Homework: Analysis of the BigFoot Research Organization (BFRO) Sasquatch Reporting Database
Due: Friday, March 8, 2024 12pm PT

---

1. Overview

Figure 1: The BFRO dataset (BigFoot Research Organization): mysterious reports of BigFoot/Sasquatch sightings in and around the United States over the last century and associated dataset of 5,467 rows x 28 columns and the most frequently occurring topics in the dataset show as a word cloud (bottom middle). More can be found at http://bfro.net.

In this assignment we will explore several of the topics discussed in the early portion of class – Big Data – MIME types and their taxonomy – Data Similarity – and so forth. To do this, we will leverage the dataset highlighted in Figure 1 – a set of 5,467 reported sightings of "Big Foot" or sasquatches, which are mysterious human/ape like creatures, dating in part from January 1959 to December 2021. The posts contain several features which are highlighted below:

- Report Type (either a 'Report' as in an actual sighting as reported by an eyewitness or otherwise, or 'Media article' in which the report is highlighted by an accompanying media article.)
- Id (a numerical identifier for the report)
- Class (there are three types of classes of BigFoot reports, 'Class A' is the highest quality, and has been validated potentially by secondary sources and other witnesses or phenomena. 'Class B' is the next most reputable type; and 'Class C' is potentially hearsay or secondary unverified sourced report. You can read more here: https://www.bfro.net/gdb/classify.asp)
- Submitted Date (the date that the report was submitted)
- Headline (a headline / short text description of the report)
- Year (the year that the report corresponds to)
- Season (Fall, Winter, Summer or Spring)

- Month (the month that the report corresponds to)
- State (the state in the United States where the report corresponds to)
- County (the local county in which the sighting occurred)
- Location Details (any extra information or details about the location)
- Nearest Town (any information about the closest town to the sighting)
- Nearest Road (any information about the nearest road to the sighting)
- Observed (how the BigFoot was observed, text field)
- Also Noticed (any other features noticed during the sightings, could be regarding the location, or the BigFoot itself)
- Other Witnesses (if there were other witnesses, and what did they say. You can expect this field to be filled most of the time for Class A or Class B sightings)
- Other stories (any other stories of sightings that could be related to this sighting)
- Time and conditions (associated time and lighting and conditions that may affect the sighting)
- Environment (any environmental conditions that could have affected the sighting)
- Follow up (whether a follow up was performed by any BFRO investigators)
- Follow up report (if any BFRO or other follow up report was made)
- Date (date of the sighting)
- Author (report author)
- Media Source (if any, names of the associated media sources)
- Source Url (if there is a link to a URL for associated media reports)
- Media Issue (the volume number of issue number of associated media)
- Observed.1 (text information describing the BigFoot creature)
- A&G References (information about the surrounding area of sighting)

The BFRO dataset is a rich dataset with high variation in its features and properties. For example, as can be seen from Fig 2., most of the sightings for BigFoot occur during the summer months, followed by the Fall and Winter. Additionally, nearly 4 times the sightings are considered class B and class A (total) which are the high quality, highly sourced and verified sightings, compared with the categories of Class C (heresay) and 'Unknown' in the dataset. Given BFRO's stated goals of providing high quality BigFoot research reports for the community to study, and also given public lexicon and domain knowledge about BigFoot sightings, we can convince ourselves that both the quality of the sightings as well as the timing in terms of months where BigFoots are most seen somewhat make sense, even from a cursory analysis of the data.

