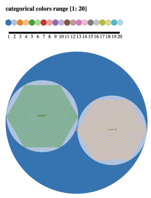
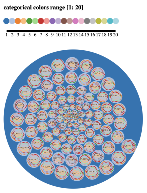
DSCI 550: Assignment 1 Report (Team 9)

1) What we notice about the dataset while completing the tasks:

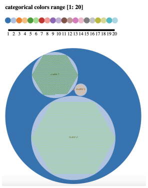
We noticed that the Bigfoot dataset contained sightings in 49 out of the 50 US states (excluding NaN values). There has not been a single Sasquatch sighting in Hawaii, this could be due to its remote location in the Pacific Ocean. Perhaps Sasquatch is a terrestrial creature that is unable to travel across the ocean. 90.05% of reports have listed the season for sighting. Out of those, Summer had the most sightings (37.87%) followed by Fall (30.19%), Spring (16.79%), and Winter (15.12%). Data cleaning and preparation involved handling a diverse set of challenges, such as standardizing the "Year" column, which included a mix of simple and complex entries (like '2012' and 'mid-80s', respectively) and missing values (nan). Complex year entries were converted into an average year value by creating a list of corresponding years for each and averaging them. For missing year data, an attempt was made to impute values using the Submitted Date, although this approach was limited by the potential discrepancy between the sighting and submission dates. Geographical matching to the nearest national park required fetching location coordinates via the Google Maps API, with specificity enhanced by state, county, and route details. The task of extracting the total number of witnesses from the 'Other Witnesses' field involved developing a custom function to parse a variety of descriptive entries, acknowledging the difficulty in accurately determining witness counts in every case. Merging weather data with the original database posed another layer of complexity, requiring a method to aggregate frequently updated weather information with the less frequently updated original data. This was achieved by using geographic (state and county) and temporal (year and month) keys to aggregate climate data by region and season, averaging numerical weather data, selecting the most frequent occurrences for categorical data, and choosing representative values for all other information.



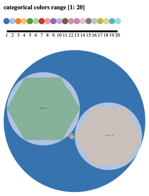
1



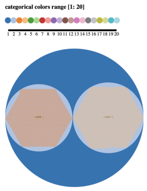
2



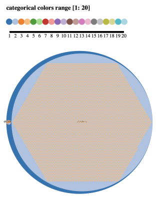
3



4



5



6

In the example above, although the clusters were generated based on the same chunks of data, the number and structure of clusters significantly differ for Jaccard (1), Edit distance (2), and Cosine (3). The main reason for that is that they are simply calculated differently based on not the same (as it was concluded from the exploration below) features.

To figure out features based on which columns have the most impact on the cluster generation, the following technique was implemented: columns with specific data from the JSON chunks were replaced by just nulls ("") and the results were compared with the visualizations based on the same chunks of data, but without deletion of those columns. According to the result, Jaccard clusters are likely to depend on numerical data, as when columns with other types of data (such as text) were deleted, it had almost no impact on them. For instance the picture 4 above is how Jaccard visualization looks after deletion of text columns from the data (almost identical to the one above, where no columns were deleted). However, after deletion of columns containing numerical data specifically, the visualization changed to the picture 5. In addition, the Cosine is likely to be based on features related to text, as after deletion of text columns the visualization changed to 6 (only one huge cluster and 1 small one, so almost all the data in one cluster).

2) Clearly explain which datasets were joined with the BFRO data and how were the features extracted from each dataset.

Dataset 1: People are usually skeptical about believing in Big Foot because the person reporting the incident could be hallucinating under the influence of alcohol, opioids, or other drugs. Therefore, we chose to merge the dataset on Alcohol consumption in the USA from 1977 to 2021, published by the National Institute on Alcohol Abuse and Alcoholism (NIAAA). This dataset contains consumption of beer, wine, spirits, and all alcoholic beverages. Since the person claiming to have seen the BigFoot could have drunk any of these beverages, we chose to merge the data for all beverages. Additionally, the data also included Population, Gallons of ethanol per capita, and Decile for per capita consumption for two age groups (14+ and 21+). Every state had a different legal drinking age before the Federal Uniform Drinking Age Act was passed in 1984, also teenagers have a tendency to consume alcohol illegally using fake IDs. Due to this reason, we chose to include data for the population of 14 years & older. Since the data was stored in a text file, it was relatively easy to extract, clean and merge the data using python code in a jupyter notebook. The data was merged according to matching state location and year in both datasets.

