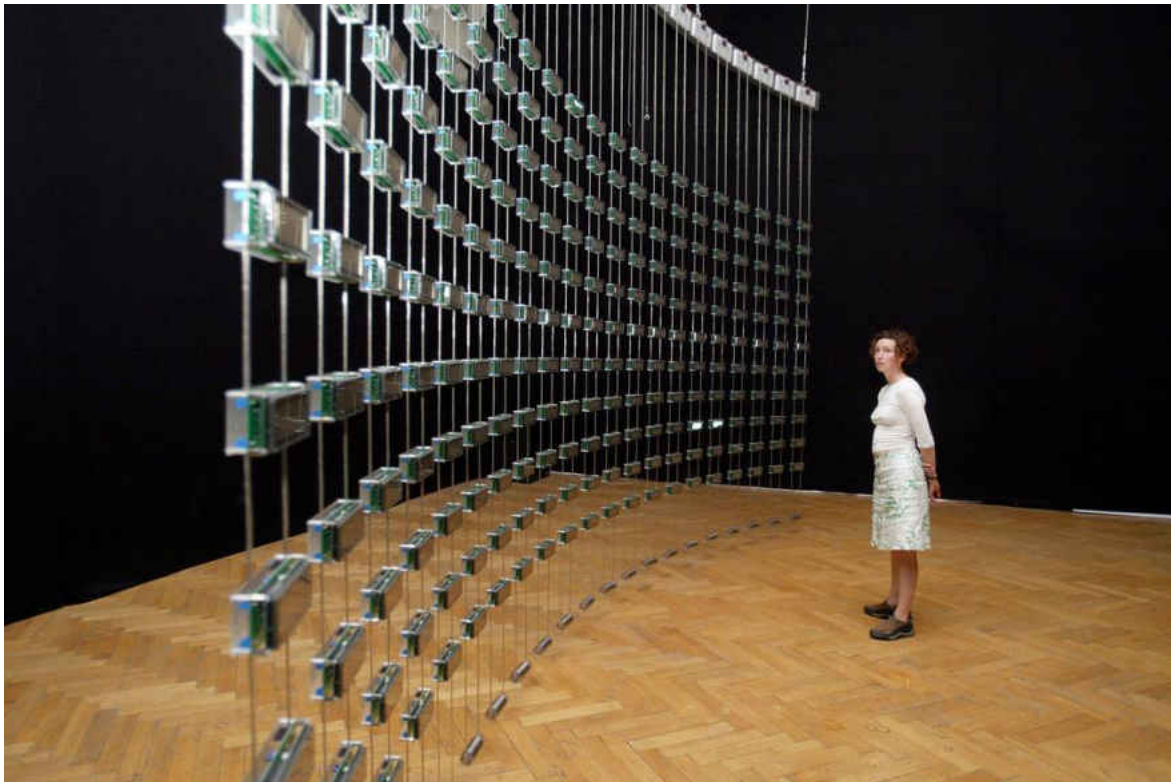


Figure 3: The Listening Post, by Mark Hansen and Ben Rubin. (Source: <https://www.flickr.com/photos/arselectronica/5453987194>). CC BY-NC-ND 2.0

MIMMI

MIMMI is an installation artwork created jointly by design groups INVIVIA and Urbain DRC in Minneapolis, Minnesota. The work is a large, inflated, suspended sculpture originally located at the Minneapolis Convention Center Plaza. One of the work's primary artists, Allen Sayegh, is a faculty member at Harvard's Graduate School of Design in the Architecture department. He describes *MIMMI*:

Cloud-like in concept, the sculpture hovers 30 feet above the ground, gathering emotive information via twitter from Minneapolis residents and visitors to the plaza. *MIMMI* analyzes this information in real-time, creating abstracted light displays and triggering misting in response to this input, creating light shows at nighttime and cooling microclimates during the daytime. Whether the city is elated following a Minnesota Twins win or frustrated from the afternoon commute, *MIMMI* responds, changing behavior throughout the day and night.

Similar to the two previous works, *MIMMI* is a project that collects and uses mass amounts of social media data to reflect content back to an audience. Fig. 4 shows an installation image of the *MIMMI* object floating above the crowd's heads. What I find so unique about this project is its limitation to a space of 15 miles around the Minneapolis plaza. By only sourcing tweets from such a small location, the results can provide much more relatable content and context to the people who visit. This may also allow residents to see their city in a new way, as the art object reflects the city's mood and activity.



