DSCI 550 Assignment #3 Report, Spring 2021
Dr. Chris Mattmann

Building Visual Apps to Explore Fake Scientific People and Literature using Data Science: Creating Data Insights

By Group Banana: Katie Chak, Sarah Pursley, Madeleine Thompson,
Claudia Winarko, and Amber Yu

1. D3 Visualizations

1.1 Why select these visualizations? How are they answering and showing off the features from assignments 1 and 2?

We created a circular barplot to display the most frequently used attacker titles, a bubble chart to display the most frequent words used in the emails separated by attack type, a time-series plot of unemployment data in the country and year that the email was sent, a heatmap comparing the day of the week and time of day each email was sent, and a histogram of the results of three different similarity/distance measures when we ran Tika Similarity on our original email set. These visualizations combine the data that we collected in assignments 1 and 2 and allow us to see the greater trends across this massive dataset. For example, it is clear that most emails in this dataset were sent on a workday (M-F) between the hours of 8AM and 12PM, and that most attackers used a simple title of “Mr.” or “Mrs.”. See the appendix for copies of these visualizations.

2. ImageSpace

2.1 Did ImageSpace allow us to find any similarity between the fake attacker images that previously was not easily discernible?

Yes, ImageSpace is able to help us find similarities between images that are based on more intricate and subtle features, which we were unable to identify before. Since there are a total of 800 generated faces that we needed to process, it was very difficult to screen through all of them to find similarities manually. Therefore, ImageSpace was a great help to process a large number of images in a reasonable amount of time.

However, we found that the algorithm for Image Space similarity is a Black Box, and was not easily understandable for why it selects certain images. There are many features that the program is taking under consideration, such as color, facial feature, index, etc. Since we do not have a full understanding of the algorithm, it could make mistakes occasionally that are not explainable. Thus, we cannot fully rely on the Image Cat system to detect similar images, and having a human double-check the similar images is an essential step to ensure accurate and unbiased results.

2.2 What was easy about using it? What wasn't?

We encountered many challenges when setting up ImageSpace using Docker. It is not compatible with some of the latest versions of Mac hardware, and we discovered that it needs specific environment and setups. When using command lines to install, our team had to keep tweaking the code, and we navigated through the directory with caution. Because ImageSpace is dependent on Docker, running it successfully requires a certain level of familiarity with Docker and its operations. ImageSpace also needs memory, and we see potential difficulties to run it on everyday personal laptops, especially if the dataset is large. In order to run the system efficiently, a higher GPU and RAM would be more optimal.

We also discovered a set of very useful features in the ImageSpace system. ImageCat, ImageSpace, and SMQTK integrate seamlessly to process the images and to find similar counterparts. Its site has an user-friendly interface, and the experience design is intuitive, which is friendly for researchers with or without computer science backgrounds. Although running a large amount of images through the system is time-consuming, separating the dataset into smaller sets makes the searching significantly faster. For day-to-day analysis of smaller datasets, we believe that ImageSpace presents impressive abilities to find a decent amount of similar images in a short amount of time.

3. GeoParser

3.1 What type of location data showed up in our data? Any correlations not previously seen, e.g., from assignment 1?

GeoParser was able to locate the points of 6,060 out of the 7,178 fraudulent emails in our dataset. Most of the locations were found around the regions of North Africa (Morocco and Algeria) as well as the European region across North Africa (Portugal, Italy, Monaco). Further, there were a significant number of points situated in the US, with most points along the East Coast and California. We have also discovered that points scattered around the Central and Southern part of Africa, as well as around Asia.

There are some correlations that were not previously seen from previous assignments. For instance, in assignment 1 we were instructed to find the attackers' location, which resulted in most of the locations situated in Central and West Africa (Nigeria and the Democratic Republic of the Congo) as well as the US. However, GeoParser's analysis of the fraudulent emails' text extracted locations that did not align with the results from assignment 1. Instead, most of the extracted points were situated in North Africa and Europe. See the appendix for examples of GeoParser's visualization of the fraudulent emails dataset.

