**INFO 3300 / 5100**

# *PROJECT 2 - INTERACTIVE DATA VISUALIZATION*

**FINAL REPORT**

Team Members

Han Gao (hg359), Ruidi Peng (rp474), Ronin Sharma (rrs234), Songwen Liu (sl2356)

**Due: Friday, November 19, 2021**

# PART 1: PROJECT IDEA

## I. Motivation

The use of social media has consistently increased over the past decades. Twitter, created in 2006, has become one of the most popular social media platforms. By 2021, it has **211 million**[1] active users. On average, there are about 5**00 million tweets**[2] sent every day. Information exchange and spread happen quickly on this platform. However, those tweets tend to vary significantly across users in aspects of contents and word choices. This fact raised our team's interest in exploring ***differences and similarities in tweets among Twitter users who come from different professional backgrounds***.

## II. Description

Our project focuses on the investigation of Twitter users from 6 different professional industries, who are **Sports Analysts**, **Actors/Actresses**, **NBA Players**, **Musicians**, **Entrepreneurs**, and **Cornell Faculties**. The selection of professional industries is based on our team members' interests. We selected 5 Twitter users accounts from each industry (in total, we got 30 users) and analyzed the top meaningful words that are most frequently used among them. The users are

- ***Sports Analysts:*** Skip Bayless, Shannon Sharpe, Pat McAfee, Max Kellerman, Steven A. Smith
- ***Actors/Actresses:*** Emma Watson, Anna Kendrick, Ryan Reynolds, Matt LeBlanc, Robert Downey Jr
- ***NBA Players:*** Lebron James, Kevin Durant, Giannis Antetokounmpo, Stephen Curry, Joel Embiid
- ***Musicians:*** Steve Aoki, Cousin Stizz, Travis Scott, Justin Bieber, Lady Gaga
- ***Entrepreneurs:*** Mark Cuban, Travis Kalanich, Evan Williams, Arianna Huffington, Tony Robbins
- ***Cornell Faculties:*** John W Sipple, Mor Naaman, Soumitra Dutta, Maureen Hanson, Geoff Coates

The project consists of three parts. **The first part** focuses on analyzing the common words in tweets, which are most frequently used by 30 users and referred as "top words shared by all

[1] https://www.socialmediatoday.com/news/twitter-rises-to-211-million-active-users-though-longer-term-growth-target/608958/.
[2] https://www.brandwatch.com/blog/twitter-stats-and-statistics/.

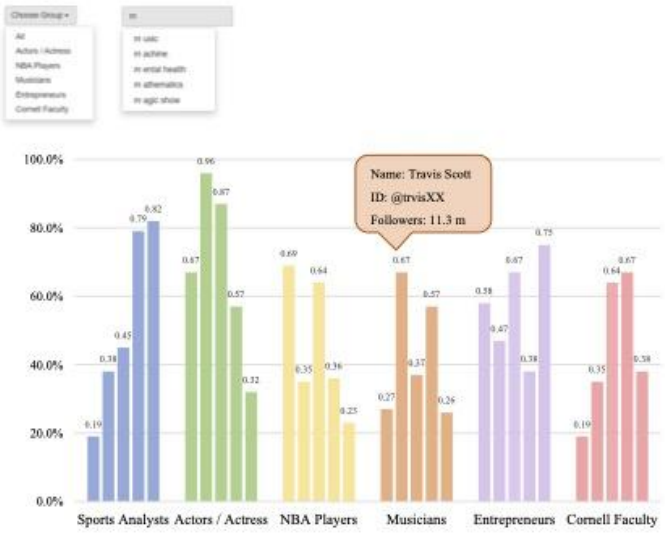
users" in the latter part of the report. To investigate differences in word usage habits across six user groups, the first part also covers the common words most frequently used in each user group. **The second part** explores different word usage habits across different users. It presents the percentage of each user's tweets containing the top words shared by all 30 users and shared inside each user group. **The third part** analyzes the frequency of state names mentioned by 5 users from the Sports Analyst user group. We attempt to understand whether there is a correlation between each sports analyst's focus area and his/her tweet contents.