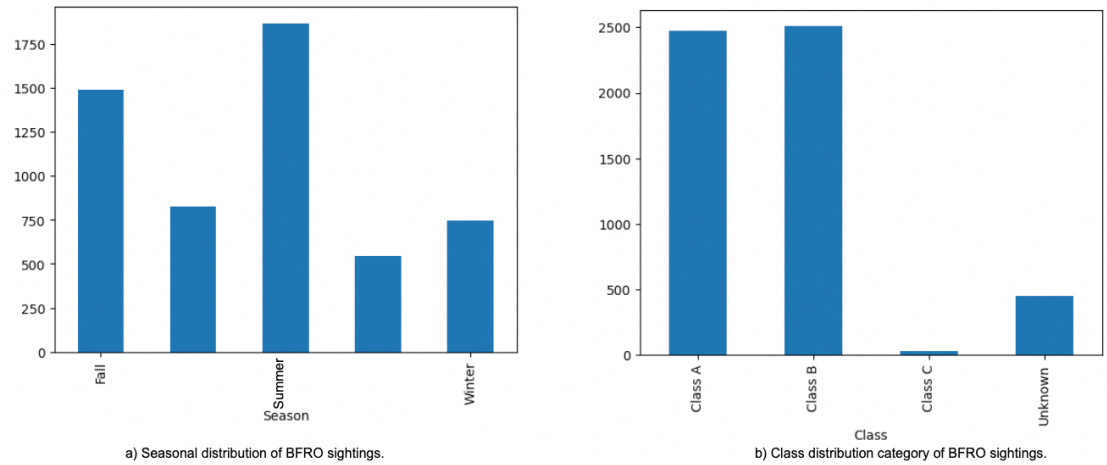


Figure 2. Seasonal and class distribution based on sightings in the BFRO dataset.

One of the other important elements of the BigFoot dataset is all the ancillary report information including location, time of day, and whether other witnesses saw the sighting. These should give you confidence in your review of the sightings, but also should be something that may bring some questions to mind. For example, in Class A sightings that occurred in the evening portion of the day, combined with low visibility, what other source information could be used to verify if the BigFoot was a creature not known to us rather than a traditional sighting of a bear? How could someone tell this in low lighting situations, in the evening?

One of the important fields is "Other Witnesses" however you will see that the text is written sometimes as numbers ("4") and other times by mentioning who was there, or how many ("Myself and daughter. We were listening to music, talking and just enjoying the sights and view while driving."). This field could help, along with other portions of the sighting to compare and build confidence in the sighting. For example, we can use this information to explore: are sightings with multiple witnesses more likely to be classified as 'Class A' sightings even at night and in low lighting settings? Are single witness sightings more likely to have a follow up performed? You could use the number-parser Python tool to extract out the numerical value of number of witnesses: https://github.com/scrapinghub/number-parser as an example.

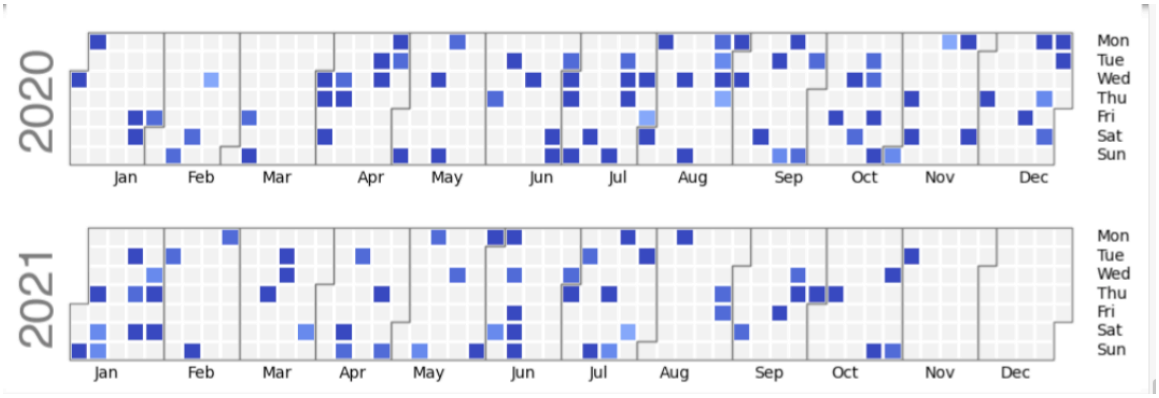
Additional insight can be gleaned from exploring the temporal properties of sighting. Are there particular days of the week with more BigFoot sightings? What happens if you slice it down by Class A (high quality sourced) sightings, for example? Examine figure 3 closely to determine if you can see any patterns.