4. Individual Contributions

Katie Chak - Researched and ran ImageSpace. Gathered and modified required command lines for the repositories.

Sarah Pursley - Researched and created four of the five D3 visualizations, formatted all D3 visualizations, and pre-processed data.

Madeleine Thompson - Pre-processed json and tsv data for use in D3 visualizations, created a Solr index for this data, created the histogram comparison of different similarity/distance metrics with Tika-Similarity data in D3, and formatted D3 visualizations.

Claudia Winarko - Installed GeoParser and ran it against our TSV data and location data from assignment 1 and 2.

Amber Yu - Ran ImageSpace, organized and wrote report.

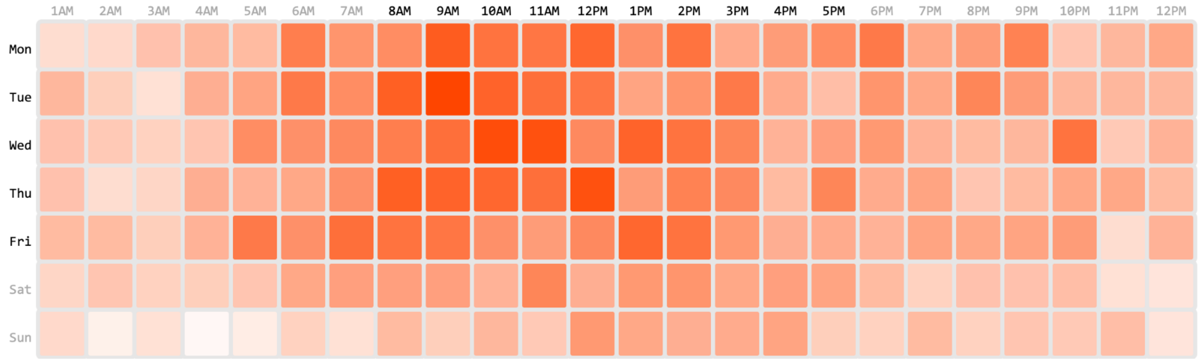
Appendix

D3 Visualizations:

Heatmap of Email "Sent" Time Signatures

A heatmap comparing the most popular times for attackers to send emails, broken down by day of the week and time of day.

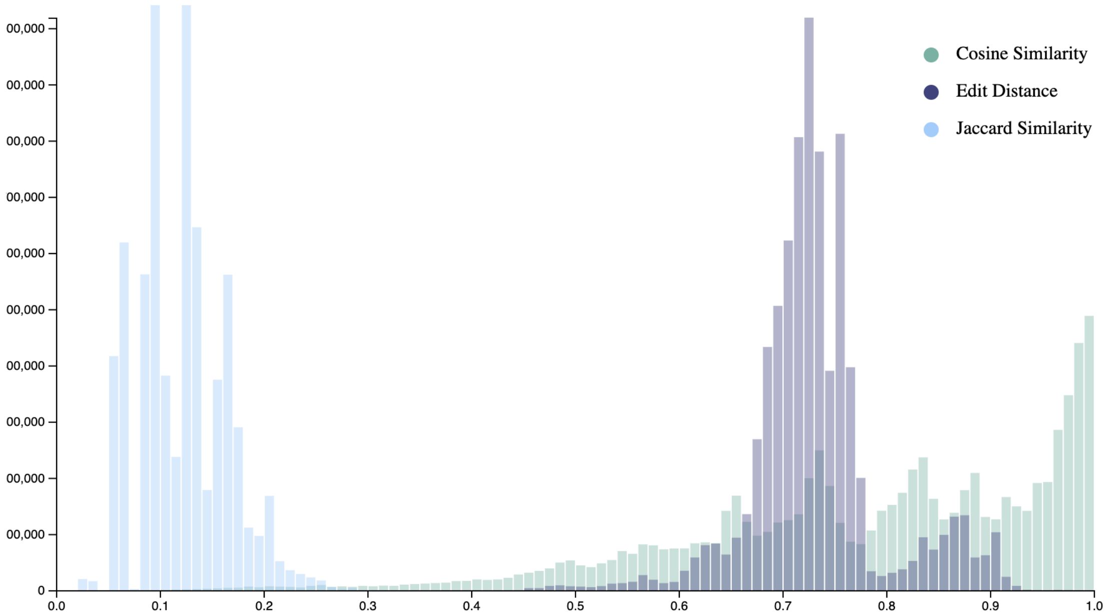
The darker the square, the higher frequency of emails sent during that time period.