## III. Design Sketch

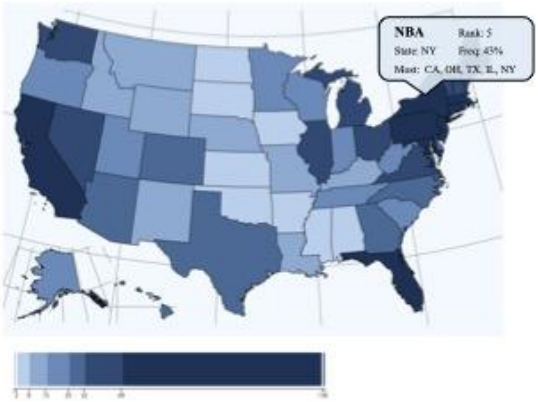
Based on the focus of each part, we decided to use three different types of graphs to visualize data, which are **Word Cloud**, **Bar Chart**, and **Choropleth Map**. While brainstorming the project idea, we did a design sketch as a guide for the latter coding and design process. The design sketch is shown below.

- **Graph 1: 50 Most-searched Words**


- **Graph 2: Frequency of Occurrence by Users**



- **Graph 3: Frequency of Occurrence by States**



# PART 2: DATA SOURCES & PROCESSING

## I. Data Source

In this project, we used two data sources: ***Twitter API*** and ***U.S. topoJSON*** dataset.

- ***Twitter API*** allows us to scrape tweets from individual Twitter users. We used the **Snscrape Web Scraper** to collect the data. This allowed us to provide a Twitter username and obtain a fixed number of tweets for that user. We obtained one million tweets for each individual among the 30 users that we chose.

- ***U.S. topoJSON*** data file, encoding the topology of the United States, is used to draw the U.S. map in the third visualization. The data file, ***us-smaller.json***, was provided by the professor during the lecture.

## II. Data Processing

### STEP 1: DATA CLEANING

After collecting the data, we performed some steps to clean the data. The web scraper provided us with a lot of data about each tweet, but we didn't need all of this information. We filtered the data by obtaining the following features:

- ➢ Twitter User Object
- ➢ Date Object
- ➢ Content
- ➢ Reply Count
- ➢ Retweet Count
- ➢ Like Count

For each user, we extracted the Twitter **username** from the Twitter User Object. We also extracted the **week**, **month**, and **year** attributes from the date object and stored them as individual features within the dataset. Lastly, we created an updated JSON file for each user. The cleaned data files can be found in the folder, *cleaned-data*, or be accessed by the link: https://github.com/roninsharma25/info-5100-project2/tree/main/cleaned-data.

## STEP 2: COUNT WORD FREQUENCY

The first part of the project visualizes the most frequently used words among 30 Twitter users, so we processed the collected data to count the frequency of each word. We used ***pandas*** to manipulate the data and get the frequency count for each word.

We started by reading tweets from each user's cleaned data file. After reading in all *.csv* files as a data frame, we stripped away all columns except the ***content*** column. To filter out meaningless words, such as "so," "not," "oh," and etc, ***a stopword file*** containing 700+ English common stopwords was used to create a stopword list. We iterated through each word in each tweet *content*, counted its frequencies, and stored them in a Python dictionary shown below

```
[('shannonsharpe', 78198),
 ('now', 25405),
 ('like', 20000),
 ('think', 16002),
 ('time', 14521)]
```

Then, we sorted the dictionary in descending order of frequency. Eventually, we converted the dictionary to a JSON file named ***word-freq-sorted-full.json***, which is a more d3 friendly format.

## STEP 3: SELECT MEANINGFUL TOP WORDS

The second part of the project analyzes the differences and similarities of top words that are most frequently used by the 30 users. Even though we added a stopword list in the previous step to filter out meaningless words, we still gathered some meaningless nouns and verbs, such as "take," "do," "have," etc. Therefore, we decided to **manually select** top words from the word-freq-sorted-full.json data file to create a list of **10 words** that are most frequently used by the 30 users and stored them in a JSON file named ***top-words-AllUsers.json***. Similarly, we also found **5 top words** shared by 5 individuals in each user group and stored them in separate JSON files. The final top-word lists for all users and each user group are shown below

- ***All Users:*** ["game", "love", "thanks", "win", "undisputed", "lol", "happy", "bro", "haha", "damn"]
- ***Sports Analysts:*** ["lebron", "nfl", "patmcafeeshow", "brady", "broncos"]
- ***Actors / Actresses:*** ["love", "aviationgin", "happy", "Deadpool", "thejudge"]

- ***NBA Players:*** [“lol”, “bro”, “happy”, “win”, “congrats”]
- ***Musicians:*** [“steveaoki”, “music”, “album”, “aokijump”, “party”]
- ***Entrepreneurs:*** [“uber”, “twitter”, “business”, “work”, “world”]
- ***Cornell Faculties:*** [“data”, “cornell”, “research”, “students”, “congrats”]

