DSCI 550
04/27/2023
Professor Mattmann
Assignment 3
Team 12: Jai Agrawal, Daniil Abbruzzese, Todd Gavin, Tania Dawood

Pixstory Visualizations

For Assignment 3, we used our analysis of the original PixStroy dataset used in Assignment 1 to create an interactive set of visualizations using the Data Driver Documents (D3) framework. Our objective was to use the MEMEX Image Space and GeoParser applications to explore our data and find similarities to pull insights. Additionally, we created a mini site to display our visualizations and it can be found here: Team 12 Visualizations

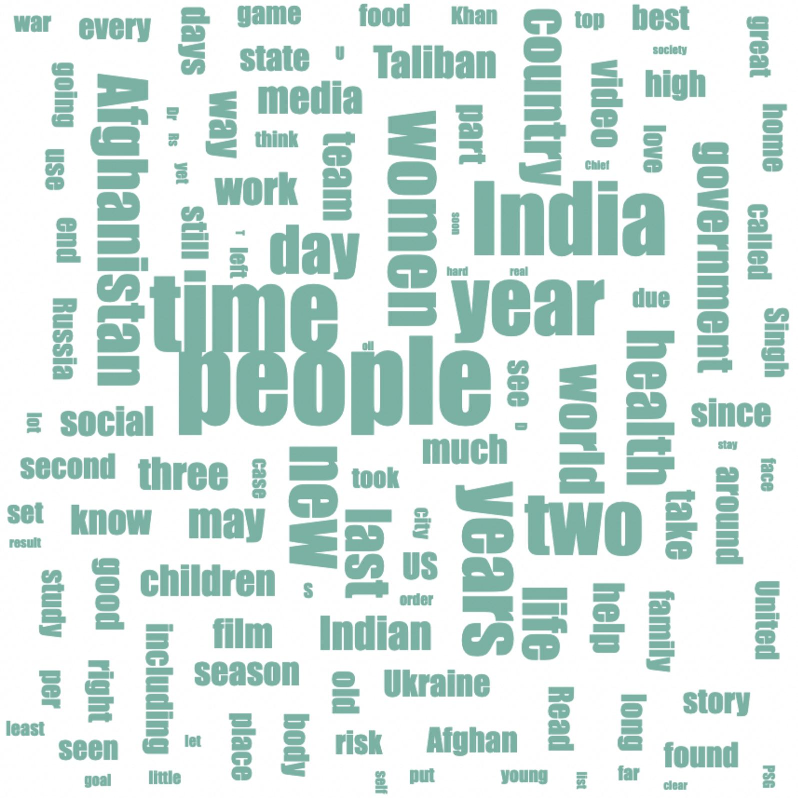
For our visualizations, we decided to use a Word Cloud to test our hypothesis on the toxicity of PixStory, a Bubble Chart to dig deeper into the demographics and interests of our users, a word map to identify the most toxic countries in terms of toxic posts, a bar chart to get further information on interests and finally a hierarchical bar chart to better understand the proportion of narratives belonging to the different languages detected in assignment 2 on PixStory. Overall, we used features such as Tika Lang Detect, Translated Narrative, and finding the average age of posters to analyze the data and determine if PixStory users were adhering to the platform's initiative of being a clean social media platform or not. Our analysis from assignments 1 to 3 has helped us prove that PixStory users are, in fact, adhering to the policies, and it is not a toxic platform.

Once we created all our visualizations, we were able to pull a number of insights from each which are discussed below. (Question 1 and 2)

Word Cloud

We decided to use a Word Cloud to test our hypothesis, which was that if the most popular words used by posters were flagged by the sarcasm, hate speech or GLAAD flags we created in assignment 1, the dataset was more likely to be toxic. In order to create the word cloud, we first created subsets of our dataset of 20,000 rows each and extracted all the words used in the “Narrative” section of each Pixstory post. Once we had all the narratives extracted, we then removed stop words from our data provided by NLTK as well as created a list of our own stop words. This is an important step when creating a word cloud as it helps improve the accuracy of the cloud. WIthout removing the stop words, our word cloud risks being populated by words like “I”, “The” and “And” as those are some of the most commonly used words in social media posts. Once we had completed all the pre-processing steps, we used the observable D3 word cloud visualization to generate our cloud.