Histogram of Different Tika-Similarity Measures

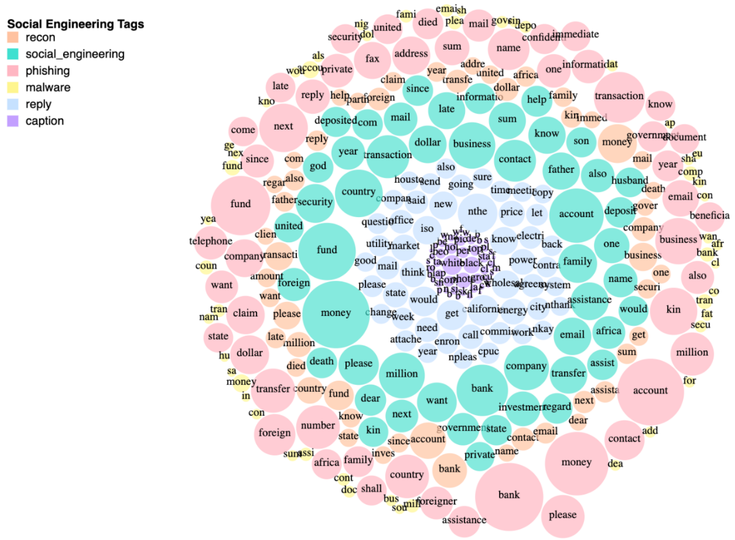
A histogram comparing the similarity scores of Cosine Similarity, Edit Distance, and Jaccard Similarity.

This figure will take approx. 1 minute to load, please be patient.



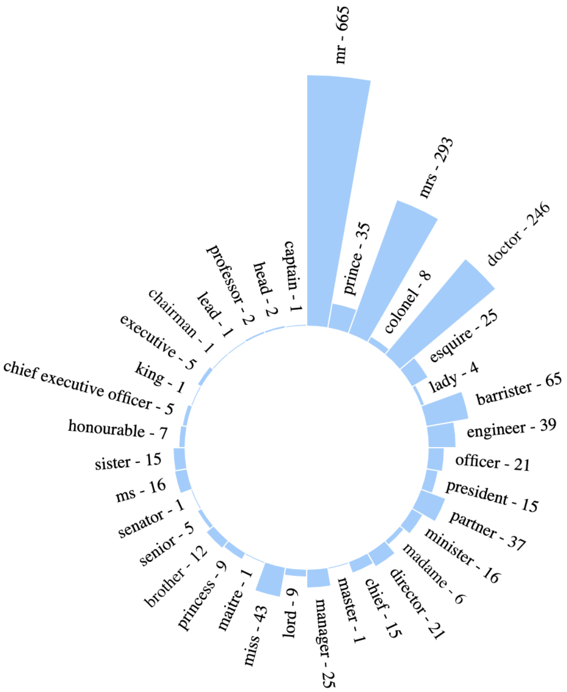
Bubble Chart for Content by SE Type

A bubble chart to display the most frequent words in the fraudulent email contents. Words are separated by the social engineering tags we identified in task one.



Frequently Used Attacker Titles

A circular barplot to display the most frequently used titles by attackers in the fraudulent email corpus.