## STEP 4: COMPUTE POST PERCENTAGE

Since the selected 30 Twitter users have different numbers of followers and tweets. For the second part of the project, **to avoid BIASES**, we decided to visualize the **percentage of posts** containing the top words instead of frequencies. We used ***pandas*** to complete the task. We used a counter for each individual. We looped through all tweets for each user and incremented the counter for each tweet that included the top words. Then, we converted the count to a percentage by dividing by the total number of tweets for the user. We stored users’ Twitter accounts, top words, and their corresponding percentages in a dictionary, which is shown below

```
{
        RealSkipBayless: {
                undisputed: 0.1814873651544259,
                amp: 0.03426555342101374,
                win: 0.10981601891532437,
                ...
        }
        ...
        maxkellerman: {
                undisputed: 0.0017457084666860634,
                amp: 0.14562118126272913,
                ...
        }
        …
}
```

Eventually, we converted the dictionary into a JSON file named ***top-words_percent.json***.

## STEP 5: COUNT STATE NAME MENTIONED

For the third part of the project, we analyze the frequencies of each state name mentioned by 5 Sports Analysts, so we need to count the number of times each of the Sports Analysts mentioned a state in a tweet. We started by creating an empty dictionary. The keys were states and the values were also empty dictionaries. We populated each sub-dictionary with each user and the

percentage of times that user tweeted about a state. Again, we looped through all the tweets for each user to perform this task. Here is an example of the final JSON structure:

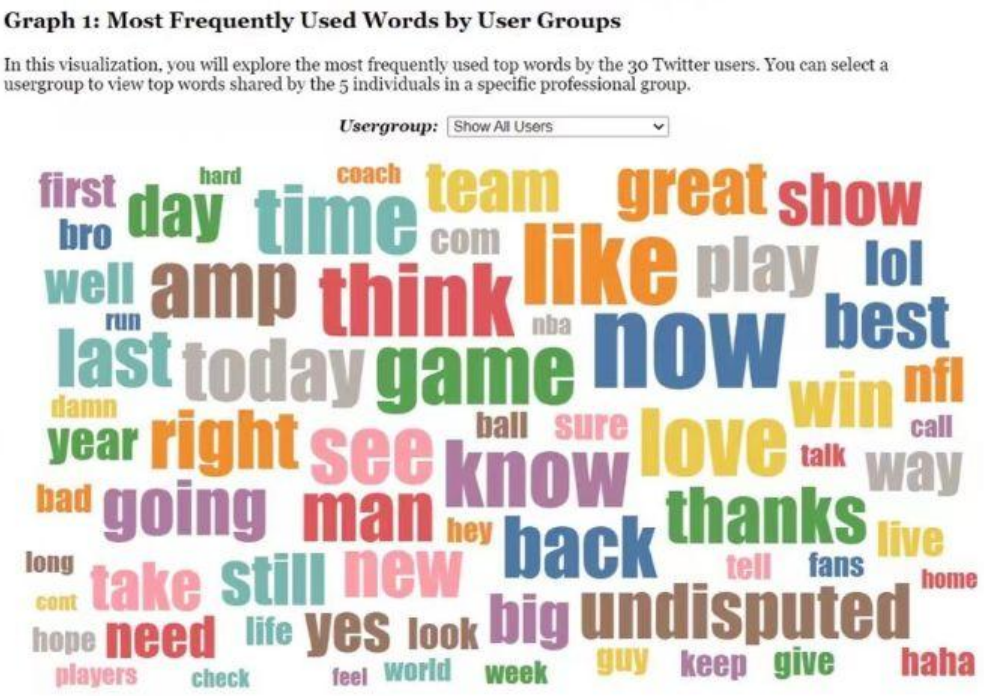
```
{
        New York: {
                Twitter User #1: Tweet Mention Count,
                Twitter User #2: Tweet Mention Count
        }

        New Jersey: {
                Twitter User #1: Tweet Mention Count,
                Twitter User #2: Tweet Mention Count
        }
}
```