2. Objective

Exploring the BFRO dataset, some things may jump out at you. First off, regardless of low- or high-quality sightings, Saturdays seem to be a strong date for observations. What might be the reason for that? Diving down into the data, one of the things that may jump

out is that the weekend is usually when humans are out backpacking and hiking in the mountains and more likely to see a BigFoot. Looking at national park data, would it be possible to validate this by comparing the number of visitors of national parks during the weekend, compared with other days of the week, looking for patterns and trends? For example you could use the publicly available data here: https://www.nps.gov/aboutus/visitation-numbers.htm and compare. Additionally, could the fields for nearest town, time and conditions and environment help to convince yourself of this intuition? Of the non-weekend days, just from looking at the recent years' temporal properties, middle of the week tends to be a strong sighting data as well. What could the reason for this be? Could you compare and look at data for the sighting location and look at visibility data for that region and date/time? Is it possible that even during the Summer and Fall months, that middle of the week had greater visibility because the days are a bit longer and there is greater chance to see a BigFoot? There are weather datasets such as US Weather Events (2016-2022) on Kaggle (https://www.kaggle.com/datasets/sobhanmoosavi/us-weather-events) that you could use to compare the visibility in the region and area of the sighting and cross reference that with the date/time and other conditions.

A) Subset: 2020 - 2021 of Day of week distribution of BFRO sightings: All Classes Sightings (A, B, C and Unknown)



B) Subset: 2020 - 2021 of Day of week distribution of BFRO sightings: Only Class A (high quality, highly sourced) sightings shown.

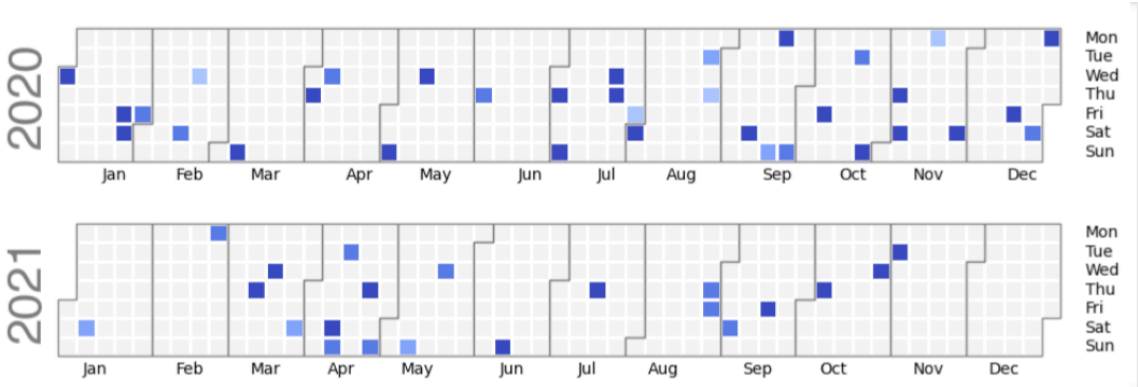


Figure 3. Temporal breakdown of BFRO subset data (2020-2021) – all classes of sightings shown in A) and only high quality (Class A) sightings shown in B).

What ways could you explore this, by looking beyond this rich dataset and by applying lessons learned from class thus far where we have been studying the 5 V's, MIME types of associated datasets, and large datasets and their characteristics? Could you join for example, the US National Park Services (NPS) visitation numbers by year dataset? What trends emerge? Do the BFRO sightings correspond to visitation numbers in the National Parks? In high visitation years do we get more sightings? Additionally could you join the US Weather Events dataset at least for the overlapping years 2016-2021 and then compare the weather events to the count of additional witnesses and see if there are any correlations between weather events affecting the class type of the sightings, or what the witnesses saw? Can you deduce location as a factor of number of posts? Are certain sighting locations BigFoot "hotspots"? What other features can you think of?

You will choose at least three additional publicly accessible datasets along these lines to join the BFRO data to, and you must add at least three new features per dataset that you join. The datasets you select may not all belong to the same MIME top level type – that is – you must pick a different MIME top level type for each of the three datasets you are joining to this BFRO dataset.

Once the data is joined properly, you will explore the combined dataset using Apache Tika and an associated Python library called Tika-Similarity. Using Tika Similarity, you can evaluate data *similarity* (as discussed during the Deduplication lecture in class; and also during data forensics discussions). Tika similarity will allow you to explore and test different distance metrics (Edit-Distance; Jaccard similarity; Cosine similarity, etc.). And it will give you an idea of how to cluster data, and finally it will let you visualize the differences between different clusters in your new combined dataset. So, you can figure out how similar BFRO sightings are, given their locations, witness counts, geographic proximity to certain towns, including national parks and their visitor stats, along with sighting times, and other features, and explore your new augmented BFRO dataset. For example, you may ask, how many highly sourced class A sightings of BigFoot came when there was low visibility in the Fall season and if the sightings were in a large visitation national park year with at least 3 witnesses?

The assignment specific tasks will be specified in the following section.