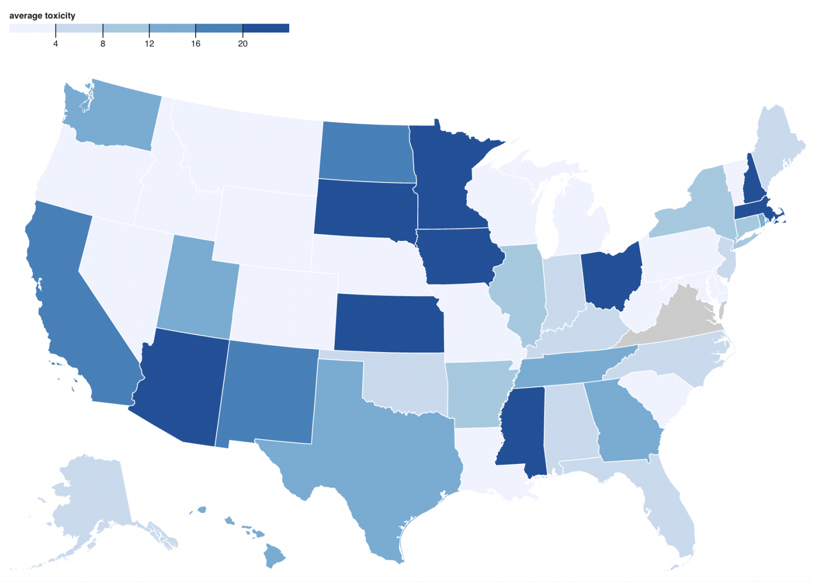
Our visualization showed that some of the most commonly used words in the PixStory dataset were "India", "People", "Afghanistan", "County", and "Life", among others as shown below. Based on this analysis, we were able to determine that users are adhering to PixStory's goals of a clean and non-toxic social media platform. Through this visualization, we can also tell that users are often posting about world news as we can see "Ukraine" and "war" are also common words used in posts.



Word Cloud Testing Toxicity

U.S. State Choropleth

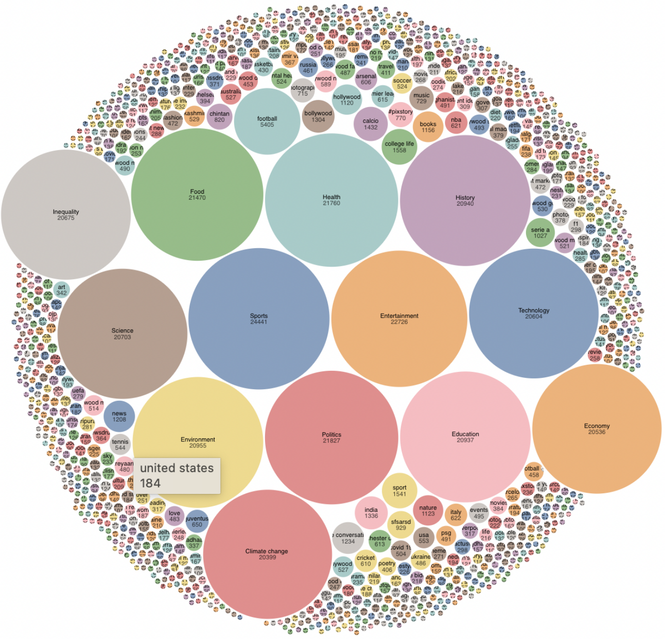
To gain some more insight into the toxicity of PixStory, we created a U.S. State Choropleth to determine if there is a variation in toxicity levels across the different states of the US. The goal was to determine if there are any particular states with more individuals who create toxic posts than others. The approach was to calculate the average toxicity score for each state and map it to a color scale, where the darker the color, the higher the average toxicity score. We found the states with the highest toxicity scores were New Hampshire, Massachusetts, Ohio, and Kansas. The second most toxic states were found to be Minnesota, Iowa, and South Dakota, and Arizona and the least toxic states were Oregon, Nevada, and Wyoming. Further analysis could allow us to identify which topics tend to draw the most toxicity in these 'toxic states'.