# PART 3: VISUAL DESIGN RATIONALE

## I. Plot 1 - Word Cloud

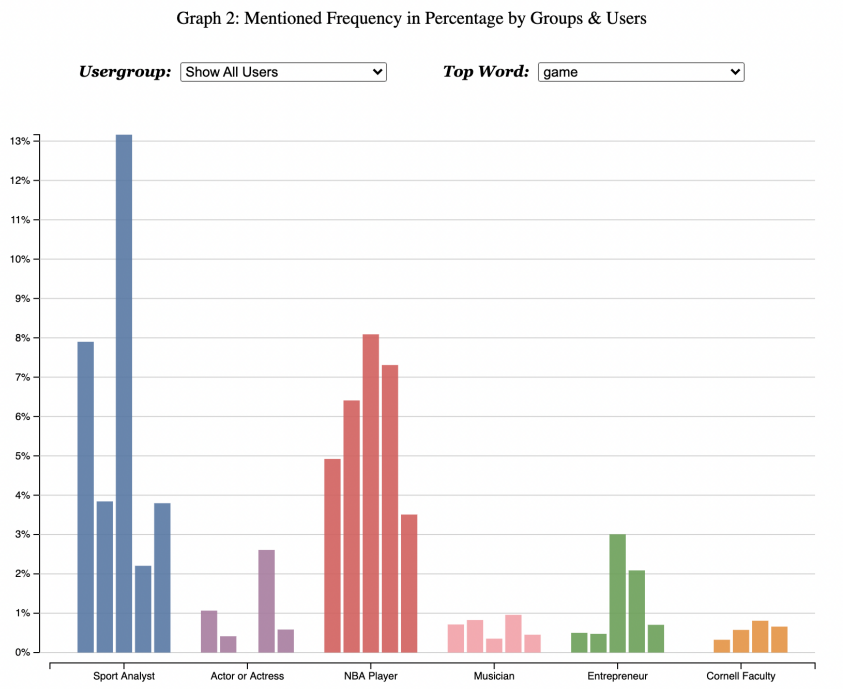
The first plot utilizes a Word Cloud to visualize the top words most frequently used in tweets by the 30 users that we selected. The final outcome is shown below

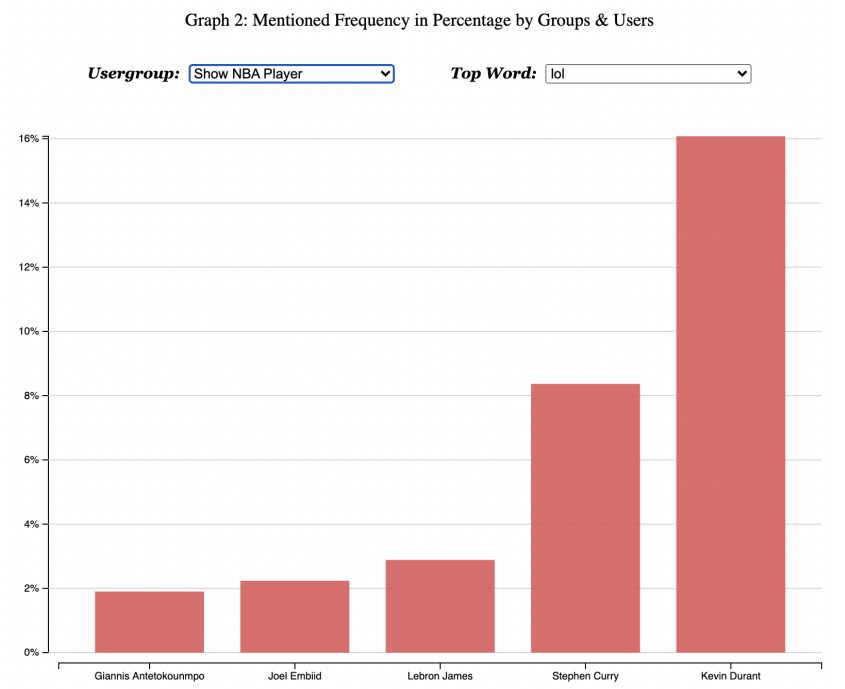



- **Marks:**
  - Texts
- **Channels:**
  - Font sizes vary corresponding to frequencies of the word mentioned in tweets
    - We chose to user ***d3.scaleLog()*** function to compute the scale for the font size, considering frequencies of top words are polarized. The *d3.scaleLog()* function will make difference more **visually visible**.
- **Colors:**
  - The word color is generated by the ***d3.scaleOrdinal()*** function. To make the word cloud more visually appealing, we created a color palette with 10 different colors with various hues.
    - Color Palette: ["#4e79a7", "#f28e2c", "#e15759", "#76b7b2", "#59a14f", "#edc949", "#af7aa1", "#ff9da7", "#9c755f", "#bab0ab"]

