Bigfoot Assignment 3 Report: Team 9
DSCI 550: Spring 2024

Creating Data Insights

In this assignment, we explored creating data insights for our bigfoot dataset through visualization, similarity analysis in Image Space, and Geographic Analysis and Visualization with the MEMEX GeoParser. By doing so, we were able to build on our work throughout this semester, create data insights, and further learn about our bigfoot sightings dataset.

Visualizations

We made five interactive visualizations using D3.js to show relationships and features we have discovered in our bigfoot dataset throughout this semester.

Line Chart

In the first assignment, one of our group's conclusions was that bigfoot hotspots seemed to overall have milder weather conditions and higher national park visitation counts. In order to explore the relationship between the nearest national park visitation count and bigfoot hotspots versus non-hotspots further, we decided to create an interactive line chart where there is one line for all hotspot sightings and one line for all non-hotspot sightings. The year values are on the x-axis and we have the national park visitation count on the y-axis which corresponds to the nearest national park to each sighting location and is provided by year. By default, the line chart displays the median value for each group by year; however, it also enables viewing the mean, minimum, and maximum values. Interestingly, the relationship between the nearest national park visitation count and hotspots is less clear from this visualization. This may be due to a difference in analysis approach from the first assignment. That being said, by looking at the median values, there is a higher visitation count for 2017 and 2018 for hotspots than non-hotspots. In 2001, they had the same median value. In other years, the lines are relatively closely aligned which is also the case for the mean values. Overall, the mean for non-hotspots is higher for all years; however, they are very closely aligned in 1990, 1994, 2015 and 2018. The maximum values are significantly higher for non-hotspots across all years than hotspots; however, for minimum values, it is notable that the hotspot sightings visitation count is significantly higher for all years than for non-hotspots.

Particularly for the median and mean values, if we compare the graph with the most popular sighting years in the dataset, there may be more to learn about this relationship. For comparison, the most popular year was 2000 with 220 sightings. In 2001, there were 181 sightings, in 2015 there were 91 sightings, in 2018 there were 56 sightings, in 2017 there were 66 sightings, in 1990 there were 59 sightings, and in 1994 there were 78 sightings. This indicates that the years that the median and mean values were high for hotspots were all popular years for sightings which may suggest a relationship. That being said, limitations to the dataset in determining this relationship, such as the granularity for visitation counts being limited to years rather than seasons and years, as well as the nearest national park not being in the exact county location of the sighting, may still prevent us from concluding anything about this relationship. A statistical test, such as a t-test would be a potential good future step to enable stronger conclusions regarding the relationship between hotspots and visitation counts.

Bubble Map

We selected a bubble map because we thought it was the most direct, simple, and intuitive way to show how abundant the “Number of sightings” is per county. This was one of the first visualizations we thought about since it answers one of the most basic questions we had since the start of the project: “In which type of locations are sightings more common?”. Features we added such as from the Census data, state mammal, and national park data all made us think of this question.

As for the results, we can see that the coast has by far the greatest number of sightings, especially:

California and Washington for the West area, Michigan and Ohio for the North-East area, and Florida for the South-Eastern area. This shows a clear correlation between population density and number of sightings, with Texas being the only prominent outlier (since New York still has a moderate amount). Sadly, the number of 'state mammals' is quite inconsistent in these areas so, contrary to what we previously believed, it does not seem to have any direct correlation with the number of sightings. A second important factor we found was the environment, more specifically we found out that high density sighting areas are found mainly in forest areas, which would explain the low number in Texas.

To summarize, highly populated areas with a moderate to high amount of forest area, are the most important conditions for a sighting to occur.