3. Tasks

1. Download and install Apache Tika
    a. Chapter 2 in your book covers some of the basics of building the code, and additionally, see https://tika.apache.org/2.9.1/index.html
    b. Install Tika-Python, you can pip install tika to get started.
        i. Read up on Tika Python here: http://github.com/chrismattmann/tika-python
2. Download the BFRO dataset
    a. We will provide you a Dropbox link in Slack for each validated team
    b. Make a copy of the original dataset (because you are going to modify/add to it in this assignment)
3. Create a combined TSV file for your BFRO dataset

    a. Convert the CSV to TSV (here's a simple example of how to do this with Python https://unix.stackexchange.com/questions/359832/converting-csv-to-tsv)
4. Add and expand the dataset with the following features
    a. Add a new feature called "National Park Visitation Count" and join the associated visitation counts data from national parks https://www.nps.gov/aboutus/visitation-numbers.htm – add any overlapping years and park locations with the BFRO dataset.
    b. Add a new feature called "Witness Count" and use https://github.com/scrapinghub/number-parser to add the count. This should be a numerical feature capturing the stated number of witnesses.
    c. Add a new feature called "Multiple Witnesses". Set this to True if the Witness Count is greater than 2, and the sighting is a Class A or Class B sighting. If the sighting is an Unknown sighting, set this to True, and rank the sighting as a Class B sighting. If the sighting is a Class C sighting, set this to True. If Witness Count is less than 2, set this to False.
    d. Add a new feature called "BigFoot Hotspot" and set this to True if the location of the sighting is in the top 10 locations with the most sightings in the dataset. Set this to False, if the location falls outside of the top 10 sighting areas.
    e. Join in the 14 columns from the US Weather Event dataset for the overlapping sightings 2016-2021. https://www.kaggle.com/datasets/sobhanmoosavi/us-weather-events
5. Identify at least three other datasets, each of different top level MIME type (can't all be e.g., text/*)
    a. Check out places including: https://catalog.data.gov/dataset (Data.gov)
    b. For each dataset, develop a Python program to join the data to your new BFRO dataset
        i. For each non text/* dataset, be prepared to describe how you featurized the dataset
    c. Each dataset that you join must contribute at least three features (in addition to the features you are adding described in part 5)
    d. For each feature you add, be prepared to discuss what types of queries it will allow you to answer and also how you computed the feature
6. Download and install Tika-Similarity
    a. Read the documentation
    b. You can find Tika Similarity here (http://github.com/chrismattmann/tika-similarity)
    c. You will also need to install ETLLib, here (http://github.com/chrismattmann/etllib)
    d. Convert the TSV dataset into JSON using ETLLib's tsv2json tool
    e. Compare Jaccard similarity, edit-distance, and cosine similarity using Tika Similarity
        i. Compare and contrast clusters from Jaccard, Cosine Distance, and Edit Similarity – do you see any differences? Why?

    f. How do the resultant clusters generated highlight the features you extracted? Be prepared to identify this in your report.
7. Package your data up by combining all of your new JSONs with additional features into a single TSV (tab separated values) file where the columns represent the features and the rows are the instances of email attack.
8. (**EXTRA CREDIT**) Add some new D3.js visualizations to Tika Similarity
    a. Currently Tika Similarity only supports Dendrogram, Circle Packing, and combinations of those to view clusters, and relative similarities between datasets
    b. Download and install D3.js
        i. Visit http://d3js.org/
        ii. Review Mike Bostock's Visual Gallery Wiki
            1. https://github.com/mbostock/d3/wiki/Tutorials
    c. Consider adding
        i. Feature related visualizations, e.g., time series, bar charts, plots
        ii. Add functionality in a generic way that is not specific to your dataset
        iii. See gallery here: https://github.com/d3/d3/wiki/Gallery
        iv. Contributions will be reviewed as Pull Requests in a first come, first serve basis (check existing PRs and make sure you aren't duplicating what some other group has done)

4. Assignment Setup

4.1 Group Formation

You can work on this assignment in groups sized at **minimum 2, and maximum 6**. You may reuse your existing groups from discussion in class. Please fill out the group details in this https://forms.gle/dAtyJGN7P6DAtMGx6 after class on Thursday, February 8th. Only one form submission per team. If you have any questions, contact Deepanshu via his email address with the subject: DSCI 550: Team Details.

