Homework: Analysis of Sightings UFO Data and Unintended Consequences
Due: Friday, March 2, 2018 12pm PT

---

1. Overview

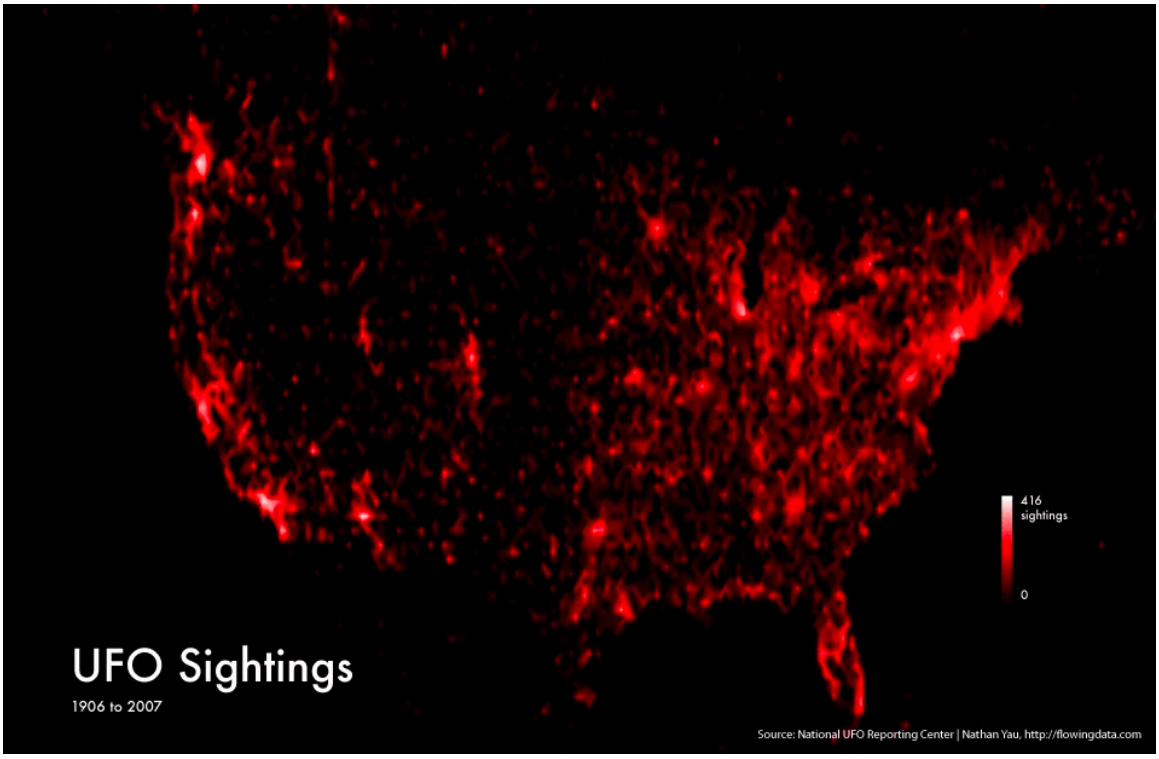

Figure 1: UFO Sightings -
http://flowingdata.com/2011/07/07/where-the-aliens-are-flying-their-ufos/

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 60,000+ UFO ("unidentified flying objects") sightings from the years 1906 to 2007. The data is summarized in a sighting count density heat map shown in Figure 1.

As can be gleaned from the heat map, the data occurs in the continental United States (don't worry – in future assignments, we will leverage data from elsewhere on UFO sightings – including other places in the world). The data is formatted according to a schema which will be provided to you and is in Tab Separated Values (TSV) format (MIME type: text/tab-separated-values). The data includes the dates on which the UFO sighting occurred, and when the observer reported the sighting (may not be the same date – for example someone may have seen a UFO, and been afraid to report it for example – until later deciding to simply do so). There is also data about what type of shape the UFO took on – a triangle; a sphere; an orb, etc. The dataset also includes the duration of the

sighting (5 minutes; an hour, etc.), and finally a free text description by the observer of what happened, and what they believe to have seen.

As you can imagine, there are plenty of things we could do with this data. Some of them fall into to the realm of the paper we discussed from IEEE Computer in 2013 regarding "Big Data's Unintended Consequences". If you recall, one of the key points of this paper was that companies and governments were moving away from data silos, and instead focused on how data could be *joined* together to form even more comprehensive, statistically relevant, and accurate data on *everyone* – in short, increasing its value, and potentially its veracity, in addition to its volume, and potentially its variety.

Another topic frequently discussed in class thus far has been the framework of the "Five V's" – volume, velocity, variety, veracity and value. We have done several class/group activities so far and thought about many datasets in the real world and how to classify and leverage them along this framework. From judicial data about inmates and prisoners, to health data, to Twitter data, to restaurant review data, there have been extremely useful discussions and points made. You will leverage those discussions, class lectures, and material discussed, in this assignment.

