Homework: Building and Exploring UFO Sightings using Data Science: Creating UFO Data Insights
Due: Friday, April 27, 2018 12pm PT

---

1. Overview

Figure 1: Current figure of http://irds.usc.edu/ufo.usc.edu, soon, http://ufo.usc.edu

In the third assignment, you will create an interactive set of visualizations that show off your data using the Data Driven Documents (D3) framework. This may include maps of UFO sightings compared to airports from assignment 1. It may include similarities of various sightings based on the features you generated. It may include information extracted from the UFO stalker sightings images, and the British UFO files OCR pipeline you built. In addition you will deploy the MEMEX Image Space open source application and the MEMEX ImageCat ("Image Catalog") to explore your UFO sightings images and find similarities between them.

You and your team will take these visualizations, and create a comprehensive "mini site" part of the broader http://ufo.usc.edu as shown in Figure 1 to live on as an example of the great work you did in exploring and investigating UFO sightings using data science.

2. Objective

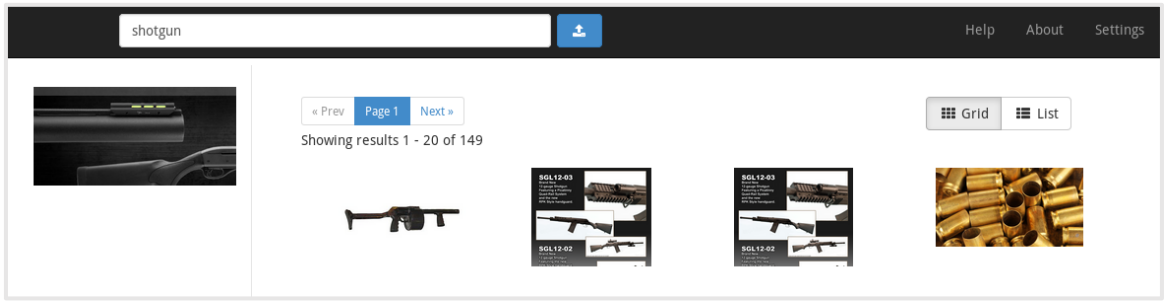
The objective of this assignment is to persist and make the great detective work and data science work you did exploring UFO sightings from the UFO awesome dataset; the British UFO sightings OCR dataset; and from the UFO stalker image dataset.

You will use explicitly the Data Driven Documents framework (D3) and its set of gallery visualizations used to explore and interact with your data.

We have built a template web site at http://irds.usc.edu/ufo.usc.edu and at GitHub at http://github.com/USCDataScience/ufo.usc.edu. You can explore the website and styles there. Your job on this is to add your work specifically under the Explore Visualizations tab and under the Gallery section of the website, by team number similar to e.g., if you click here on the Polar IRDS website ("Polar Data Insights"), here: http://polar.usc.edu/html/d3-examples.html. You will create a snapshot image of your team's work (that best represents your data and hard work e.g., like http://polar.usc.edu/images/team28.png), and then use this to link to your actual website with your D3 visualizations. You should make the visualizations connected together, e.g., such as the landing page here: http://polar.usc.edu/html/team28mime/index.html; you should use the USC IRDS UFO Data Insights template, and you will create a pull request to http://github.com/USCDataScience/ufo.usc.edu adding your team's JSON data, and your TSV v2 data.

You may need to summarize your TSV v2 data; to aggregate it so it displays well in your visualizations, or to prepare the data for interaction. In doing this you must choose to ingest your TSV data into Apache Solr or ElasticSearch, and then connect your D3 to those services. If you end up using one of these services, we will have a Solr and ElasticSearch service available at http://ufo.usc.edu and you may submit us a JSON dump as part of your assignment that we can load into it after the assignment is over and when you turn your assignment in so that your visualizations may live on.

Additionally, in continuing with our content extraction from Multimedia theme, you will also explore and install the ImageSpace open source application built on the MEMEX program (http://github.com/nasa-jpl-memex/image_space) as well as the ImageCat ("ImageCatalog") application (http://github.com/chrismattmann/imagecat/).



ImageSpace is an investigative forensic tool allowing you to search and compare images based on similarity using a variety of algorithm plugins including the Social Media Query Toolkit (SMQTK) http://github.com/kitware/SMQTK, and the Fast Library for Approximate Nearest Neighbors (FLANN), https://www.cs.ubc.ca/research/flann/. ImageCat is an ETL/ingest application that can ingest 10s of millions of images, extract their EXIF metadata and perform OCR on them using Tesseract and Apache Tika. The

ETL/ingest performed is into an Apache Solr index. The resultant index is used by ImageSpace.