U.S. State Choropleth Measuring Toxicity

Bubble Chart

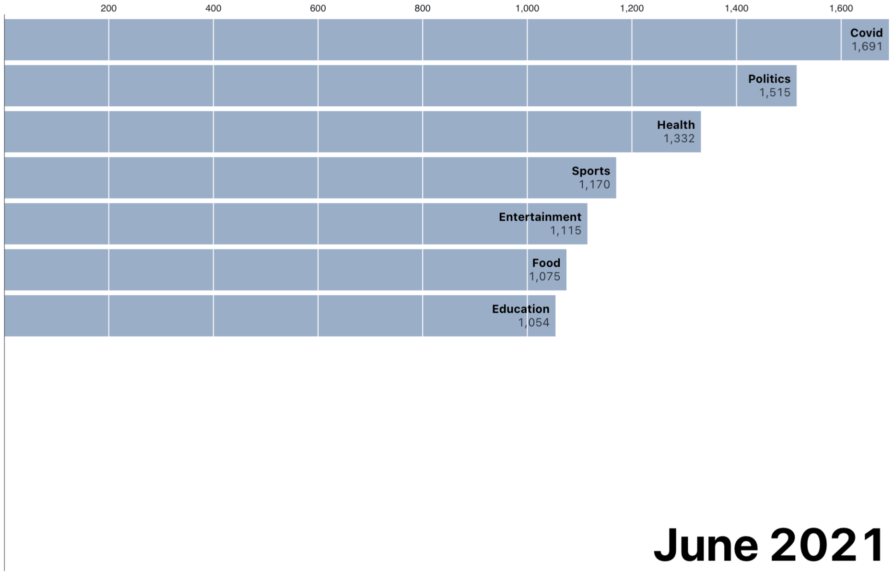
To dig deeper into the demographics and interests of PixStory users and how the platform was being used, we decided to create a bubble chart. To create the bubble chart, we separated the interests of each post and created a dataframe to represent interests, number of related posts and the average age of posters on those interests. Once we had our dataframe ready, we were able to use the D3 bubble chart visualization to represent interests and how many people posted on those interests. The bubble chart showed that the average age of posters on a majority of interests were around 23 to 25 years old, and the most popular interests across posts were: sports, entertainment, politics, health, food, environment, history, education, science, and inequality. The fact that inequality was an interest that many posters had, showed us that the PixStory platform is a progressive form of social media, allowing its users to talk about real world issues in a positive way. Additionally, sports-related content, particularly football and specific sports teams, were popular among the audience. The dataset also revealed interests related to specific countries, suggesting a diverse and global audience. Some interests were more specific to the audience, such as those related to Bollywood or COVID-19 and mental health. Lastly, there are some interests that appear to have a lower sentiment score, including Taliban Government and Alt News. Overall, these insights also help further prove that PixStory is not a toxic platform. Additionally, insights can be used to inform marketing or content strategies by understanding the demographics and interests of the target audience.