## II. Plot 2 - Bar Chart

The second plot is a bar chart visualizing the percentage of tweets that contain the top words for the 30 users that we selected. The final outcome is shown below



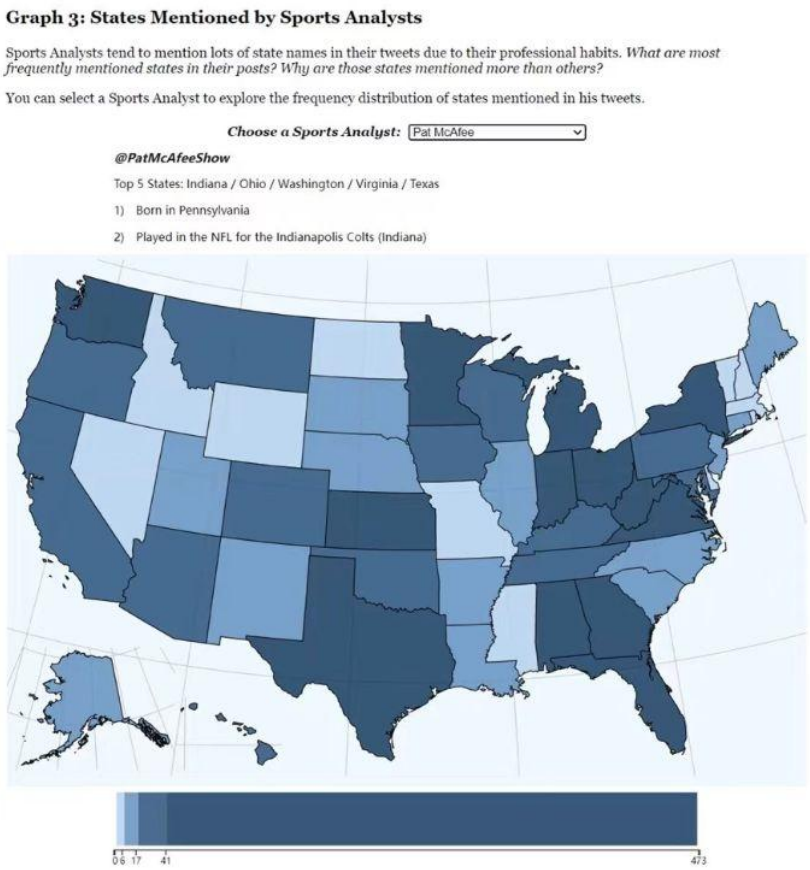



- **Axes:**
    - The x-axis represents **6 user groups** when the option “Show All Users” is selected. It represents **5 individuals** in a user group when the option “Show [Usergroup]” is selected.
    - The y-axis represents the percentage of users’ tweets that contain a specific top word selected by the viewer.
- **Marks:**
    - Rectangles, Gridlines
- **Channels:**
    - Vertically aligned positions vary according to the percentage of each user’s tweets that contain the top word.
    - Horizontally aligned positions vary according to the user group that each user belongs to.
    - Bar colors vary corresponding to the user group.
- **Colors:**
    - The bar’s color is generated by the ***d3.scaleOrdinal()*** function. We created a color palette with 6 different colors to categorize users by groups. To achieve an internal consistency with other plots, the 6 colors were picked from the color palette created in Plot 1.
        - Color Palette: ["#4e79a7", "#f28e2c", "#e15759", "#59a14f", "#af7aa1", "#ff9da7"]

## III. Plot 3 - Choropleth Map

The third plot is a choropleth map demonstrating the frequencies of each state mentioned by the 5 Sports Analysts in their tweets that we selected. The final outcome is shown below.

- **Marks:**
    - Boundary lines, graticules
- **Channels:**
    - The state color saturation varies corresponding to the frequencies of being mentioned by Sports Analysts.