Figure 4: MIMMI, by INVIVIA and Urbain DRC. (Source: <http://www.gsd.harvard.edu/project/mimmi>).

While *We Feel Fine*, *Listening Post*, and *MIMMI* exist in completely different spaces, they all present content that echoes the voice of the Internet. Each project achieves astounding technical feats, making use of modern physical and digital technologies. I look to these works as major inspirations for what is possible to accomplish using publicly available Internet data. As individual works, each demonstrates that data can be used in drastically different ways: through digital, physical, audible and visual representations. I feel that artworks created using public data provide such unique opportunities for viewers to see and interact with each other in ways that were not possible just a short time ago.

APPLICATION ARCHITECTURE

During the early stages of my research into the types of tools and programming languages I needed for development of this project, I experimented with Processing to create various small artworks and code sketches. Processing is an open-source computing language and integrated development environment (IDE) that provides developers with the ability to quickly code and run their project in a self-contained application. The tool is designed to help non-programmers learn how to write an application using the Java language framework to easily provide graphical results. As a starting point, the Processing framework allowed me to prototype some ideas, however, I quickly realized that a much more dynamic system was needed to build a complete web application. At the advice of my advisor, I turned to the MEAN stack application framework.

MEAN is another open-source set of tools designed for building dynamic web applications and websites. The acronym represents four major components that make up the application's architecture: MongoDB, ExpressJS, AngularJS, and Node.js. The first component in the stack is MongoDB, which is a NoSQL database providing developers with the ability to store documents in JavaScript Object Notation (JSON) styled formats. These documents are grouped into collections based upon their similarity of context or purpose, which are then stored in a single database. The second tool in the acronym is the ExpressJS web application framework. Express is the server controller of the application, handling the server requests and responses. The third item in the stack is the AngularJS framework, which is used to control the information sent to the user interface of the application and helps present dynamic content through the manipulation of traditional HTML. Finally, the last tool in the stack is the Node.js framework. Node is a JavaScript runtime environment that makes use of various modules to give more functionality to web applications.

The modules from Node are stored and accessed from an included tool called the Node Package Manager (NPM). NPM provides users with the ability to upload and download modules and libraries written in JavaScript for various uses in web applications.

A few modules that I used in my application are Mongoose, Bootstrap, Bower, Jade, and node-twitter. A possible use for these tools is to expand the capabilities of other major components of the web application. For example, Mongoose is a package that works alongside MongoDB, providing developers with a way to model the data that will be stored in the database through the use of a schema, or outline. Another important module is node-twitter, which is a web client library used to access Twitter's REST and Streaming API. A major reason why the MEAN stack is popular for web development, is it allows programmers to work in a single language: JavaScript. In many other production environments, a project may require 3 or more programming languages to create a dynamic application. By working with tools that use the same programming language, a single developer can code the database content, as well as the client-side and server-side content.

To run the application, the express server is started with a command to execute the server-side code, initiating the connection with the Mongo database and the creation of various collections from Mongoose. Using the node-twitter library, a set of Twitter developer credentials is sent to Twitter.com which provides my application with access to the API. With that access, I am able to request a stream of content from Twitter in real-time, based upon a set of options I can choose. As the stream of content is stored in the database, I can begin using the data for other purposes. First, the tweet content matching the options I specified in the Mongoose schema is run through a function designed to clean the data. I initially remove any usernames from the tweets, which are always prefixed with an "@" symbol. The next items to be removed are any hyperlinks to websites, making sure to find both http and https links. I then search for and remove any html shortcut code

that creates symbols in tweets, such as the ampersand or other symbols not part of the standard ASCII set. All punctuation is removed from the tweets, leaving only the full words. The final content to be removed from the tweets are the words most commonly used in the English language, including most pronouns, conjunctions, prepositions, and some adverbs. The remaining content left in the tweets consists mainly of nouns, verbs and adjectives that provide valuable contextual information. At this point each tweet is separated into a collection of individual words that counted in order to find out how many times each word was used in the Twitter conversation. The whole process of cleaning the tweets and storing the new data occurs once every second, during a 20-minute collection period in which new tweets continue to be collected and stored.