Bubble Chart for Topics

Racing Bar Chart

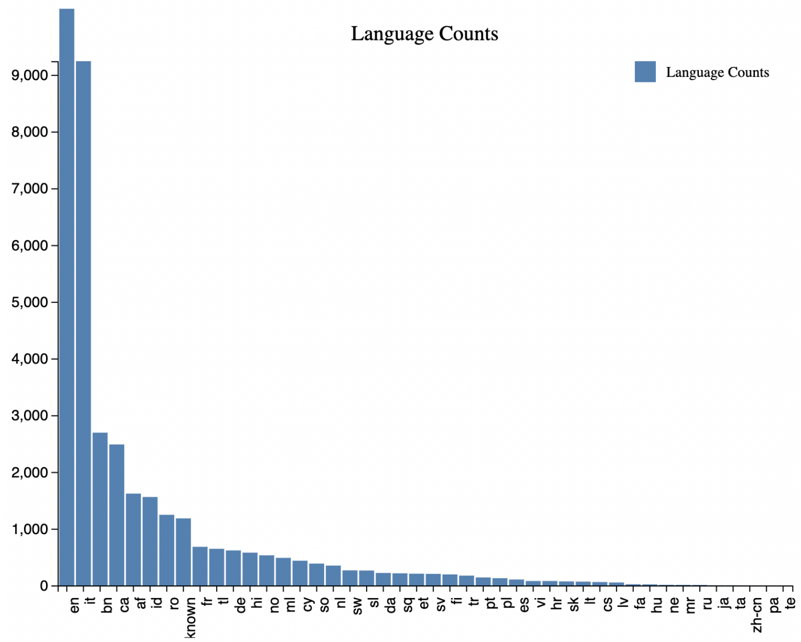
We then created a Racing Bar Chart to track the engagement levels of the most popular interests over time. As shown in the bubble chart, the most popular interests included sports, health, entertainment, food, politics and education. Our goal was to determine if any particular interests spiked during specific event months. For example, we would expect the engagement levels of the 'health' interest to increase during the months of a COVID-19 outbreak. Some interesting insights from our visualization include that 'health' was in fact the most talked about topic in 2020 during the height of the pandemic. Then in January 2021, the US Capitol raids caused the 'politics' topics to spike, spiking even farther than the 'health' topic. Entertainment" saw a significant increase in July 2021, becoming the second most popular interest. Then in September 2021, the 'Sports' topic overtook 'Health', becoming the most popular interest, and by the end of the month, 'Health' had become the most popular topic once again, possibly due to the abortion ban in Texas. Finally in November 2021, 'Entertainment' spiked to the most popular interest, followed by 'Sports' in January 2022, and it remained the most popular interest for the majority of the year.



Race Bar Chart for Topics

Hierarchical Bar Chart

Lastly, we decided to create a hierarchical bar chart as a means of displaying the proportion of narratives belonging to different languages detected on PixStory. We chose this visualization because it enabled us to identify the languages spoken by PixStory users, the frequency at which the languages were used on the platform, which we detected in assignment 2. In order to create the bar chart, we used the 'Narrative TikaDetect" feature from our dataset to count the number of occurrences of each detected language. Our findings revealed that aside from English, Italian, Bengali and Catalan were the most commonly used languages. We found this surprising as these languages are not the most widely used languages on the internet and we expected, for example, French to be a more common language, which it was not. Furthermore, the bar chart provided us with information on what parts of the world PixStory is a popular platform, and we found it was not as popular in Europe, barring Italy and Spain, as German, Slovakian, Slovenian, Danish, and Dutch had very low language counts. However, we did find that Hindi, Bengali, Afrikaans, and Indonesian had high counts, indicating that PixStory may be popular among a demographic in less developed countries.



Hire Article Bar Chart for Languages

Similarity with Image Space

Furthermore, when we conducted a similarity analysis using Image Space, we found that Image Space was able to find similar images, when the images were more simple, compared to when the images had a lot more objects. For example, when looking for similarities for an image showing an athlete kicking a soccer ball, we were able to find similar images and even an exact match. However, when running Image Space on an image with more objects, for example, Valdemir Putin sitting at a desk, we did not find greatly similar results. While we did find the exact match, the other results were not as closely related to our original image as they were with the soccer image. This suggests Image Space's similarity algorithm may have limitations in detecting similarities between more complex images, and further development may be needed to improve its accuracy in such cases. It is also important to note that the extent to which Image Space can help find similarity between images may depend on the level of complexity of the images and the algorithm used to analyze them. (Question 3)