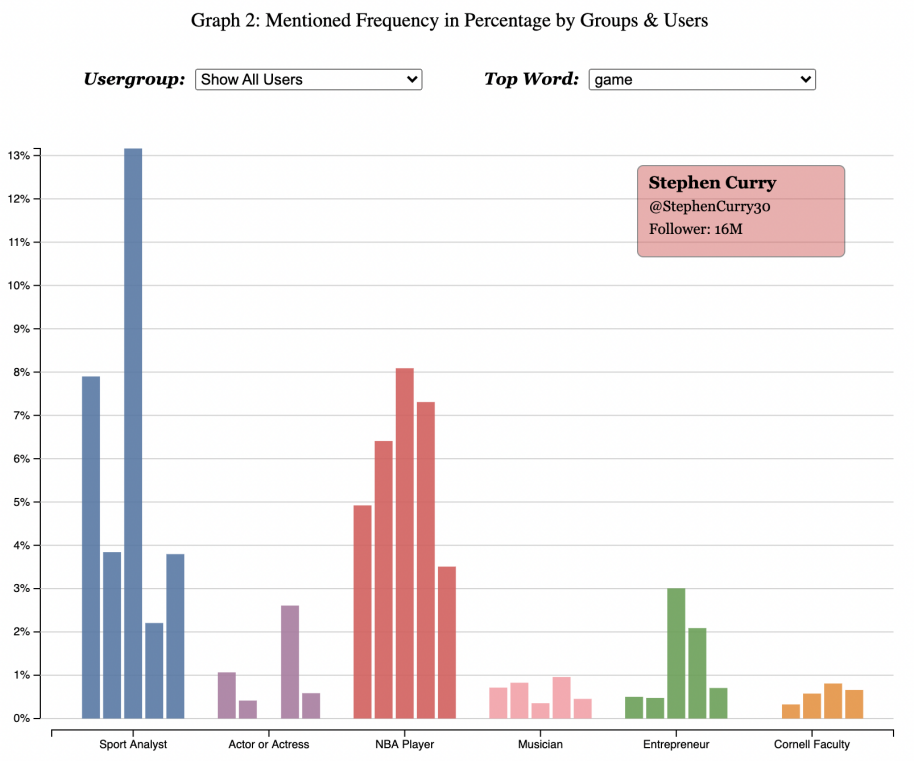

- **Colors:**
  - The color scale is built by the ***d3.scaleQuantile()*** function, which splits the domain into intervals so that the same number of points fall into each interval. We chose this function, considering the data points are not evenly distributed.
  - We created a color palette with 4 colors varying in saturation. We chose the same blue as Plot 2, which was used to represent the Sports Analyst group, to achieve internal consistency.
    - Color Palette: ["#B3D0ED","#5A8BBC","#1B4875","#082F56"]
- **Color Legend:**
  - The color legend was created to help viewers interpret colors in the map.
- **Information Boxes:**
  - We decided to include an information box on the top of the map to provide viewers with information about the top 5 states frequently mentioned by Sports Analysts and each analyst's fun facts. We thought that this would help viewers understand why each Sports Analyst mentions some states more than others.

# PART 4: INTERACTIVE DESIGN RATIONALE

## I. Plot 1 - Word Cloud

The Word Cloud plot includes two interactive elements: a **Dropdown Menu** and an **Animation**. The **dropdown menu** is used to allow viewers to select a user group from six available options and view top words frequently mentioned by 5 users in this group. This interaction allows viewers to understand differences in word usage habits among individuals from different professional industries. We intentionally chose the dropdown menu to avoid mistyping or misspelling errors. The **animation** was designed as a transition between viewers' selections of user groups to notify viewers that they have successfully selected an option.