Dataset 2): One of the most obvious reasons for why a sighting can occur more in one location or time than another, is its population density. Because of this we decided to extract the Resident Population, Population Density, and Density Rank features from a census dataset (application/pdf). First we used the "tika" to parse the content of the pdf so it was readable, but since the output provided wasn't correctly structured we had to do some cleaning to the data, mainly removing anything unrelated to the data columns and making compound named states (E.g. New York) into single words. Afterwards we just created a dataframe with the three features, state, and years (1905-2021) and combined it with the bigfoot dataset using the 'State' and 'Fixed Year' as the key to combine them. Year's results were organized by their closest decade (E.g. 2000's data: 1996-2005).

Dataset 3): The image dataset consists of 73 images of state mammals. A state mammal is the official mammal of a U.S. state as designated by a state's legislature. Along with the images, some side information is provided, such as the name of the mammals, the corresponding states of the mammals, and the years when it was officially announced as the state mammal. So we collected the following three types of data: State; State Mammal; Year, and manually made it into three columns, with each row representing information of one state only. To make the dataset more related to BFRO data, we manually created a new column called 'Foot Size', indicating the foot size of the corresponding state mammals. The 'Foot Size' data has three values: Big; Small; None. 'Big' and 'Small' refer to a relatively big and small foot size of the state mammals, respectively. 'None' refers to an unlikeliness that the state mammal could ever leave a footprint on the ground, predominantly because they are marine mammals. The method of determining the 'Foot Size' is by manually searching all the mammals on Google and categorizing them according to their authorized description. Afterwards, we developed a Python script to join the state mammal image dataset into the BFRO dataset, matching the 'State' data of each sighting report (See Task5PNG.py). So in the reports_task5.tsv file, following after the 'State' column are 'State Mammal' 1/2/3, 'YEAR PUBLISHED' 1/2/3, and 'SIZE' 1/2/3.

3) What questions did your new joined datasets allow you to answer about the BFRO data and its sightings and additional features previously unanswered?

In our study on national parks, a slight positive relationship was found between Bigfoot sightings and the number of park visitors, implying more sightings occur in parks with more visitors, likely due to increased outdoor activity. Similarly, analysis of weather data revealed that regions with more Bigfoot sightings tend to have slightly higher rainfall but generally experience milder weather conditions, making them ideal for hiking. This suggests that areas popular for outdoor activities, with good hiking conditions, report more sightings. By analyzing the census features, we found a weak correlation between the Population Density and the number of reports, which showed that most sightings were not done in states with the biggest density, they were rather centered in states in the upper middle size. To be more specific, the Census Rank of the top ten states (by number of sightings) would be around 19.8. Although the reason why sightings are less frequent in the less dense states could be easily explained by saying it's a population/sighting rate effect, we can't say the same about the states with bigger density. For this situation, the most logical explanation would be that since the residential area of the state is bigger, wild environments (and activities related to it) where bigfoot could be found, are less abundant. By analyzing the image dataset and its related data such as ‘Year Published’ and ‘Size’, the data showed that the earlier the state mammals are published, the more likely a bigfoot sighting report would occur in that corresponding state. Furthermore, the more state mammals with big foot size the state has, the more sightings were reported within that state. This may be related to some cultural reasons. A state mammal was widely considered a symbol or representation of the state. The announcement of the state mammal with a big breed size may raise citizens’ awareness of a potential wildlife existence, thus increasing the possibility of a bigfoot report, though they were not necessarily all from an unknown creature.

4) What did the additional datasets suggest about “unintended consequences” related to sightings of BigFoot?