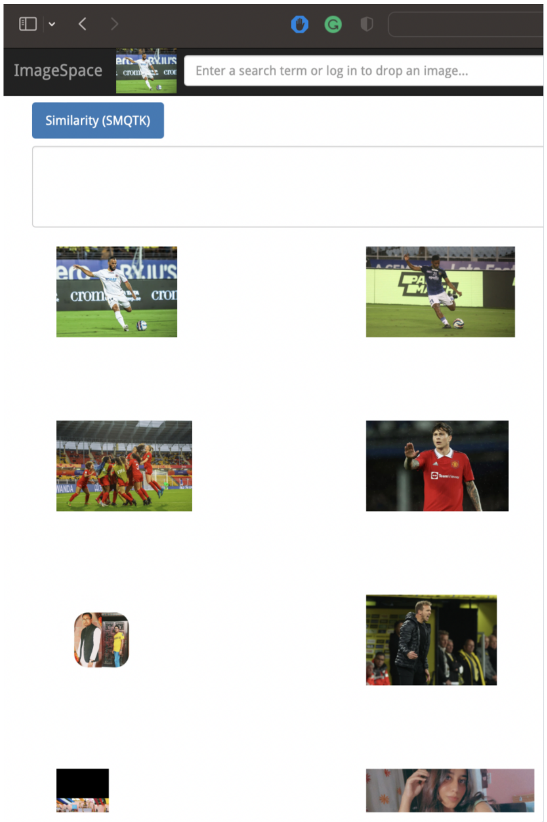


Image Space Similarity for Soccer Player

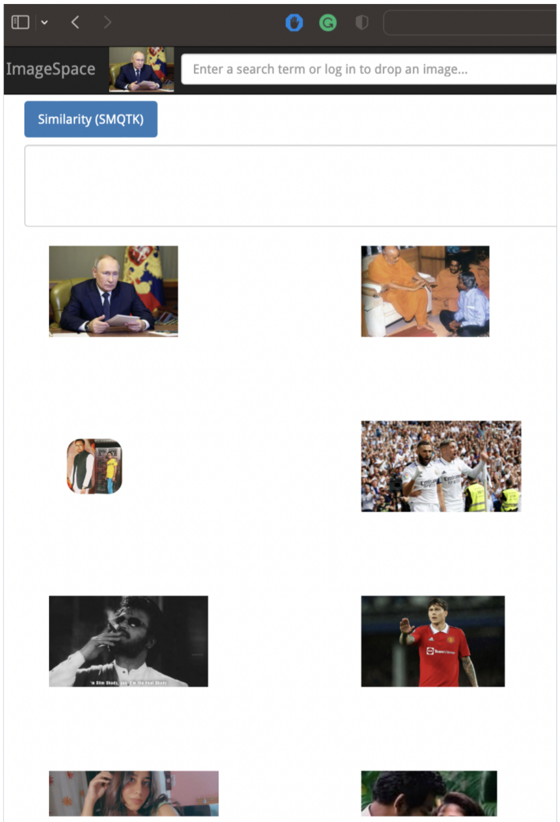
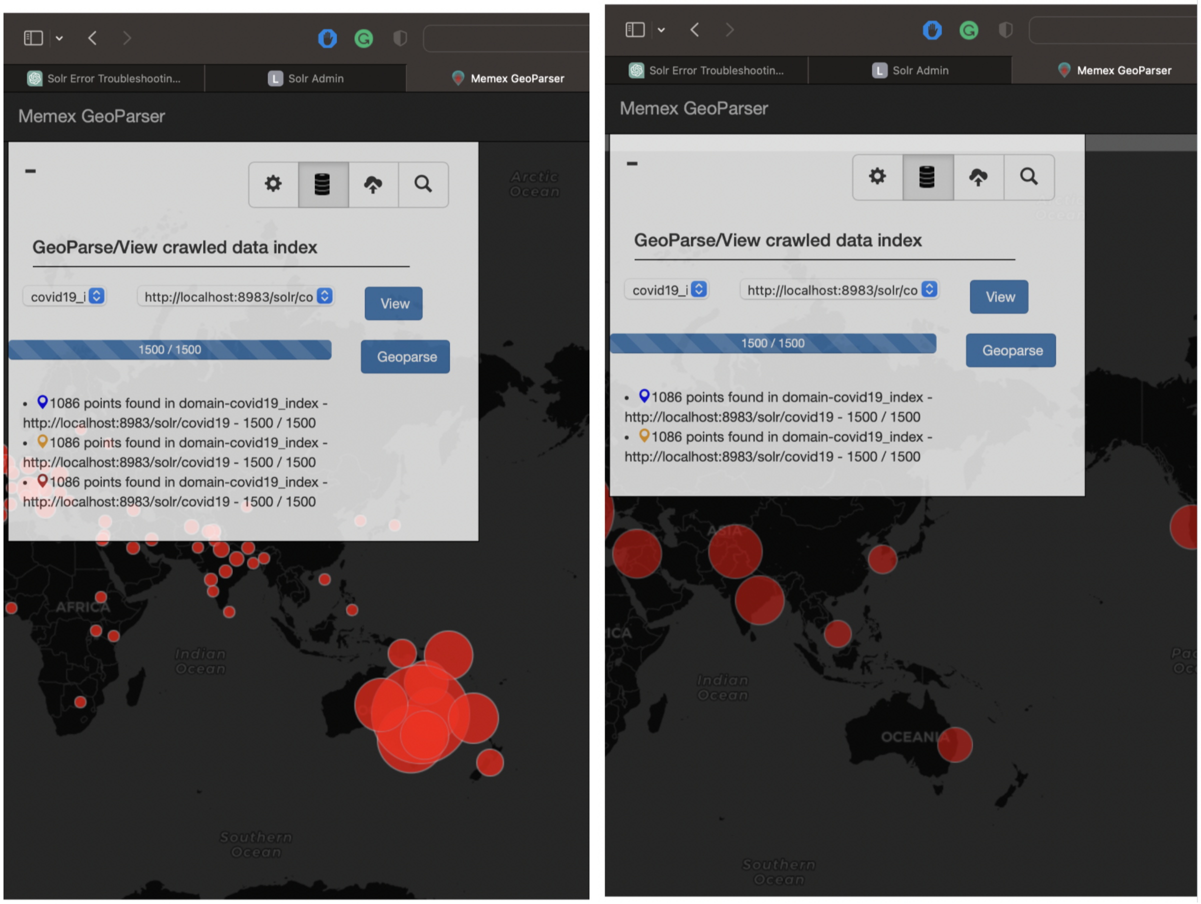