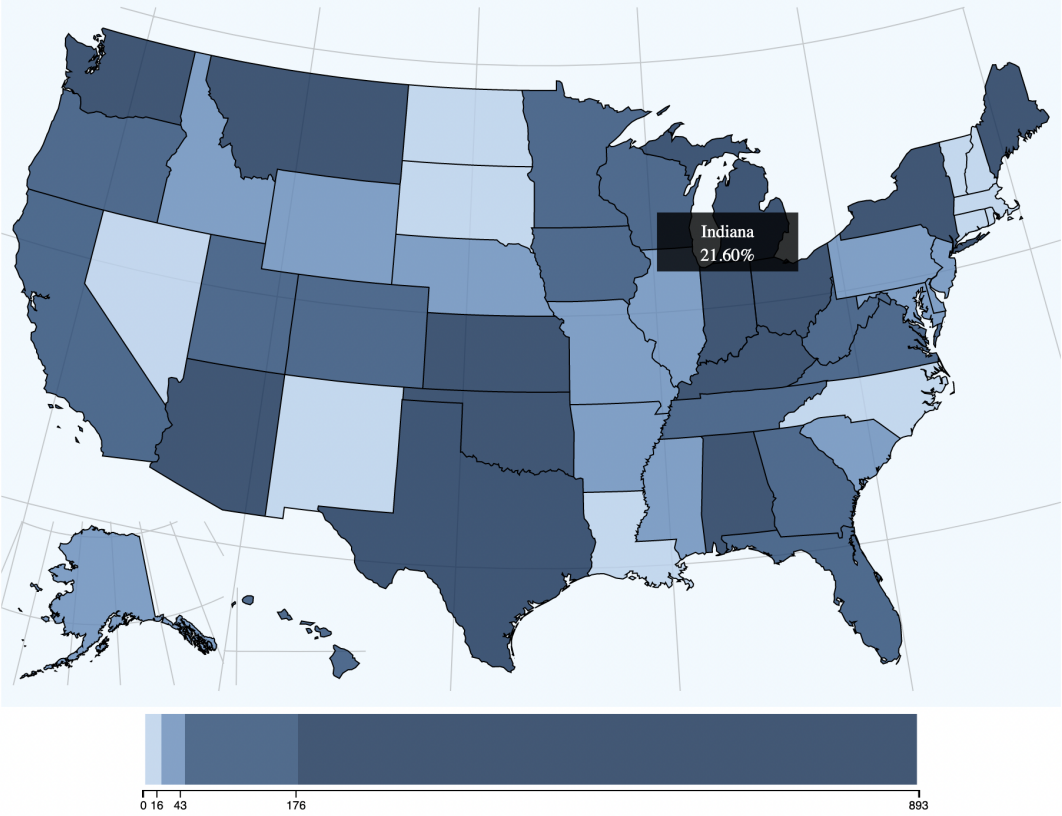
## II. Plot 2 - Bar Chart




The Bar Chart plot also includes two interactive elements: two **Dropdown Menus** and a **Mouseover Event**. Viewers can select a specific user group and pick a top word to see the percentage of tweets containing the word from 5 users in the group. This interaction allows users to easily view the visualization of groups or words that they are interested in.

When viewers mouse over a bar, the plot will show a pop-up box with the corresponding user's real name, Twitter account, and the number of followers. The information helps viewers know more about the user. We decided to use a pop-up box so that viewers won't be overwhelmed by too much information.

## III. Plot 3 - Choropleth Map




The Choropleth Map includes two interactive elements: a **Dropdown Menu** and a **Mouseover Event**. Viewers can select a specific Sports Analyst from 5 available options to view the frequency of states mentioned in his/her tweets. We designed this interaction so that viewers can see the frequency distribution of each analyst. Besides, when viewers mouse over a state, they will see a pop-up box showing the state name and percentage of tweets mentioning it. This interaction provides viewers with more information about the data distribution without overwhelming them.

# PART 5: THE STORY

## I. Story 1 - Positive Words Shared by Public Figures

Nowadays, online hate speech has become a severe problem, which negatively affects people's mental health conditions and social stability. However, while exploring the most frequently mentioned top words in their tweets, we found that lots of positive words had high usage frequencies, such as "thanks," "love," "congrats," "wins", etc. This is a positive signal suggesting possible mitigation of online hate speech. The 30 Twitter users that we selected are mainly public figures with large numbers of followers, who can bring huge impacts on the public. If they continue using those positive words on social media, they can build an example for the public of talking and sharing positively, encourage more online positive speech, and possibly reduce hate speech. However, ***to what extent public figures can impact the public online speech*** should be examined by more in-depth studies and data analyses.

## II. Story 2 - Impact of Professional Background on Word Usage Habits

While exploring differences in the most frequently used top words among different professional groups, we found that users' professional backgrounds would affect their tweet contents and word choices. While tweeting, users tend to mention their professional-related content. Actors/Actresses often introduce TV shows and films in their tweets, such as "*The Judge*" and "*Deadpool*". Musicians use words, such as "music" and "album" often. Entrepreneurs frequently mention business brands in their tweets, such as "Uber" and "Twitter." Cornell Faculties used words, "Cornell," "students," and "research," frequently.

Moreover, users' professional backgrounds also influence their tweeting or speaking styles. Sports Analysts and NBA Players have similar tweeting styles, who use lots of casual words and abbreviations, such as "bro," "lol", and "haha." This may be caused by the overlap between their professional industries. In contrast, Cornell Faculties tend to not use those casual, social media words. Especially, only one Cornell Faculty used "damn" in his tweets, and the usage frequencies in percentage are lower than 0.5%. This may be due to that their professions require them to be rigorous and strict.

## III. Story 3 - Reflection of Personal Information in Tweets

While looking at the Chroplopleth Map showing the frequency distribution of state names mentioned by the 5 Sports Analysts, we found that **Maine**, **Indiana**, **Washington**, **Texas**, and **Ohio** are the top 5 states most frequently mentioned in their tweets. These states are the top sports-active states with great amounts of athletes. This fact provides a possible explanation of why these 5 states are frequently discussed by Sports Analysts.

