Homework: Spatial Search using Apache Solr, SIS and Google Maps
Due Date: May 7, 2014

---

1. Introduction

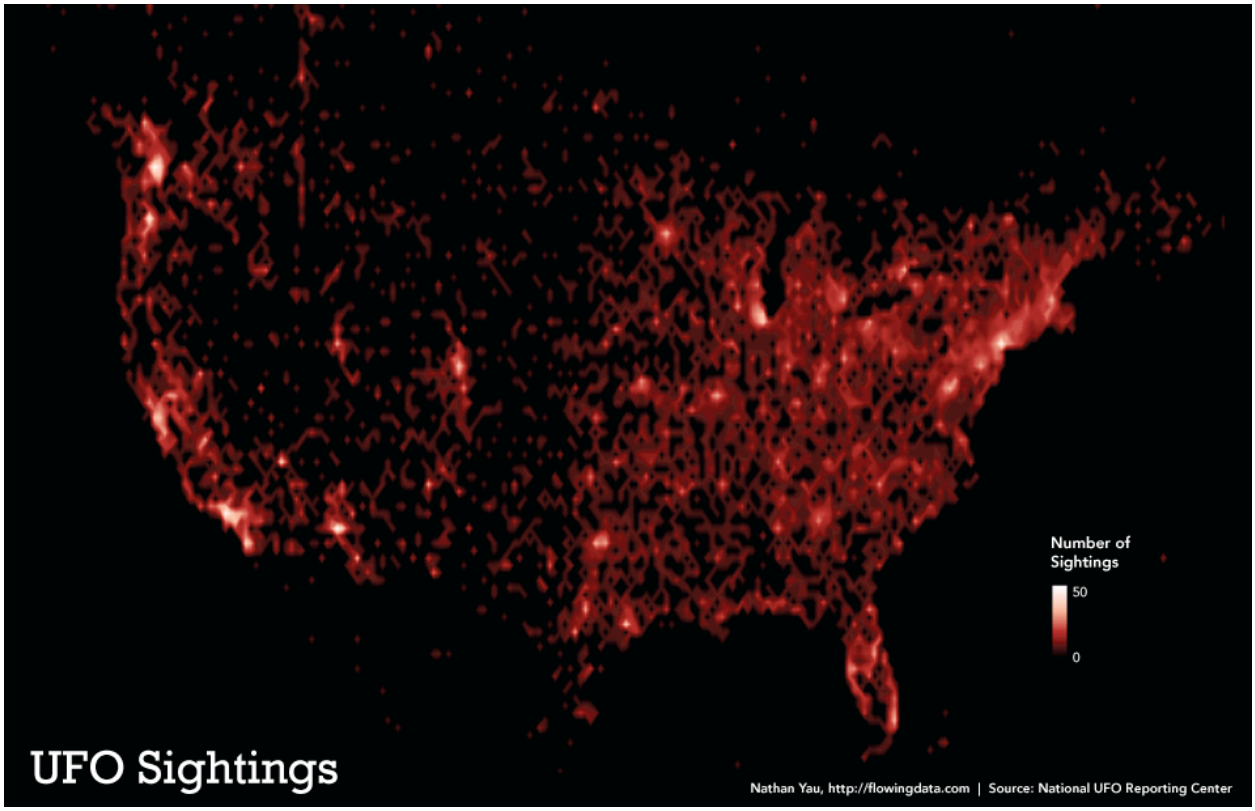
So, we're at the end of the road here with assignments. Let's recap what you've done so far:

1. In the Tika homework, we gave you vault.tar.gz and we asked you to leverage Apache Tika and really dig into those secret PDF files from the FBI to figure out how many of them *actually* had to do with UFOs.

2. In the Nutch homework, we showed you how we originally got vault.tar.gz. You used Apache Nutch to crawl the World Wide Web, to download all of the PDF content from our mirror of the FBI's vault website. You then extracted the PDF files from the segmented, distributed binary data that Nutch stored in Apache Hadoop.

Now it's time to look for another important and emerging property in the Search Engine research domain: spatial search and visualization. Wouldn't it be cool to visualize and determine where those PDF documents that you downloaded, and decimated, were located across the United States? Do you think that UFO related documents typically center on airports? Do you think that the JFK documents are centered on Texas, where he was shot and where he died? What about the documents that the FBI collected on the Branch Davidians and Waco, TX? Where were those located? Do you think that spatial location of the PDF data will generally make sense and coincide with the location of the event? Time to find out.

Nathan Yau, who runs the http://flowingdata.com blog decided to do a similar experiment, with a different dataset from the National UFO reporting center. Check out this cool map he made.



2. Overview

In this assignment, you will:

- Set up the Apache Solr search engine technology, which, along with what you learned in Homework 3, will allow you to load and tag your PDFs from `vault.tar.gz`.
- Leverage the Geonames.org dataset, http://geonames.org/ which maps names and geographic locations to latitudes and longitudes, in order to geo-tag each of the PDF files with their approximate location of interest.
- Load the geo-tagged data into Solr, making it available for spatial search.
- Develop a web page that enables searching the PDF files for conspiracies. Your web page will need to integrate Google Maps to plot locations of the search results on a map.
- (Optional) Dump your Solr data via its GeoRSS response writer into an emerging Apache project called Apache SIS. SIS provides a quadtree index and web service to perform point-radius and bounding box searches of associated GeoRSS data. It also easily plugs into Google Maps, thus letting you visualize the results of your queries.

