Homework: Building Visual Apps to Explore Fake Scientific People and Literature using Data Science: Creating Data Insights
Due: Friday, April 30, 2021 12pm PT

---

1. Overview

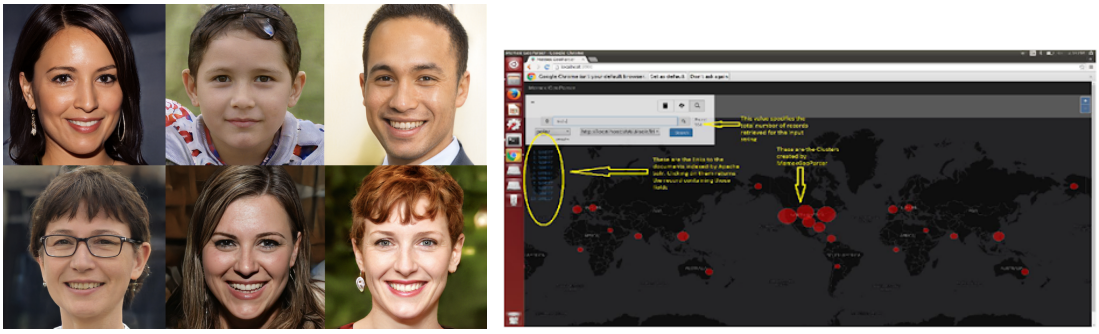


Figure 1: Examples from http://thispersondoesnotexist.com and locations from MEMEX GeoParser.

In the third assignment, you will create an interactive set of visualizations that show off your active social engineering defense (ASED) interactions and work you've done through the first two assignments using the Data Driven Documents (D3) framework. This may include maps of attack origin locations compared to age groups from assignment 1. It may include similarities of various attackers and victims based on the features you generated. It may include information extracted and generated from your PhishIris dataset, and/or emails that you automatically generated using GPT-2, and the fake attacker pictures you built. In addition, you will deploy the MEMEX Image Space open-source application to explore your generated fake attackers and find similarities between them and you will deploy the MEMEX GeoParser application to explore the locations present from your original data, and your newly generated attack data.

You and your team will take these visualizations, and apps and create a comprehensive "mini site" to demonstrate as an example of the great work you did in exploring and investigating how to perform active social engineering defense (ASED) 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 active social engineering defense (ASED).

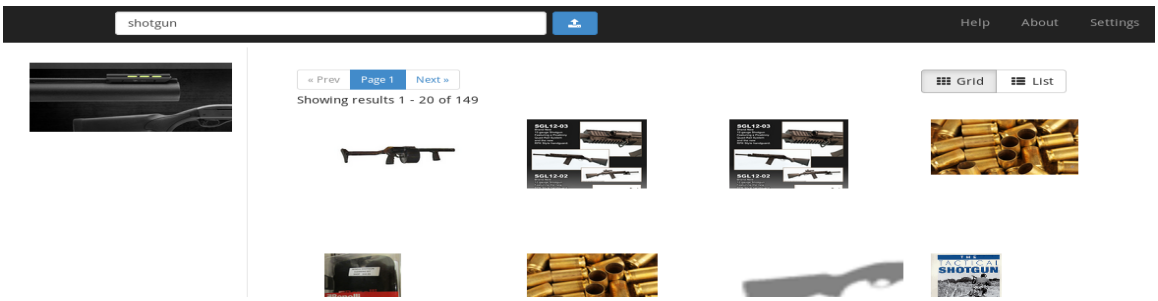
You will use explicitly the Data Driven Documents framework (D3) and its set of gallery visualizations used to explore and interact with your data.

We have built several template web sites in the past in the IRDS group, for example, see the one from 2018 for UFO research 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 use this as a reference and add your work specifically under the Explore Visualizations tab and under the Gallery section of the website, by team name. 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.