2. Objective

Looking at the UFO sightings data, you may ask yourselves: "what other data is available that could be joined with this information" to affect its Five V's – intentionally, or unintentionally. For example, if you read the blog post linked above in the caption for Figure 1, you will see that the author somewhat dismisses these sightings as simply being indicative of people seeing "things" but not necessarily, e.g., "alien life" due to their close proximity to major cities with U.S. airports. In short, to the blog post author, these "sightings" are nothing more than exotic aircraft and lights, seen in the sky due to the observer living close to an airport. This is an interesting question, and will form the first part of your assignment which will involve finding an appropriate dataset of airports in the United States – and their location(s) – and then joining that data to the UFO sightings dataset. You will add new features to the dataset of "Airport Name" and "distance to the airport in miles" corresponding to the sighting location and its distance to the closest airport. You may also add other features (description of airport – e.g., "This airport is a modern airport with close ties to e.g., the US military, and located in Las Vegas, Nevada….").

What other datasets could you join the UFO sightings data to? For example, what about census population per closest city from the sighting location? What about weather phenomena in that spatial location, on that date? Maybe potentially add features related to the proximity to an air force base within X miles (where you set the threshold for X) of the sighting? Finally, what about traffic data in the closest county, city, metropolitan area to the sighting, to give you an idea of the observer's potential for actually seeing the UFO while stuck in traffic, versus, in a field near one's house, or while driving fast….etc.

You will choose at least three publicly accessible datasets along these lines to join the UFO 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 UFO sightings 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 different UFO events are within the data, and ask questions of your new UFO dataset. For example, you may ask, how many sightings were on the same day, within the same type of object, in which the sightings were all visible for longer than an hour, and where the population density of the closest city is under 50,000 people, with very little traffic, and a clear night sky that day?

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 http://tika.apache.org/1.16/gettingstarted.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 and install D3.js
    a. Visit http://d3js.org/
    b. Review Mike Bostock's Visual Gallery Wiki
        i. https://github.com/mbostock/d3/wiki/Tutorials
3. Download the UFO sightings data from the Dropbox
    a. https://www.dropbox.com/sh/s1lgh3fjtc5d12x/AADZUx0SVmBmw76SYeMylF2Sa?dl=0
    b. Make a copy of the original dataset (because you are going to modify/add to it in this assignment)
4. Begin by finding an appropriate dataset of major US airports and their cities, e.g., you can scrape:
    a. https://en.wikipedia.org/wiki/List_of_airports_in_the_United_States
    b. You can also directly use the airport codes dataset: https://github.com/datasets/airport-codes
    c. Write a Python program to add additional fields per sighting for:
        i. Airport Name
        ii. Distance to the airport in miles (from the location of each sighting) – you may need to use a Geocoder for this, see:

1. https://www.movable-type.co.uk/scripts/latlong.html (for the theory)
2. https://pypi.python.org/pypi/geopy (use GeoPy to implement lat/lng, or point distance)

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 UFO sightings 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 two you are adding from the airport dataset described in part 4)
    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. Compare Jaccard similarity, edit-distance, and cosine similarity
        i. Compare and contrast clusters from Jaccard, Cosine Distance, and Edit Similarity – do you see any differences? Why?
    d. How to the resultant clusters generated highlight the features you extracted? Be prepared to identify this in your report
7. (**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. 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 5. You may reuse your existing groups from discussion in class. Please email your group ONE time (from the main point of contact) to Simin, our TA after class on Thursday, February 15. If you have questions contact Simin via her e-mail address on the class website with

the subject: CS 599: Team Details.

4.2 UFO sightings dataset

Access to the sightings data is provided in the Dropbox. The dataset itself is 81Mb (compressed). 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-1.16.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/1.16/api.

Since you will be using Tika Python, you will want to read up on the Tika REST API, here: https://wiki.apache.org/tika/TikaJAXRS. 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 UFO sightings previously unanswered? What clusters were revealed? What similarity metrics produced more (in your opinion) accurate measurements? Why? What did the additional datasets suggest about "unintended consequences" related to UFO data? You should also clearly explain which datasets you used to join the UFO data and how you extracted the new features from each dataset.

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

1. Do UFO sightings only occur in rural areas?
2. Are UFO sightings mostly (greater than 75%) occurring in areas within 25 miles of an airport?
3. What do population demographics tell us about the areas in which UFOs occur?
    a. Densely populated? Sparsely populated?
4. What insights do the "indirect" features you extracted tell us about the data?
5. 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?

6. Submission Guidelines

This assignment is to be submitted ***electronically, by 12pm PT*** on the specified due date, via Gmail csci599spring2018@gmail.com. Use the subject line: CSCI 599: Mattmann: Spring 2018: BIGDATA Homework: Team XX. So if your team was team 15, you would submit an email to csci599spring2018@gmail.com with the subject "CSCI 599: Mattmann: Spring 2018: 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.
- Include your updated dataset TSV. We will provide a Dropbox location for you to upload to.
- 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_CSCI599_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 csci599spring2018@gmail.com.

Important Note:

- Make sure that you have attached the file the 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