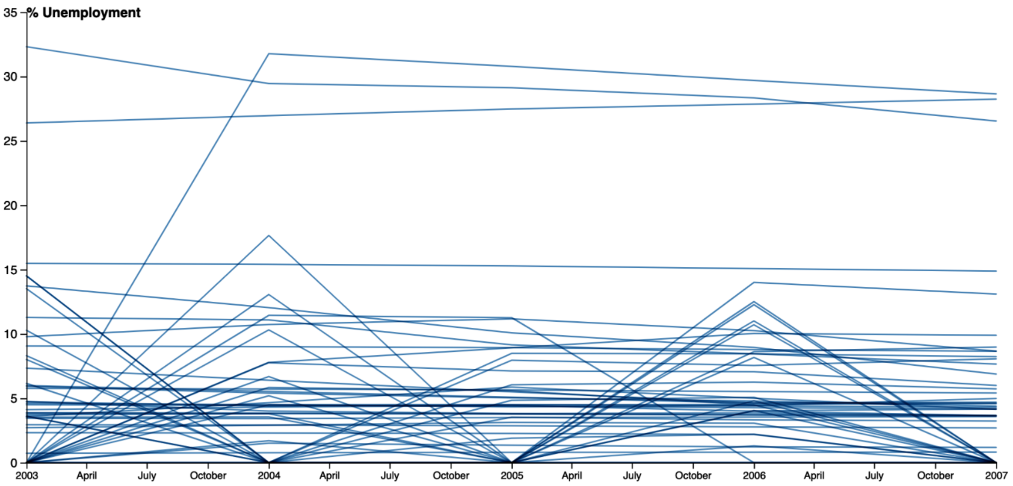


2003 - 2007 Unemployment Rates by Country

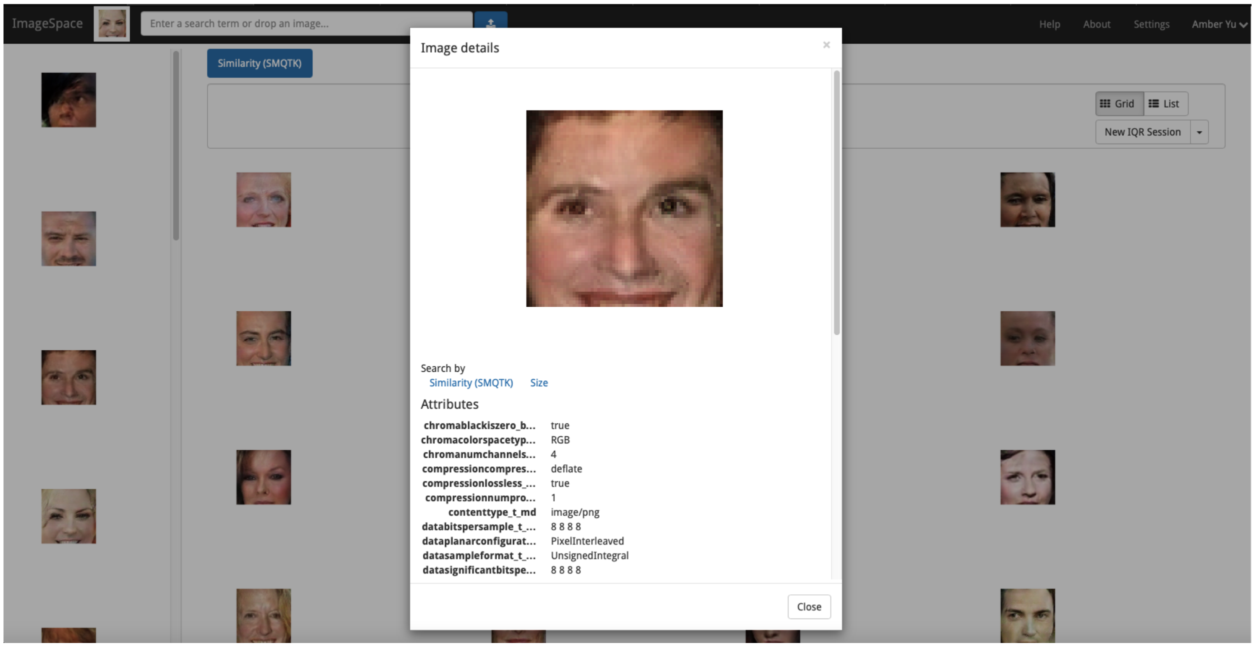
Added from an external dataset in Assignment 1, we joined recent global unemployment data to our TSV.