Thus far the content has been collected and stored on the server-side of the application. In order to display the information, the content needs to be moved to the client-side of the application. This problem is solved with the third item in the MEAN stack: AngularJS. Angular provides the ability to request information from the database for use in the development of client-side content. Rather than relying on the use of static data, or data that doesn't change or update, I can fetch data that is constantly transforming and present it in a more dynamic way. A simple output for data using Angular is the presentation and updating of a table or list, such as a to-do list. The application then waits for a user to submit an entry to the to-do list, storing the content in the database as an object. Angular then requests the objects stored in the database to display them as an ordered list using HTML. Adding or deleting items can dynamically change this list, as Angular updates what the user sees on their screen in real time.

For a graphic display of items on the user's screen, a tool is needed to visualize the information. For this project, I decided to use another JavaScript library called D3, or Data-Driven Documents. D3 is a tool that works excellently with Angular to create dynamic, interactive data visualizations through

HTML, CSS and SVGs (Scalable Vector Graphics). D3 is a popular resource due to its gentle learning curve and quick implementation. The combination of Angular and D3 allows developers to bind data to visual objects and manipulate those objects based upon changes in the bound data. D3 visualizations come in a multitude of forms, from bar charts and scatterplots to hierarchical trees and geographical mapping, all of which are capable of animation and user interaction. In order to display the data collected from Twitter, I created a series of rule sets in D3 based upon various design considerations.

DESIGN CONSIDERATIONS

“Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.”

- Edward Tufte, *The Quantitative Display of Information*, 1983

Early in the process of designing the graphic visualization for this application, I understood the need to focus on how viewers would interact with the information and how that information should reflect the data I was collecting. It is important to me that the information provided in the visualization is understandable to people who contribute heavily to social media as well as those who may not participate. The same concept applies to the viewers who have interacted with data visualizations before and the viewers who haven't. With this idea in mind, I began to develop the rules and limitations for how the data should be delivered.

At the beginning of the design stage, I knew that I would be working with mass amounts of data, especially if it was being collected in real-time. This meant that I needed to find a method of visualization that could simplify big data into small information. During this investigation, I was reminded of a popular method of textual visualization that is familiar to social media communities. The “tag cloud” is a tool that displays words at varying sizes to represent the word's popularity or frequency of use. A good example of an early version of a tag cloud is the comparison between the 2002 and 2011 State of the Union Address (Fig. 5). This image allows viewers to compare the content of both speeches to determine what topics both presidents focused on during their addresses. From the tag cloud, I can ascertain that President Bush's speech focused heavily on national security and terrorism, while President Obama's speech focused primarily on the economy.

A second version of a tag cloud can be seen in Fig. 6, which displays the content from the 2015 State of the Union address. This version of the tag cloud is similar to the first, in that it displays the words in varying sizes based upon how many times the word was said in the speech. However, I don't think this version is as successful as the tag cloud in Fig. 5. The 2015 tag cloud displays many words rotated on a 90-degree axis, making it difficult to read. The words are also spaced very close to one another, which also hinders the viewer's ability to understand what is being displayed. While the tag cloud has been designed in many different ways, it still displays content using the same concept: the larger the word, the more it was said.

Using the concept of a word or tag cloud, I began prototyping a graphic using D3 as my visual foundation. Prior to the design phase of the project, I constructed the part of the application that collects, cleans, and organizes the data into individual words and their frequency of use. This provided me with the ability to use example datasets in the D3 prototypes. An example visualization, shown in Fig. 7, displays a test of a radial graphic using data from a #humanrightsday trend on Twitter. The concept for this graphic involved the representation of the top ten words used in the collected data, as well as a space for linked descriptive words. After deliberation with my advisors, I decided that this method of information display didn't accomplish my goals of displaying large amounts of data in a dynamic and engaging way. I began researching methods of using data in real-time to create graphics that change as the data is updated. The D3 tools provided me with the ability to bind data directly to the graphic objects so that as the data changes, the objects also change. With the ability to update the graphics over time, I shifted my attention to how the objects would interact with each other.