The correlation between increased visits to national parks and Bigfoot sightings, as well as the allure of sighting locations for outdoor activities and favorable weather, and the link to areas with higher alcohol consumption, may not be causal. Accurate statistical analysis is necessary to avoid incorrect conclusions. Regarding the image dataset, mammal images might be misleading regarding the size of footprints, so analysts must verify to prevent errors.

5) What similarity metrics produced more (in your opinion) accurate groupings? Why?

According to the exploration above, different similarity metrics are primarily based on different columns/data, so it is not very rational to compare them on accuracy, as they primarily display different features. However, we may claim, e.g., that Jaccard metrics is likely to be efficient for the numerical data, as it seems likely to be primarily based on numerical features.

7) Does the time of day of the sighting matter?

Most of the sightings have occurred between 10 pm to 3 am. A few have occurred during dawn or dusk (7pm or 5 am). Very few sightings have occurred during the day, especially afternoon (12pm to 4pm). Since, visibility is low during nights and most of the sightings occurred in secluded areas such as national parks or forests, the time of day likely played an important role. It could either suggest that BigFoot is a nocturnal mammal which only ventures outside during the night or people are simply mistaking some animal during the night to be a BigFoot.

8) Are specific locations more likely to be “BigFoot Hotspots”?

Washington State and California states account for about 21% of the total number of witnesses. And both of these states are located on the West Coast. When comparing datasets focused on Bigfoot Hotspots with those that are not, by paying particular attention to the number of visitors to national parks and meteorological data, it appears that Bigfoot Hotspots tend to have mild weather and a higher number of national park visitors. This suggests that Bigfoot Hotspots may be regions favored for outdoor activities.

9) Are specific keywords apparent in the reports that are less sourced? For example, do class C reports tend to have familiar keywords?
Most common keywords were similar in all classes (E.g. 'heard', 'miles', 'area', 'creature'...) but Class C seemed to lack some important keywords that made it look like the sighting was poorly verified, among the missing keywords you could find: investigation', 'investigator' and 'follow-up'. Additionally, most Class C reports had keywords like: 'family', 'brother', 'dad', 'friends', 'deer', 'dog', 'bear' and 'animal'. Probably due to the fact that many of them are second-hand sightings and the high probability of having confused a specific animal with bigfoot, which explains why class C is the most unreliable.

10) Is there a set of frequently co-occurring that define a particular sighting or class of sighting?
Having a 'Follow-Up' report seems the most important event for there to be a Class A or B sighting. If we add to that weather 'Severity', we have that a report is more than two times more probable of being Class B than Class A if the severity is 'severe' or 'moderate'. But what we found the most intriguing was that the keyword 'heard' is also more than two times more frequent in Class B (likely only a noise report).

11) What insights do the "indirect" features you extracted tell us about the data?
When working with noisy numerical data, such as Witness Count, it's hard to see patterns because the noise hides them. Using indirect features like Bigfoot Hotspots, for data aggregation can lessen this noise effect, allowing for clearer group comparisons. Through this, we could identify that Hotspots, marked by the indirect feature of Bigfoot Hotspot, typically feature mild weather and attract outdoor enthusiasts.

12) What clusters of sightings made the most sense? Why?
The clusters for the Jaccard distance made the most sense since it produced clusters which contained similar features based on specifically categorical data.

13) Also include your thoughts about Apache Tika – what was easy about using it? What wasn't?
Easy:

1. Extraction of text and metadata from diverse types of files, including images and videos, which can be useful for transferring different file types into a consistent format for further analysis.
2. Compatibility with multiple data types, including text, image, video, audio, application, etc, allows us to expand our choices on MIME types when adding new features from other datasets.
3. Apache Tika is constantly being updated, getting accustomed to different programming languages, especially Python, which is relatively the easiest one among all the other languages. As a result, Apache Tika can meet the needs of programmers of different levels.

Difficult:

1. As a sophisticated toolkit, Apache Tika may introduce unnecessary bias and complexity when it comes to some simple tasks.
2. Tika Similarity requires working with json files a lot, and they are not easy e.g. to divide into chunks or delete columns from, which required implementing separate scripts
3. It might be difficult for the beginners to understand how Apache Tika works and due to the lack of a user interface or an application.