These are the top 51 countries of origin for the fraud emails and their corresponding unemployment rates over the time from 2003-2007

Details: X-axis is a continuous date scale and Y-axis is unemployment rates. Please hover each entry to see which line is each country and the specific unemployment rate for that selection.

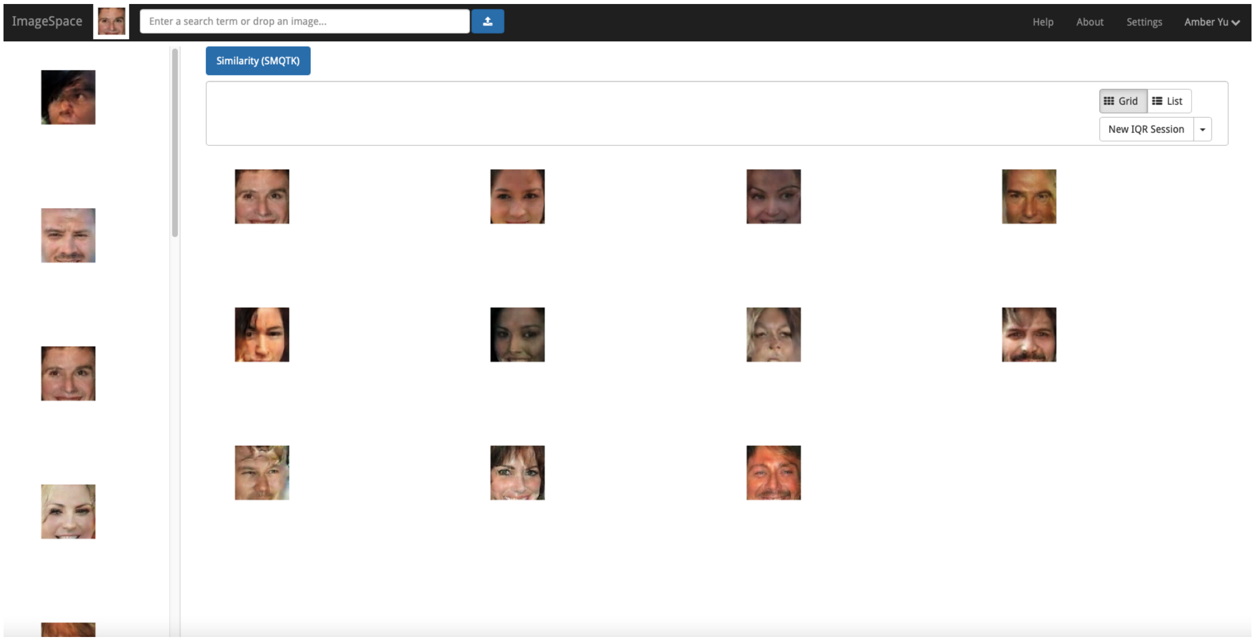


ImageSpace Screenshots:



Attributes

chromablackiszero_b...	true
chromacolorspacetyp...	RGB
chromanumchannels...	4
compressioncompres...	deflate
compressionlossless_...	true
compressionnumpro...	1
contenttype_t_md	image/png
databitspersample_t_...	8 8 8 8
dataplanarconfigurat...	PixelInterleaved
datasampleformat_t_...	UnsignedIntegral
datasignificantbitspe...	8 8 8 8