3. Implementation

3.1 Identification of 3 conspiracies

First, we'd like you to pick 3 of the following conspiracies that we have identified in the FBI's vault dataset:

1. JFK
2. UFOs
3. Watergate
4. Waco, TX
5. J. Edgar Hoover

These will form the basis for investigating and visualizing the vault dataset.

For each conspiracy that you choose, come up with a list of search terms that you could use to find documents related to that conspiracy.

3.2 Downloading and Configuring Apache Solr

The next step will involve grabbing the Apache Solr search technology. You can grab it from:
http://www.apache.org/dyn/closer.cgi/lucene/solr/

Follow this install guide to get Solr installed on top of Apache Tomcat:
http://wiki.apache.org/solr/SolrTomcat

Other documentation can be found in the Solr wiki:
http://wiki.apache.org/solr/

3.3 Geo-tagging the PDF documents in Solr

Next, we'll want to geo-tag all[1] of the PDF documents from `vault.tar.gz` with a geographic location. To accomplish this, we're going to use Tika to grab out the most frequently occurring terms in the document. In a similar fashion to the Tika homework your job is to write a program that will use Tika to extract the text from the PDF document, and compute the list of unique words in the document, and sort them by their frequency of occurrence, from highest to lowest.

Armed with this list, we'll now try to use the geonames.org dataset to geo-locate the document based on the highest occurring term, as described in https://issues.apache.org/jira/browse/SOLR-2073[2]. If geonames.org can't locate the term, move onto the next term. Repeat this process until you have a latitude and longitude for the document. Once you have the latitude and longitude, index the document in Solr, using one of the many available Solr client APIs:

- http://wiki.apache.org/solr/SolJava
- http://wiki.apache.org/solr/SolPython
- http://code.google.com/p/solr-php-client/

3.4 Build a web page for searching PDF documents and visualizing the results on a map

Create a web page that provides a simple user interface for querying your Solr index. Your interface should accept multiple search terms, and be able to plot all documents in the search results on a map based on the geographic location that they were tagged with earlier, using the Google Maps API:
http://code.google.com/apis/maps/index.html

You can check out the Solr AJAX JavaScript client to get an idea of how to query Solr using JavaScript:
http://evolvingweb.github.com/ajax-solr/

This IBM DeveloperWorks article also shows how to link Solr with Google Maps:
http://www.ibm.com/developerworks/opensource/library/j-spatial/

Make use of your search engine to query for each one of your three conspiracies, using the search terms you came up with earlier.[3] For each conspiracy, you should either generate a map with markers for each document, or produce a heat map (akin to Nathan Yau's map) based on the location of the documents.

3.5 Hints

This is a longer assignment, far more complex than the previous ones, so get started early. There are a **lot** of moving parts here.

The hardest part of the assignment will be the term frequency generation and the geo-tagging with Geonames.org data. Focus on that portion of the assignment first.

---

[1] Do this with the entire dataset, not only limited to documents relevant to your selected conspiracies.
[2] A former CS572 project, from William Quach!
[3] At this point you should also check each of your search terms to make sure they do help to find relevant results!

4. Options (Extra Credit)

Figure out how to dump the Solr metadata into the Apache Spatial Information System (SIS) technology, which you can find here: http://incubator.apache.org/sis/. See the SIS README file here:
http://svn.apache.org/repos/asf/incubator/sis/tags/0.2-incubating/README.txt

The README will explain how to install SIS on top of your Apache Tomcat instance. Once you have SIS installed, find the GeoRSS ResponseWriter for Solr that Professor Mattmann and W. Quatch made here: https://issues.apache.org/jira/browse/SOLR-2074

Install the GeoRSS response writer and plug Solr into SIS. Once you have GeoRSS from Solr loaded into your SIS Location Service, perform various queries (bounding box and point/radius) and play around with the data. Look for some trends or patterns in the data set, and make a list of 3 such discoveries you make from exploring your data spatially with SIS. Be prepared to demonstrate with examples when presenting your assignment for grading.

5. Assessment

This assignment is to be submitted ***electronically, by 12pm PT*** on the specified due date, via Gmail to chris.mattmann@gmail.com. Use the subject line: CSCI 572: <Horowitz|Mattmann>: Spring 2014: Solr Homework: <Your Lastname>: <Your Firstname>. So if your name was Lord Voldemort, and you were taking Dr. Horowitz's course, you would submit an email to chris.mattmann@gmail.com with the subject "CSCI 572: Horowitz: Spring 2014: Solr Homework: Voldemort: Lord" (no quotes).

- Prepare a report as a PDF file (`Lastname_Firstname_TIKA.pdf`) and include it in your submission. Your report should detail all of the important aspects of your assignments, highlights that you'd like the instructors to review, and a link to your demo.
- Compress all of the above into a single zip archive and name it according to the following filename convention:

  <lastname>_<firstname>_CSCI572_HW_SOLR.zip

  Use only standard zip format. Do **not** use other formats such as zipx, rar, ace, etc.

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.