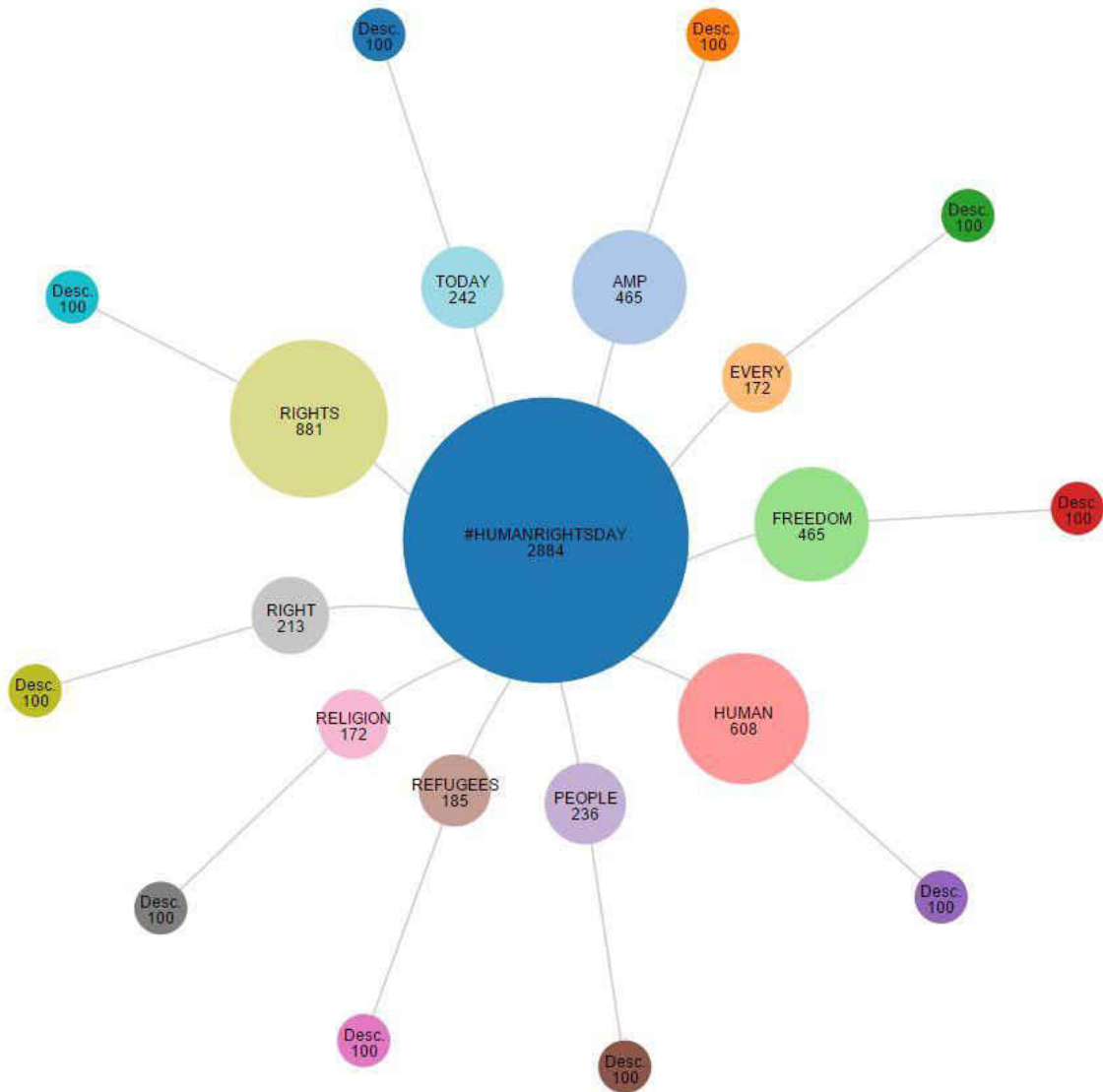


Figure 7: D3 Radial Graphic Prototype, Image by Author.

D3 was an invaluable tool in the process of creating dynamic graphics using real-time data. The visualization library includes a variety of capabilities that developers can use to craft moving and interacting objects. Many of these capabilities are organized into methods of transformation called “layouts”. Scott Murray explains that “The layout methods have no direct *visual* output. Rather, D3 layouts take data that you provide and remap or otherwise transform it, thereby generating *new* data that is more convenient for a specific visual task.” (Murray, 2013) The complete list of layouts includes the bundle, chord, cluster, force, histogram, pack, partition, pie, stack, tree, and treemap methods. In Fig. 4 I made use of the tree layout, which branches out to add new layers of content. I transitioned from using the tree layout to exploring the possibilities of the force layout.

The force-directed layout provided me with the ability to simulate physical forces on the graphic elements shown on the screen. This meant that I could apply physical rules, like gravity, collision detection, and box containment to the objects in the visualization. These rules can be seen in Fig. 8, which is a recent capture of the visualization’s output. This image shows a collection of the top 100 words collected during a 20-minute timeframe, on the evening of the 2nd Presidential debate in 2016. Each circular object in the image represents an individual word, with the radius of the circle determined by the frequency of the word’s use. At the top of the visualization I have displayed both the data and the timeframe from when the data shown was collected. This provides viewers with additional context, and also serves as a timeline reference when comparing one image to others in a series. Due to the application of the force-directed layout, the objects contained in the graphic are able to move and shift around each other as their size is updated over time. The sizes of the objects are updated every 3 seconds as new data is collected and added to the databases.