Image Space Similarity for Putin

Locations

As a last step, we implemented a GeoParser to pull the locations of all PixStory users to see if we noticed any differences. When using the Geoparser application, the location data was similar to what was parsed from assignments 1 and 2. This makes sense because the underlying Geoparser application uses the same lucene gazetteer to extract location data. However, in this case as we are using the full Geoparser application, we can easily visualize the concentration and frequency of the exact locations of narrative from the Pixstory social media dataset. Using the Geoparser application, we can easily search specific locations and see where each Narrative is associated with location. (Question 4)

*Note: When referring to the screenshots, although he domain name is `covid19_index` in the screenshots, rest assured that it is actually a random sample of 1500 narratives from the main 95k row Pixstory Social Media dataset.



- 1086 points found in domain-covid19_index - http://localhost:8983/solr/covid19 - 1500 / 1500
- 1086 points found in domain-covid19_index - http://localhost:8983/solr/covid19 - 1500 / 1500
- 1086 points found in domain-covid19_index - http://localhost:8983/solr/covid19 - 1500 / 1500

- 1086 points found in domain-covid19_index - http://localhost:8983/solr/covid19 - 1500 / 1500
- 1086 points found in domain-covid19_index - http://localhost:8983/solr/covid19 - 1500 / 1500

Geo Parser Locations

Conclusion

Overall, aside from the average age of PixStory users, a lot of the insights we were able to extract from our data for Assignment 3, are not what we had previously seen. The main findings from our final visualizations were:

The most commonly used words in the PixStory dataset are related to world news and do not indicate a toxic environment.

- Some states in the US have higher average toxicity scores than others, which can be further analyzed to identify the topics that draw the most toxicity.
- The PixStory platform is being used by a diverse and global audience interested in topics such as sports, entertainment, politics, health, food, environment, history, education, science, and inequality.
    - Interest in real-world issues like inequality suggests that PixStory is a progressive platform that promotes positive discussions.
- Health was the most talked-about topic in 2020 during the height of the pandemic.

Final Thoughts on Technologies

The installation and setup of Image Space was difficult, especially when troubleshooting the installation process of a Macbook M1 with a arm64 chip instead of the recommended amd64/v8 chip. We believe that the Image Space repository should include more detailed instructions on troubleshooting the installation of Image Space and a passable index that describes in detail what certain errors mean. However, once the program was working, it was incredibly easy to navigate the application and find similar images.(Question 5)