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

3. Tasks

1. Take your TSV v2 dataset and convert the data to JSON to use in D3.
    a. You may need to write scripts to summarize your data for D3. As a start, consider using ETLlib (http://github.com/chrismattmann/etllib) and its tsvtojson tool.
2. Pick 10 visualization types from https://github.com/d3/d3/wiki/Gallery and create the associated UFO Data Insights web pages and associated JSON data to display them showing off your dataset (see Task 1). Consider similarity, consider using the questions from Assignment 1 and Assignment 2 that you answered in your reports and how the D3 visualizations will help you answer them.
    a. Develop scripts for summarizing and preparing your TSV v2 datasets for D3 JSON conversion.
    b. The scripts you write are part of your delivery for the assignment. Please provide documentation for each script that you create in order to visualize the data using D3. Make sure that your scripts are portable and there are a simple set of instructions on how to run them. Any libraries that the scripts depend on should be clearly indicated.
3. Ingest your sightings data from TSV v2 JSON you created in Tasks 1 and 2 into Apache Solr (http://lucene.apache.org/solr/) and/or ElasticSearch (http://elastic.co). Both have adequate documentation and are easily installed. If you choose this step, please identify 1 more visualization type from https://github.com/d3/d3/wiki/Gallery and link it to your Solr or ElasticSearch web service.
4. Install ImageCat (http://github.com/chrismattmann/imagecat) or you may use your own custom ingest scripts using Tika Python (http://github.com/chrismattmann/) and Tika-Server, etc. The end goal here is to write scripts that extract Tika metadata and also that provide OCR to get the image content into Solr for use by ImageSpace.
    a. Ingest your UFO Stalker image data (whatever you got in assignment #2) into ImageCat using the provided instructions and scripts or the ones you write on your own using Tika-Python/Tesseract OCR.
5. Install Image Space (http://github.com/nasa-jpl-memex/image_space). See instructions on using Docker or simply installing by hand.
    a. Point your ImageSpace at your ImageCat that you built in Step 4.
    b. Install the FLANN index: https://github.com/nasa-jpl-memex/image_space/tree/master/flann_index
        i. Update listings.txt will thel file paths to your images from the UFO stalker images
        ii. Run https://github.com/nasa-jpl-memex/image_space/blob/master/flann_index/start.sh

iii. Make sure to enable the https://github.com/nasa-jpl-memex/image_space/tree/master/imagespace_flann plugin.

c. Browse and find similar images and use the ImageSpace search index and search the Image forensics and OCR.

6. Submit your Solr or ElasticSearch index by tarring it up and gzipping it. Both your index for your non image UFO sightings along with your ImageCat indices.
   a. Name the non image index as ufo-sightings-noimage-index.tar.gz
   b. Name the ImageCat index as ufo-sightings-imagecat-index.tar.gz
   c. We will provide you a DropBox location to submit your data.
7. (**EXTRA CREDIT**) Make ImageSpace work with ElasticSearch
   a. See: https://github.com/nasa-jpl-memex/image_space/blob/master/imagespace/server/imagesearch_rest.py as an example of how Image Space is talking to Solr. Update it to be configurable to also talk to ElasticSearch.
   b. Also see: https://github.com/nasa-jpl-memex/image_space/blob/master/imagespace/server/__init__.py
8. (**EXTRA CREDIT**) Try other Image Space plugins
   a. Try the SMQTK plugin, see instructions here: https://github.com/nasa-jpl-memex/image_space/tree/master/imagespace_smqtk
   b. Try the VideoSpace plugin on your videos (if you have any) from the UFO Stalker dataset. See: https://github.com/nasa-jpl-memex/image_space/tree/master/videospace
   c. What types of videos are similar?
   d. What types of images are similar? How is SMQTK different than FLANN? Write this information in your report to receive extra credit.

4. Assignment Setup

4.1 Group Formation

You should keep the same group from your assignment one. There is no need to send any emails for this step.

5. Report

Write a short 4 page report describing your observations. In particular I am interested in answers to the below questions:

1. Why did you select your 10 D3 visualizations?
   a. How are they answering and showing off your features from assignments 1 and 2 and the work you did?
2. Did Image Space allow you to find any similarity between UFO sightings images that previously was not easily discernible based on the text captions and object identifications you did?

Also include your thoughts about Image Space and ImageCat – what was easy about using them? 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: DATAVIS 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: DATAVIS 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 convert your TSV v2 data to JSON, and also likely scripts to perform ingestion into ImageCat and/or your own Solr or ElasticSEarch.
- Include your updated Indices as specified in Task 6. 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, 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_DATAVIS.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_DATAVIS.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.
- Make sure you have your team members listed on your report
- Make sure your report is self-contained. Any plots, stats, etc. should be included in your report.
- Make sure you clearly answer and specify all the questions listed.

6.1 Late Assignment Policy

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