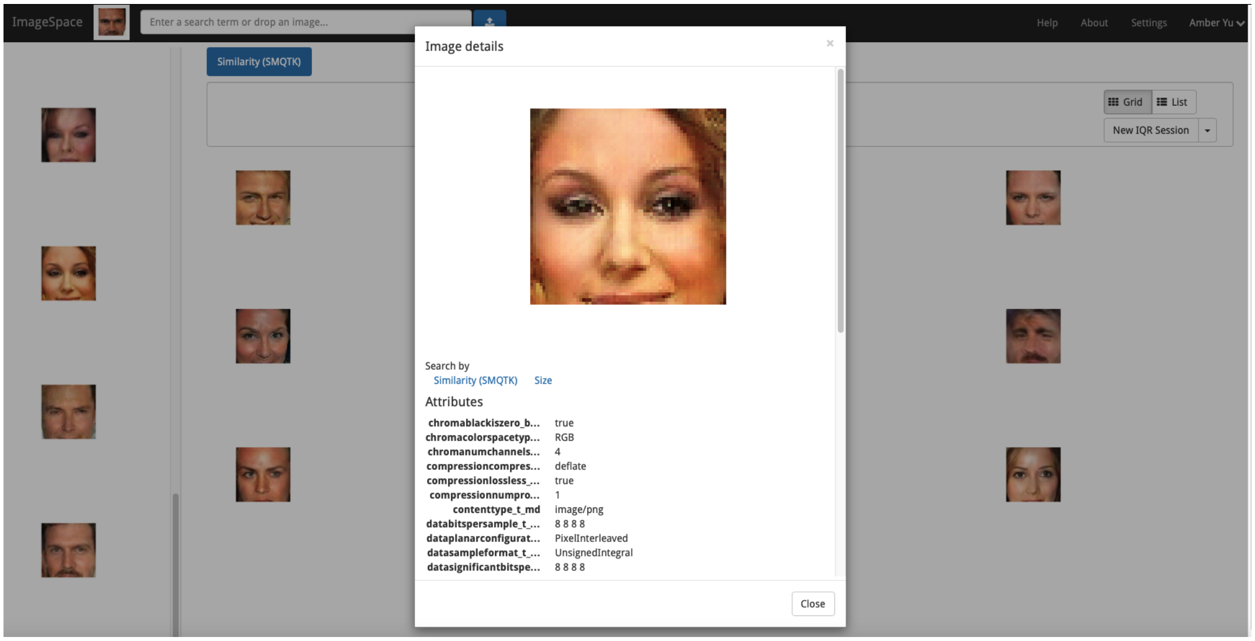
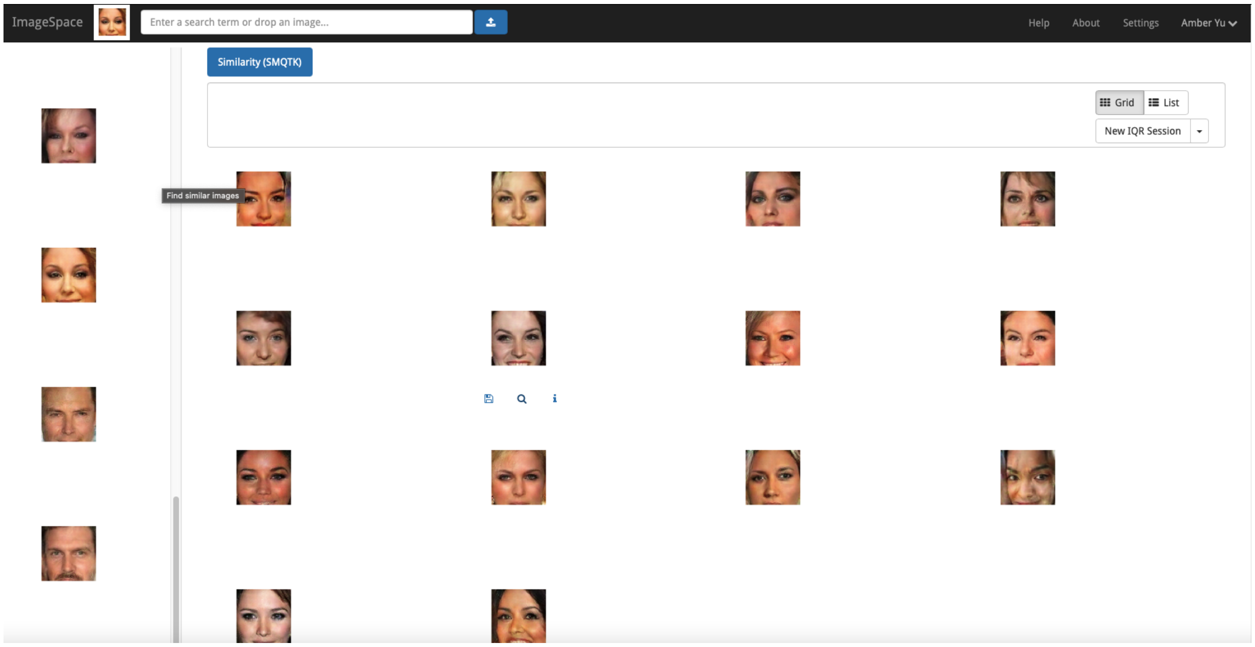