When we took a look at the top 5 states mentioned by each Sports Analyst, we found some **correlations between his personal information**, such as born cities, favorite sports teams, and job locations, **and states mentioned in his tweets**. For example, the top 5 states mentioned by Skip Bayless are Texas, Oklahoma, Washington, Ohio, and Arizona. Oklahoma is the place where Skip Bayless was born, and he is a huge fan of the Dallas Cowboys NFL team, which is located in Texas. Pat McAfee played in the Indianapolis (Indiana) Colts NFL team before, and Indiana is the top state mentioned in his tweets. What's more, both Steven A. Smith and Max Kellerman host ESPN shows in New York, which is one of the top two states mentioned by them. This discovery provides an interesting insight that **Twitter may be a mirror of a user**, which reflects his/her personal or professional background and show his/her interest.

## Appendix I - Team Member Contribution

### I. Team Role Assigned

- Han Gao: Report Writing, Plot 2 Visualization Coding, Final Presentation
- Ruidi Peng: Data Processing, Plot 1 Visualization Coding
- Ronin Sharma: Data Fetching, Data Cleaning, Plot 3 Visualization Coding,
- Songwen Liu: Design Sketch, HTML & CSS Style Coding

### II. Time Distribution:

- Data Collecting & Cleaning: 3 hours
- Data Processing: 5 hours
- Visualization Coding: 12 hours
- Visualization Styling: 2 hours
- Report Writing: 3 hours

### III. Project Timeline

| Date | Task | Team Member |
|---|---|---|
| Tuesday,<br>10/26/2021 | **1st Team Meeting**<br>*Location: Zoom*<br>• Get to know each other<br>• Assign roles<br>• Discuss design ideas | Han Gao,<br>Ruidi Peng,<br>Ronin Sharma,<br>Songwen Liu |
| Wednesday<br>10/27/2021 | Idea Brainstorming<br>• Each team member comes up with at least one project idea<br>• Write the project idea *Title*, *Data Source*, *Explanation* in the document | Han Gao,<br>Ruidi Peng,<br>Ronin Sharma,<br>Songwen Liu |

| | | |
|---|---|---|
| Thursday<br>10/28/2021 | **2nd Team Meeting**<br>*Location: Zoom*<br><br>● Discuss all project ideas<br>● Pick one final idea for Project 2<br>● Discuss and consider design elements | Han Gao,<br>Ruidi Peng,<br>Ronin Sharma,<br>Songwen Liu |
| Friday<br>10/29/2021 | Report 1 editing & submission | Han Gao |
| Tuesday<br>11/02/2021 | **3rd Team Meeting**<br>*Location: Zoom*<br><br>● Have the design sketches ready for feedbacks and discussions<br>○ Include basic *HTML & CSS style elements* in the sketch<br>● Have the cleaned datasets ready | Ronin Sharma,<br>Songwen Liu |
| Friday<br>11/05/2021 | Report 2 editing & submission | Han Gao |
| Saturday<br>11/06/2021 | Coding | Han Gao,<br>Ruidi Peng,<br>Ronin Sharma |
| Sunday<br>11/07/2021 | | |
| Monday<br>11/08/2021 | | |
| Tuesday<br>11/09/2021 | **4th Team Meeting**<br>*Location: Zoom*<br><br>● Weekly check-in with coding progress<br>● Discuss the design rationale | Han Gao,<br>Ruidi Peng,<br>Ronin Sharma,<br>Songwen Liu |
| Monday<br>11/15/2021 | Finalize coding | Han Gao,<br>Ruidi Peng,<br>Ronin Sharma |
| Wednesday<br>11/17/2021 | Finalize Styling | Songwen Liu |

| | | |
|---|---|---|
| Thursday<br>11/18/2021 | Report Writing | Han Gao |
| Friday<br>11/19/2021 | Presentation | Han Gao |

## Appendix II - Explanation of Final Submission Folders & Files

The final submission folder contains the following files:

- ***datasets folder*** contains all data files used to draw three visualizations
- ***data processing scripts*** folder contains all scripts used to collect, clean, and process the data
- ***index.html*** is the HTML page continuing the project visualization
- ***Final Outcome*** is an image file showing the final outcome of our visualization
- ***Project 2*** Final Report is a pdf file containing the report for the project