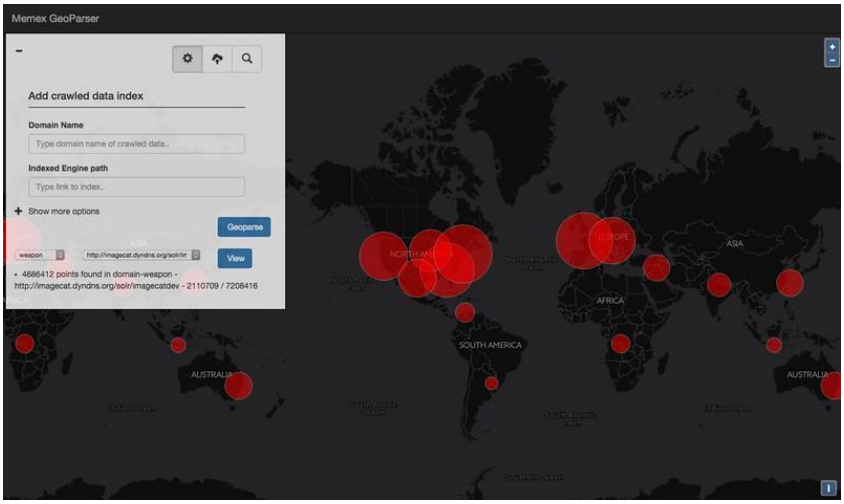
You may need to summarize your TSV data from assignment 2, and/or assignment 1; to aggregate it so it displays well in your visualizations, or to prepare the data for interaction. In doing this you must choose to ingest your TSV data into Apache Solr or ElasticSearch, and then connect your D3 to those services. You may submit 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 the 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). There is an integrated Wiki instruction page here (https://github.com/nasa-jpl-memex/image_space/wiki/Quick-Start-Guide-with-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/. The application includes a backend called ImageCat that 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.

Additionally, you will deploy the MEMEX GeoParser visual application (https://github.com/nasa-jpl-memex/GeoParser) to explore the location information in your data. GeoParser is a full stack web application that takes in documents, or data, and then analyzes all the mentions of locations in those documents, and then visualizes them on a map like below:



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

3. Tasks

1. Take your TSV 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 5 visualization types from https://github.com/d3/d3/wiki/Gallery and create the associated 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 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 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.
4. Install Image Space via https://github.com/nasa-jpl-memex/image_space/wiki/Quick-Start-Guide-with-ImageCat. **This is optional:** You can also write your own custom ingest scripts using Tika Python (http://github.com/chrismattmann/) and Tika-Server, etc.
    a. Ingest your fake attacker image data (whatever you got in assignment #2) into Image Space using the provided instructions and scripts or the ones you write on your own using Tika-Python.
    b. Browse and find similar images and use the ImageSpace search index and search the Image forensics and similarity (SMQTK).

5. Submit your Solr or ElasticSearch index by tarring it up and gzipping it. Both your index for your scientist data along with your ImageCat indices.
6. Install MEMEX GeoParser and run it against your TSV data and location data from assignment 1 and 2.
7. (**EXTRA CREDIT**) Submit Pull request and improve GeoParser, and/or Image Space. Improvements to the software will be considered for 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 4-6 page report describing your observations. In particular I am interested in answers to the below questions:

1. Why did you select your 5 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 the fake attacker images that previously was not easily discernible?
3. What type of location data showed up in your data? Any correlations not previously seen, e.g., from assignment 1?

Also include your thoughts about Image Space and ImageCat – what was easy about using them? What wasn't?
Add your individual contributions.

6. Submission Guidelines

This assignment is to be submitted ***electronically, by 12pm PT*** of **Friday, 30th April 2021**, via Gmail dsci550spring2021@gmail.com. Use the subject line: DSCI 550: Mattmann: Spring 2021: DATAVIS Homework: Team XX. So if your team was team 15, you would submit an email to dsci550spring2021@gmail.com with the subject "DSCI 550: Mattmann: Spring 2021: 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 5. We will provide a Google Drive location for you to upload to.
- Also prepare a readme.txt containing any notes you'd like to submit.

- 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_DSCI550_HW_DATAVIS.zip

  Use only standard zip format. Do **not** use other formats such as zipx, rar, ace, etc.
- If your homework submission exceeds Gmail's 25MB limit, upload the zip file to Google drive and share it with dsci550spring2021@gmail.com.
- Please try to adhere to the submission guidelines which was mentioned in the Slack channel by the TA.

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