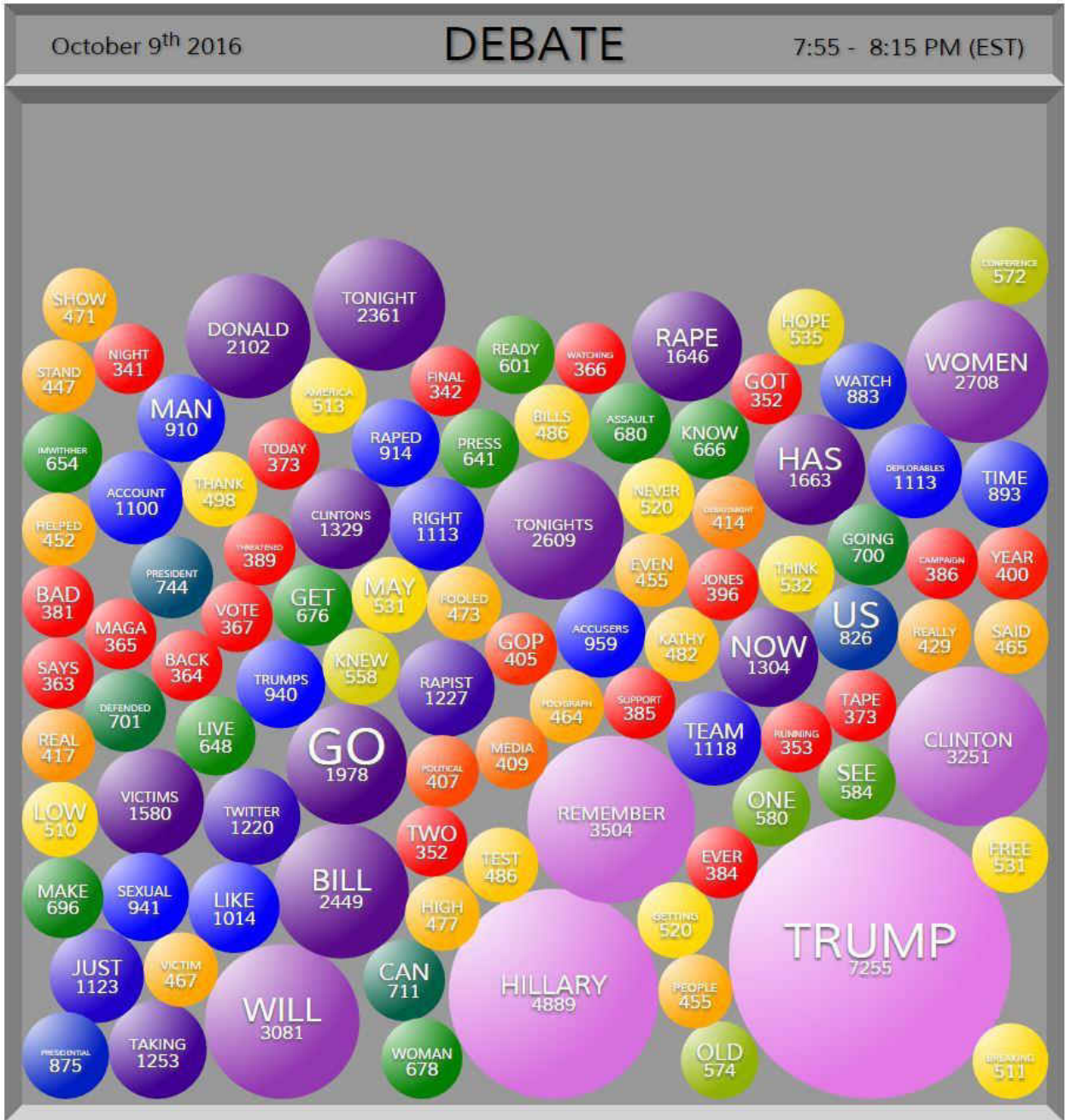


Figure 8: #Debate from October 9th, 2016 from 7:55 to 8:15pm. Image by author.