Bubble Chart

Another one of our five visualizations is a bubble chart, which indicates the frequency of different entities in "Detected Objects" that Tika Image Dockers detected in our images generated through Diffusers in Assignment 2. The chart has a positive correlation between the size of the bubbles and the frequency (count) of the corresponding entities. One of the reasons for choosing it was that the bubble chart can demonstrate explicitly the frequency of different entities detected in our AI generated images which can help us learn both about entities in the sighting descriptions and more apparently Tika entity detection relating to bigfoot. The size and the position of the bubbles can turn the vague data into visible shapes, which helps visually inform users of distribution and distinction. Furthermore, the bubble chart is categorized into five groups: "Nature", "Creature", "Vehicle", "Tool", and "Other", enabling the analysis into a specific type of objects (entities).

Enlightened by the previous analysis in Assignment 1 & 2, we are eager to know what category of objects would be prioritized in AI auto detection in terms of bigfoot sightings. According to the bubble chart we created, apparently "Nature" is the most popular category detected by Tika Image Dockers in our bigfoot sighting images . Among all the detected objects, the top five entities with the most counts are: Lakeside (735); Valley (631); Mountain bike (546); Mountain tent (405); Alp (396), and three of them (lakeside, valley, alp) belong to the "Nature" group. Within some specific categories, the most frequently appearing entities are related to mountain activities, such as mountain tent (count: 405; category: Tool) and mountain bike (count: 546; category: Vehicle). Several insights can be drawn. Firstly, reporters tend to highlight nature elements when describing the bigfoot sightings. Secondly, the Diffusers Stable Diffusion model generated a large quantity of images relating to nature, which may both reflect the sighting description as well as the model's prioritization of nature-related tokens. Thirdly, AI auto detection seems to be more sensitive to nature entities when processing images. Finally, a relatively high proportion of bigfoot sightings tend to happen during human outdoor activities, such as mountain sports, hiking, camping, and rock climbing.

Circle Packing

The Circle Packing visualization is a nested arrangement of circles such that they don't overlap and are mutually tangent to each other. This visualization was chosen simply because it enables us to understand the relationship between more than two types of related categorical data. The bigfoot report contains various information about the location such as county, state, and seasons for sightings. It was crucial to identify which counties and states have the highest sightings during a particular season. Subsequently, the visualization contains 3 nested circles. The outermost circles represent the various seasons, along with an additional circle for 'Unknown' which represents reports with missing data for seasons. The circle inside seasons represent various states where sightings have occurred, and the innermost circle represents the counties in a state. The size of all circles depends on the total number of sightings reported in that county.

The visualization denotes that summer and fall have a very similar number of sightings. However, winter and spring have relatively fewer sightings. This could simply be because people like visiting national parks and getting outside during summer and fall when it is warm instead of winter and spring, when it could snow. Therefore a smaller number of sightings might indicate reduction in tourism rather than bigfoot hibernating during winter. There was a similar trend seen in all seasons where Illinois, Washington, Ohio, and Florida have the highest number of sightings during all seasons. A similar trend was seen for counties where counties such as 'Pierce' and 'Jefferson' in Washington have the highest number of sightings across all seasons.

WordCloud

The World Coud shows the most common words in description on the sightings and their counts. This visualization is chosen because it provides a visualization of the textual data, making it easy to identify key words and themes just by analyzing the size and the frequency of the words in the cloud. Those words and themes can be very useful for further analysis, prompting deeper exploration into specific words or topics of interest.

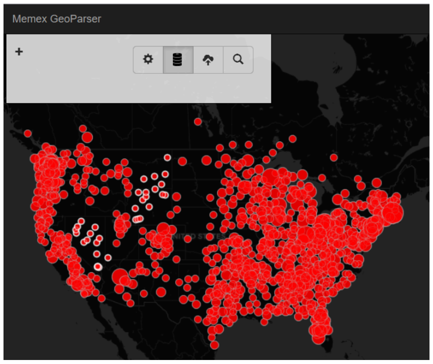
This particular word cloud analyzes a subset of textual data - 'observed', 'also noticed', 'other stories', 'environment', 'other witnesses', 'follow-up report', 'time and conditions' - related to description of the sightings themselves. The most common words include 'animal' (5608), 'creature' (6197), night (4972), and woods (5375). Several observations can be made from this analysis. For instance, focusing on factors such as time of day, location, and the type of animal/creature sighted could yield valuable insights. It's noteworthy that the term 'bigfoot' is not as frequently used by witnesses, who instead employ more ambiguous terminology such as animal and creature. Furthermore, it suggests that conducting entity analysis on aspects like 'location' and 'time' would be beneficial. In addition to that, there are common words such as 'deer', 'bear', and 'monkey', which could suggest possible animals which were confused with bigfoot.