4.2 BFRO dataset

Access to the data is provided by a dropbox link. The dataset itself is approximately 6.1Mb zipped and 16.7 Mb unzipped. You may want to distribute the data between your team-mates since the data is fairly small (for now).

4.3 Downloading and Installing Apache Tika

The quickest and best way to get Apache Tika up and running on your machine is to grab the `tika-app.jar` from: http://tika.apache.org/download.html. You should obtain a jar file called `tika-app-2.9.1.jar`. This jar contains all of the necessary dependencies to get up and running with Tika by calling it your Java program.

Documentation is available on the Apache Tika webpage at http://tika.apache.org/. API documentation can be found at http://tika.apache.org/2.9.1/api.

Since you will be using Tika Python, you will want to read up on the Tika REST API, here: https://cwiki.apache.org/confluence/display/TIKA/TikaServer. The Tika Python library is a robust REST client to the Java-side REST API.

You can also get more information about Tika by checking out the book written by Professor Mattmann called "Tika in Action", available from: http://manning.com/mattmann/.

5. Report

Write a short 4-page report describing your observations, i.e. what you noticed about the dataset as you completed the tasks. What questions did your new joined datasets allow you to answer about the BFRO data and its sightings and additional features previously unanswered? What clusters were revealed? What similarity metrics produced more (in your opinion) accurate groupings? Why? What did the additional datasets suggest about "unintended consequences" related to sightings of BigFoot? You should also clearly explain which datasets you used to join the BFRO data and how you extracted the new features from each dataset.

Thinking more broadly, do you have enough information to answer the following:

1. Are there clusters of sightings with similar features that all have multiple witnesses?
2. Does the time of day of the sighting matter?
3. Are specific locations more likely to be "BigFoot Hotspots"?
4. Are specific keywords apparent in the reports that are less sourced? For example, do class C reports tend to have familiar keywords?
5. Is there a set of frequently co-occurring that define a particular sighting or class of sighting?
6. What insights do the "indirect" features you extracted tell us about the data?
7. What clusters of sightings made the most sense? Why?

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

Note: Report should be written using 11 pt Times New Roman font, single column with single spacing.

6. Submission Guidelines

This assignment is to be submitted ***electronically, by 12pm PT*** on the specified due date, via Gmail dsci550sp24@gmail.com. Use the subject line: DSCI 550: Mattmann: Spring 2024: BIGDATA Homework: Team XX. So if your team was team 15, you would submit an email to dsci550sp24@gmail.com with the subject "DSCI 550: Mattmann: Spring 2024: BIGDATA Homework: Team 15" (no quotes). **Please note only one submission per team**.

- All source code is expected to be commented, to compile, and to run. You should have at least a few Python scripts that you used to join three other datasets, and what you used to extract additional features.
- Use relative paths {not absolute paths} when loading your data files so that we can execute your script/notebook files without changing everything.
- If using a notebook environment, use markdown cells to indicate which tasks/questions you are solving.

- Include your updated dataset TSV. We will provide a Dropbox or Google Drive location for you to upload to {you don't need to attach it inside the zip file}.
- Also prepare a readme.txt containing any notes you'd like to submit.
- If you used external libraries other than Tika Python and Tika Similarity, you should include those jar files in your submission, and include in your readme.txt a detailed explanation of how to use these libraries when compiling and executing your program.
- Save your report as a PDF file (TEAM_XX_BIGDATA.pdf) and include it in your submission.
- Compress all of the above into a single zip archive and name it according to the following filename convention:

  TEAM_XX_DSCI550_HW_BIGDATA.zip

  Use only standard zip format. Do **not** use other formats such as zipx, rar, ace, etc.
- If your homework submission exceeds the Gmail's 25MB limit, upload the zip file to Google drive and share it with dsci550sp24@gmail.com.

When submitting, please organize your code and data file as the directory structure shown:

```
Data
dataset1 {leave it empty for now}
Source Code
script1
notebook
Readme.txt
Requirements.txt
```

Important Note:

- Make sure that you have attached the file when submitting. Failure to do so will be treated as non-submission.
- Successful submission will be indicated in the assignment's submission history. We advise that you check to verify the timestamp, download and double check your zip file for good measure.
- Again, please note, only **one submission per team**. Designate someone to submit.

6.1 Late Assignment Policy

- -10% if submitted within the first 24 hours
- -15% for each additional 24 hours or part thereof