During the development of the visualization's design, I worked with both a static and a live version of the application. The static version allowed me to use a complete dataset, which was recorded and stored at a prior time. The live version relied on real-time data to be collected and added to the visualization over time. The graphic shown in Fig. 5 was created using the static version of the application. The live display is constructed over time and allows the viewers to see the changes in data as the objects move and expand.

In the live version of the visualization, the colors are determined by how long each object is displayed. The static version differs in that the colors are scaled based upon the represented word's frequency. The colors follow a transition through a color spectrum of red, orange, yellow, green, blue, indigo and violet. This series of colors was selected due to the 20-minute timeline of the visualization's construction. Using a total of 7 different colors, I can set a transition duration to shift from one color to the next every 3 minutes. This allows the objects created at the beginning of the visualization timeline to reach the violet color stage, and the objects added near the end of the timeline to possibly be at the red stage. The purpose for this time-based color scale is to provide a pattern of color similarities to the words that were collected and displayed around the same time. Based on this information, viewers may be able to link similarly colored words to find meaning or a contextual relationship.

The labels added to each object represents both the represented word, as well as the number of times each word was collected. Each label is analyzed for the length of the text, and compared to the radius of its corresponding circle. The label's font size is then changed to fit comfortably inside each object without extending outside the shapes boundary.

As viewers look at the resulting images created by the application, they are presented with a series of objects of varying sizes, colors, and labels. Each one of these attributes is created and evolves as more data is collected and presented. In Fig. 9, two series of images are shown in sequential order. The top row of images displays the 20-minute snapshots from the 2nd Presidential debate on October 9th, 2016. The bottom row displays similar results from the 3rd Presidential debate on October 19th, 2016. As these images are shown together, they provide an opportunity for the viewer to observe patterns or changes in conversation over time. Looking closer, at Fig. 10 and 11, the words and objects that are shown reflect people's reactions and concerns surrounding the debates and the elections. Through the capture and presentation of popular thoughts and words from public discourse, others can look back at these images from a new perspective to better understand our cultural values.



Figure 9: Display of #Debate and #DebateNight series. Image by Author

October 19th 2016

DEBATENIGHT

11:08 -11:28 PM (EST)