Similarity analysis with Image Space

We conducted a Similarity analysis using Image Space. It took approximately 1.5 hours for the SMQTK-service-server to process all of our 5467 Bigfoot images. Image Space allows for the rapid identification of images with high similarity by simply clicking the magnifying glass icon next to an image. This indicates a major shift from the visual inspection of image trends performed in Assignment2. For example, during Assignment 2, we observed that our AI-generated Bigfoot images often resemble gorillas and orangutans, while many other images featured bears or depicted scenes with woods, grasslands, and railways. However, through searches in Image Space during Assignment 3, we confirmed that bigfoots in the images have longer arms than typical gorillas and appear more human-like. Furthermore, we noticed a significant number of images containing other animals such as deer, cows, and dogs, as well as images combining tents with forests and depicting pairs of humans. Clicking on images that included text also allowed the extraction of other images containing text. Image Space uses OCR to extract text from these images, enabling content-based searches. However, as the text strings in AI-generated images are random, accurately targeting specific images using these strings may prove to be difficult.

Geographic Analysis and Visualization with MEMEX GeoParser

We visualized the location information for each bigfoot sighting report using MEMEX GeoParser. MEMEX GeoParser differs from the Google Map API we used in Assignment 1 in that it is capable of interpreting complex and lengthy textual input to autonomously detect and plot potential location-related entities on a map. However, this convenience can reduce accuracy, as evidenced by numerous locations erroneously plotted outside the United States, where all reports should be located.



In Assignment 1, we identified significant concentrations of reports in Washington and California with data using Google Map API, possibly related to the abundance of national parks in the West Coast region. In contrast, MEMEX GeoParser exposed us to a wider variety of location data that, while less precise, provided a different perspective on the distribution of sightings. A particular feature of the MEMEX visualization was its clustering capability, which dynamically grouped overlapping data points as the map scale changed and this feature helped to capture the general distribution of sightings. However, due to a lack of clarity in its clustering criteria, the potential new distribution suggested by the visualization could not be definitively categorized as new correlations. For example, a large circle representing a cluster appears around the Mountain States when you zoom out of the map, but this cluster turns out to be only a few sightings when you zoom in. This means that the distance threshold for clustering may be too high. Therefore, we need further analysis, such as state-based aggregation, to argue for the high-level patterns presented in MEMEX Geoparser.

Reflections on Visualization Tools

Image Space

In practice, Image Space has proven to be a fast and effective tool for image retrieval by metadata and similarity. However, we see several points where it can be improved. First, there are multiple README files scattered in the Image Space repository on GitHub, making the installation processes unclear. Also, the instructions given in the class work well for the first try of installation, but do not work for restarting docker containers or adding images. To address these issues, our README includes instructions for a reproducible installation procedure, rebooting, and adding images. In addition, we also found some improvements in the user interface. In the display, the images are small and the number of images are limited, making it stressful to browse the search results. In addition, for searches using similarity, it might be helpful to display a similarity score for each image in the search results to help users interpret the results.

MEMEX GeoParser

The pros and cons of MEMEX GeoParser are as described above. That is, although the location accuracy is relatively low and the clustering criteria is unclear, it is a useful tool with high flexibility to handle complex natural language and with the ability to perform pretty visualization. As for installation and restarting, we did not experience any particular problems.