Image details

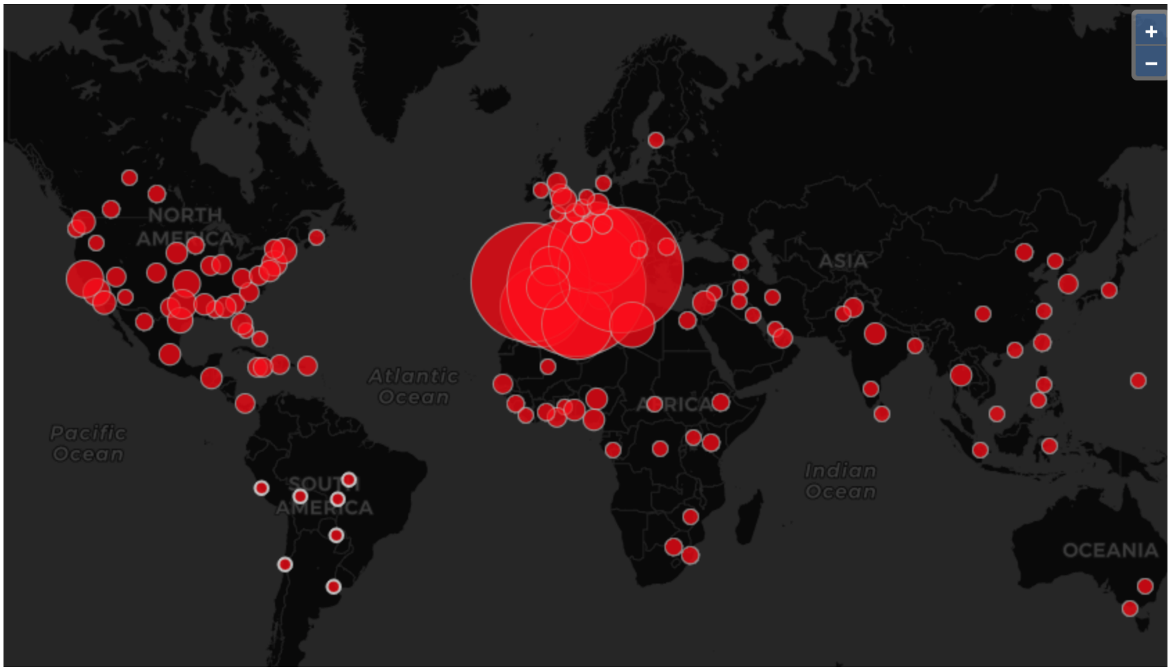
Search by
Similarity (SMQTK) Size

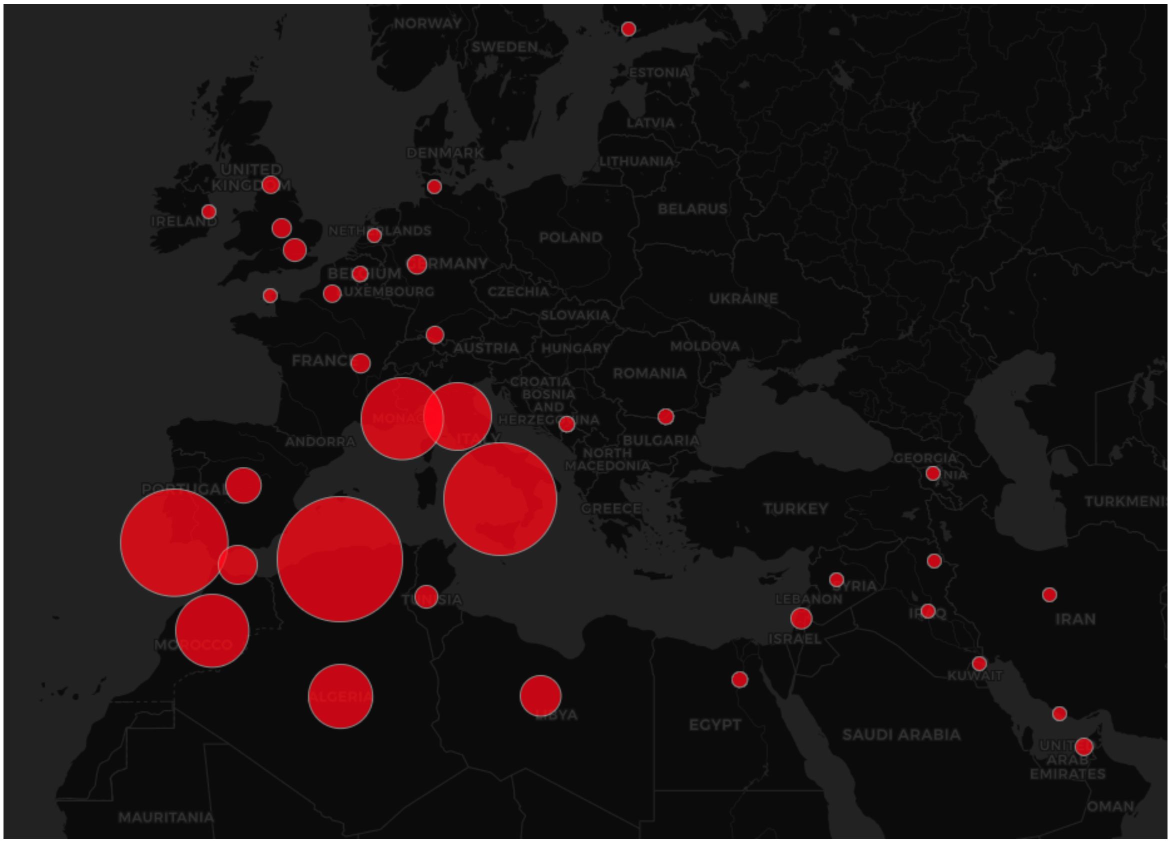
Attributes

chromablackiszero_b...	true
chromacolorspacetyp...	RGB
chromanumchannels...	4
compressioncompres...	deflate
compressionlossless_...	true
compressionnumpro...	1
contenttype_t_md	image/png
databitspersample_t_...	8 8 8 8
dataplanarconfigurat...	PixelInterleaved
datasampleformat_t_...	UnsignedIntegral
datasignificantbitspe...	8 8 8 8



GeoParser Visualizations:


+
−
NORTH AMERICA
Atlantic Ocean
Pacific Ocean
SOUTH AMERICA
AFRICA
ASIA
Indian Ocean
OCEANIA


NORWAY
SWEDEN
ESTONIA
LATVIA
DENMARK
LITHUANIA
UNITED KINGDOM
IRELAND
BELARUS
NETHERLANDS
POLAND
GERMANY
BELGIUM
LUXEMBOURG
CZECHIA
UKRAINE
SLOVAKIA
AUSTRIA
HUNGARY
MOLDOVA
FRANCE
ROMANIA
CROATIA
BOSNIA AND HERZEGOVINA
MONACO
ITALY
ANDORRA
BULGARIA
NORTH MACEDONIA
GEORGIA
PORTUGAL
GREECE
TURKEY
TURKMENI
TUNISIA
SYRIA
LEBANON
ISRAEL
IRAQ
IRAN
MOROCCO
ALGERIA
LIBYA
EGYPT
SAUDI ARABIA
KUWAIT
UNITED ARAB EMIRATES
OMAN
MAURITANIA