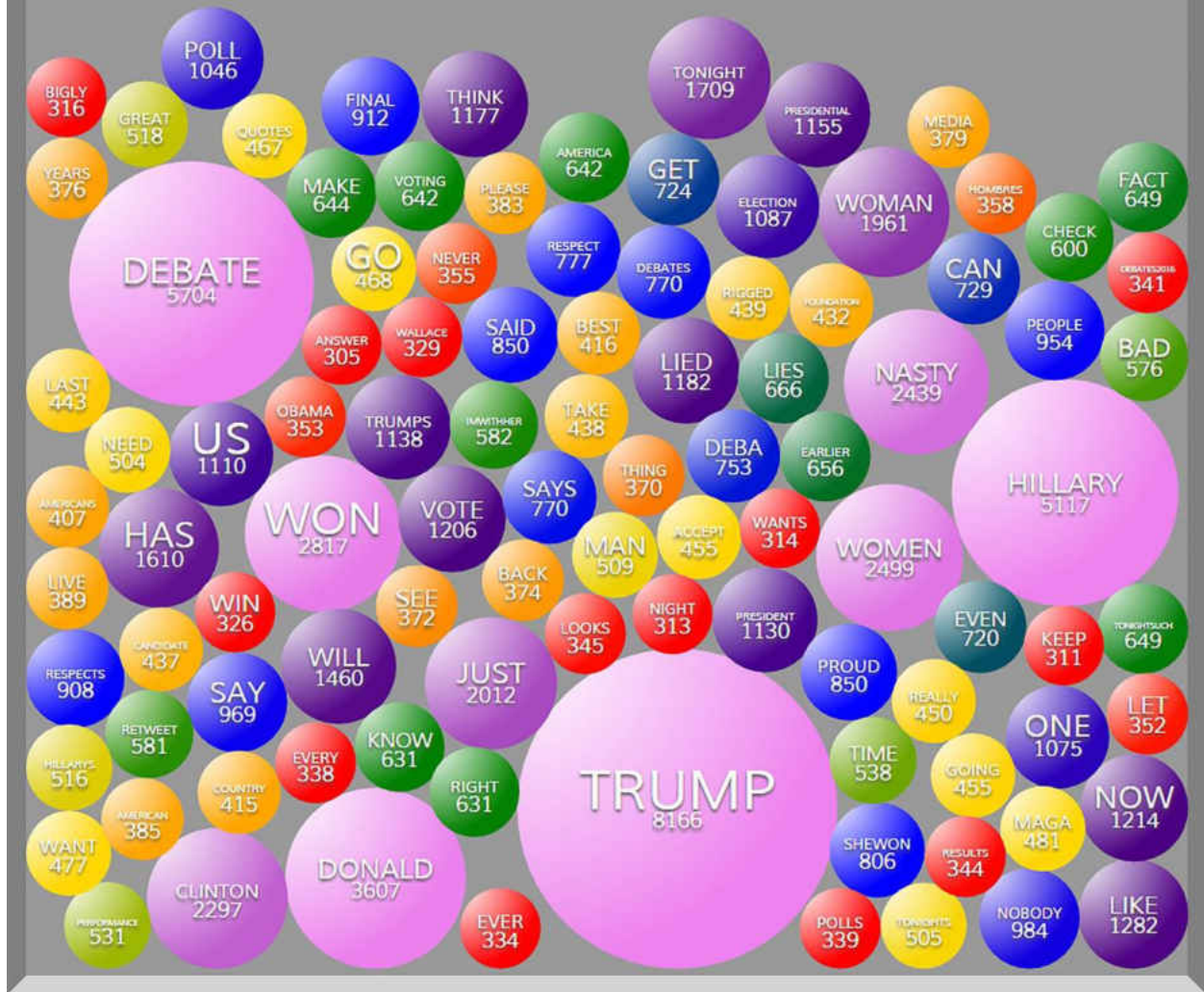


Figure 11: #DebateNight from October 19th, 2016 from 11:08 to 11:28pm. Image by author.

DISCUSSION

Looking back to the start of this project, I asked how I could creatively capture the phenomenon of a trending topic on Twitter. This question sparked an interest into exploring how data from Twitter could provide insights into the behavior and opinions of those who shared their thoughts and ideas. By collecting all of the textual content related to small parts of various topics of discussion, I desired to find a common connection among the vocal population. This exploration provided me with the opportunity to experiment with data and learn about the process of web application development, from start to finish.

With continued development in this project's future, the content produced will gain more autonomous functionality allowing it to be presented to a larger audience. Planned additions include the capability to request a list of the current trends from Twitter and select the topic with the highest tweet volume, to be the focus of the visualization for the duration of that search. The collection's duration will be set by a 20-minute timer, allowing for the generation of 3 visualizations per hour. At the end of the timer, an internal function will record the visualization to an image file and upload that file to Twitter where it can be viewed by the public. After the image is posted back to Twitter, the application will clear the collected data from the databases and restart the search for a trending topic. The complete application can be packaged and uploaded to a web application hosting site, such as Heroku.com, where viewers can observe the process of the visualization being generated. Further updates of the application may include a complete rebuild of the file structure into something more robust and modular, for the purpose of security, interaction, and the addition of more tools and capabilities.

I can see this work being transformed into other iterations at varying scales. I imagine focusing solely on posts discussing a city or a brand for an extended period of time could result in notable insights. As a resident of Orlando, Florida, I know that there are major events happening on a daily basis that hundreds of people talk about on Twitter. This type of application has the potential to be useful to many different groups of observers, including journalists searching for local insights, business concerned about their brand, or other investors looking to better understand the communities' opinions and attitudes.

As social media websites come and go, new ways of sharing and observing information will change. This concept also applies to how web applications are created. New technologies will bring more capabilities for developers to access, analyze, and disseminate information to the public in creative ways. I intend to pay attention to these changes and continue expanding my knowledge of how to use accessible social information to inform others.

CONCLUSION

At the start of this project, I knew that I was looking for a way to learn something from social networks that surrounded me. It was overwhelming, however, to try to find a place to start without getting entangled in a web of content more confusing than enlightening. By thrusting myself into learning how to develop a dynamic web application, I was able to slowly gain focus on what I was capable of and how I could best use the information I was accessing. The process of learning how to build a web application capable of collecting and displaying visual information from data was more extensive than I could have anticipated. Developers are required to pay attention to the shifting landscape of tools, libraries and frameworks that provide the ability to manage massive amounts of data and transform it into small understandable pieces of information.

Through the collection and visualization of data from Twitter, it is possible to present information that allows others to reflect on what has been said about popular topics of discussion. While most subjects that surface on Twitter's trending list do not involve the discussion of politics, major events, or social movements, they still contain popular conversation that can provide insight into how people think. This allows viewers the opportunity to gain insights into the common language and popular opinions surrounding various trends and events. By capturing these moments and reflecting them back to the Internet, I hope to contribute to how others identify themselves and their perceptions of our culture.

