**Spatial Search using Apache Solr, SIS and Google Maps**

# introduction

In this assignment, we are setting up the Apache Solr search engine technology where we load and tag the PDFs from the FBI vault. We then map names and geographic locations to latitudes and longitudes in order to geo-tag each of the PDF files with their approximate location of interest. It is then available for spatial search. We develop a webpage to plot the results of the search performed. We also implement Apache SIS which allows us to perform point-radius and bounding box searches of associated GeoRSS data.

# Team members

1. Anoop Kulkarni | Prof. Mattman
2. Garima Baweja | Prof. Horowitz
3. Nidhi Jain | Prof. Horowitz
4. Vidhi Sampat | Prof. Horowitz

# important aspects of the assignment

* The conspiracies we have picked are :

1. UFOs
2. Watergate
3. J. Edgar Hoover

* We first run the Apache Tika code we wrote in HomeWork 1 to fetch the files containing the words from our keywords. The keywords we are using are:

1. ufos
2. disk
3. flying saucer
4. vessel
5. spacecraft
6. watergate
7. crp
8. gerald
9. wiretap
10. bernard barker
11. frank sturgis
12. j. edgar hoover
13. subversion
14. kennedy
15. cointelpro
16. hoover
17. livezly
18. brown

* **We add the location field to the schema for each document by querying the geonames dataset. This location field contains the latitude and longitude.**
* While querying, we enter the query in ‘q’ part of Solr and Solr checks the same in the documents and returns the files that contains those keywords.
* We then map those documents to the correct location. This is done using the Google Map API.

# highlights

* We have used a diverse set of keywords.
* The code not only searches for 3 keywords as specified in the specification but for any number of keywords and returns documents that contains all the keywords specified.
* We have not restricted the code to return locations only within USA but all over the world.

# problems and solutions

* While trying to aces the Solr index via webpage hosted on Apache we ran into the cross-site scripting issue that we had to resolve by using snippet of PHP code as a proxy server to relay the result from the Solr index to our html page.
* We found the performance in terms of number of words extracted per document of Apache Tika to be better on Windows as compared to Linux.

# link to the demo

<http://youtu.be/CKqUke-pcEg>

# extra credit

* To integrate SIS with Solr we had to recompile Solr with the GeoRSS modifications.
* Once that was done we had to modify the schema.xml and add the following lines:

<field name="name" type="text\_general" indexed="true" stored="true"/>

<field name="link" type="string" indexed="true" stored="true"/>

<field name="location" type="location" indexed="true" stored="true"/>

<field name="pubDate" type="date" indexed="true" stored="true" multiValued="false"/>

* We queried solr with wt=georss and we got an XML output that contained the link, latitude and longitude of the documents.
* We have added screenshots of the output we got in another folder.
* A majority of the conspiracies found in the central regions of USA were related to UFOs
* Most of the conspiracies point to locations on either coasts with the central regions being very sparsely populated in terms of conspiracies.  Wyoming doesn't even have a single conspiracy in it.
* Most of the documents from the east coast were based on the Watergate scandal.

# ReadMe

* We have included the JAVA code. We run the code in 3 parts and have commented the 1st 2 parts. The current code runs on the vault and creates indexes for all the files.
* Once the indexes are ready, run the HTML and PHP on Apache Tomcat with mod PHP.
* The Solr must be up and running when you run the PHP code.
* In case of any execution issues please contact us via email or phone. Tel - 2179745588