This project is designed for anyone that wants to more easily explore the common thoughts of those that communicate on the Internet. My interest in its development stems from a desire to improve other people's ability to understand and explore cultural values via public discourse through the use of visual language.

APPENDIX: COPYRIGHT PERMISSION LETTERS

Daniel McSwain <danielmcswain@gmail.com>

to jjh ▾

Hello Mr. Harris,

My name is Daniel McSwain, and I am an MFA candidate at The University of Central Florida. I've recently started writing my thesis paper, and am looking to use an example image from your WeFeelFine project in my paper. In order to meet the university standards, I need your permission to publish your image in my paper digitally and physically in the university library collection. This is purely for academic use, and will not be used commercially in any way.

In my paper I have included this image:

<http://number27.org/assets/img/pages/wefeelfine/sentence-big.jpg>

Please let me know if you have any questions about the process, or would like to discuss these permissions any further.

Jonathan Harris via gmail.com

to me ▾

Hi Daniel,

Thanks for asking — yes, that would be fine.

I'd love to see the paper once it's complete if you feel like sharing it.

Good luck!

Jonathan

...

--

Jonathan Harris | number27.org

Request for permission to use image from MIMMI Inbox x



Daniel McSwain <danielmcswain@gmail.com>

to asayegh ▾

Hello Mr. Sayegh,

My name is Daniel McSwain, and I am an MFA candidate at The University of Central Florida. I've recently started writing my thesis paper, and am looking to use a photograph from your work in my writing. In order to meet the university standards, I need your permission to publish your image in my paper digitally and physically in the university library collection. This is purely for academic use, and will not be used commercially in any way.

In my paper I have included this image:

http://i87r32c95dp1hz05tig4px11.wpengine.netdna-cdn.com/wp-content/uploads/2016/06/MIMMI_05.jpg

Please let me know if you have any questions about the process, or would like to discuss these permissions any further.

Best Regards,

Daniel McSwain



Allen_Sayegh@invivia.com

to me, asayegh ▾

Hi Daniel,
Sure with proper credit.
Cheers
Allen

Sent from my iPhone

REFERENCE(S)

- Donath, J. S. (1999). Identity and deception in the virtual community. *Communities in cyberspace*, 1996, 29-59.
- Fine Art. (2016). Oxford Dictionaries | English. Retrieved 7 November 2016, from https://en.oxforddictionaries.com/definition/fine_art
- Ghazala, R. (2000). Designing 'Circuit-Bent' Instruments. Anti-theory.com. Retrieved 20 October 2016, from http://www.anti-theory.com/texts/Sound_on_Sound/
- Goffman, E. (1959). *The Presentation of Self in Everyday Life*.
- Hansen, M., & Rubin, B. (2000). The audiences would be the artists and their life would be the arts. MultiMedia, *IEEE*, 7(2), 6,9. doi:10.1109/93.848417
- Harris, J. (2016). We Feel Fine. Retrieved 17 October 2016, from <http://number27.org/wefeelfine>
- Kosara, R. (2007). *Visualization criticism-the missing link between information visualization and art*. In Information Visualization, 2007. IV'07. 11th International Conference (pp. 631-636). IEEE.
- Meloni, J. (2009). Wordles, or the gateway drug to textual analysis. Chronicle.com. Retrieved 11 October 2016, from <http://www.chronicle.com/blogs/profhacker/wordles-or-the-gateway-drug-to-textual-analysis/22781>
- Murray, S. (2013). *Interactive data visualization for the Web*. " O'Reilly Media, Inc."
- Sayegh, A. (2016). MIMMI. Gsd.harvard.edu. Retrieved 10 November 2016, from <http://www.gsd.harvard.edu/project/mimmi/>
- Stallio! (2005). The WordPad Effect. Retrieved Oct 17, 2016 from <http://www.animalswithinanimals.com/stallio/2005/07/wordpad-effect.html>
- Tufte, E. R. (1983). *The visual display of quantitative information* (Vol. 2, No. 9). Cheshire, CT: Graphics press.
- Twitter (2016). About.twitter.com. Retrieved 17 October 2016, from <https://about.twitter.com/company>
- Twitter Developers (2016). Dev.twitter.com. Retrieved 17 October 2016, from <https://dev.twitter.com/rest/reference/get/trends/place>