Memex GeoParser

Mazi Boustani, Madhav Sharan, Chris Mattmann, Lauren Wong
NASA/JPL - USC

About Memex

- Memex is a 3 years DARPA funded project which seeks to develop software that advances online search capabilities far beyond the current state of the art.
- Three technical areas - data gathering, tools development and technology deployment in the field.
- Revolutionize discovery, organization and presentation of search results.

http://memex.jpl.nasa.gov/

What is GeoParser

- One of Memex sub projects, it is open source
- Extract geospatial information from any type of file as well as indexed data
- Visualize extracted information on map
- Search capabilities over textual data

Example: (http://www.marriott.com/hotels/travel/rdumc-raleigh-marriott-city-center)

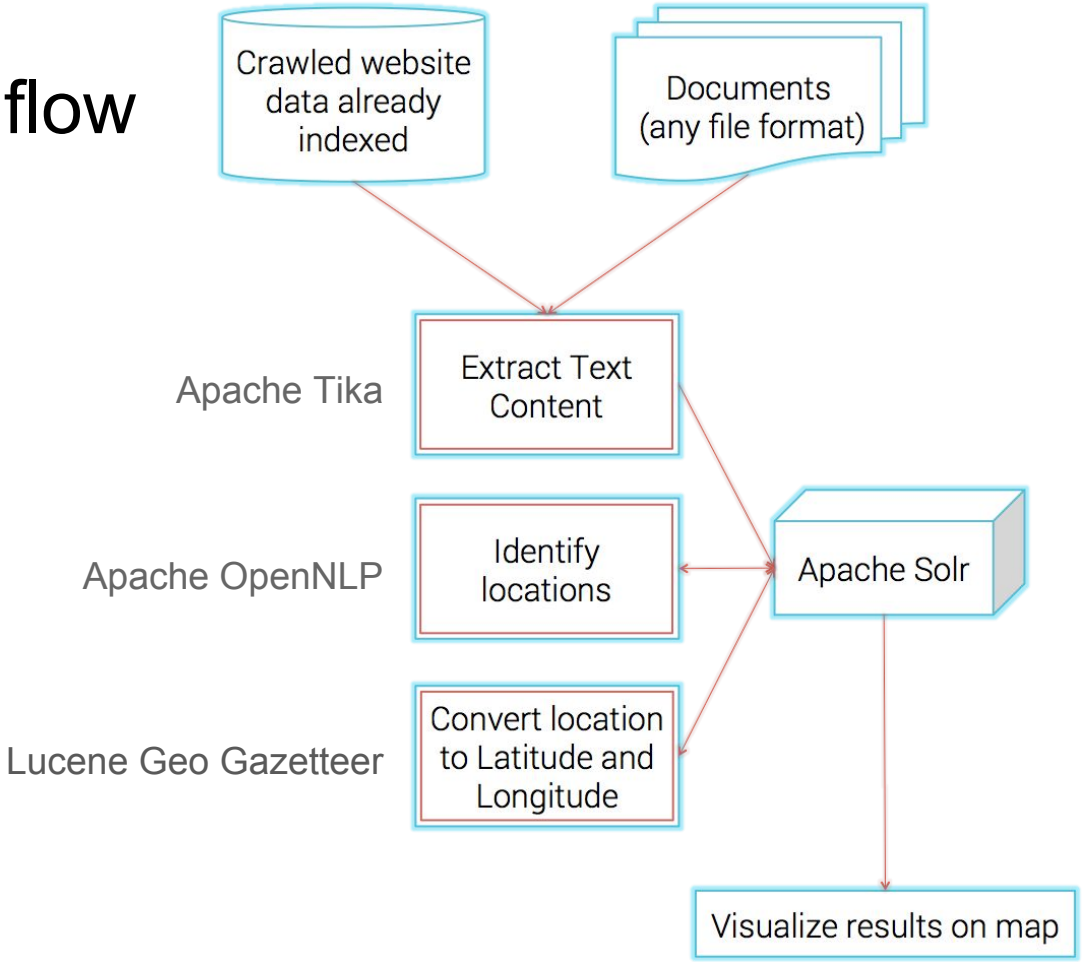
1. Madrid
2. Taiwan
3. NorthCarolina
4. Raleigh
5. HongKongSpecialAdministrativeRegion

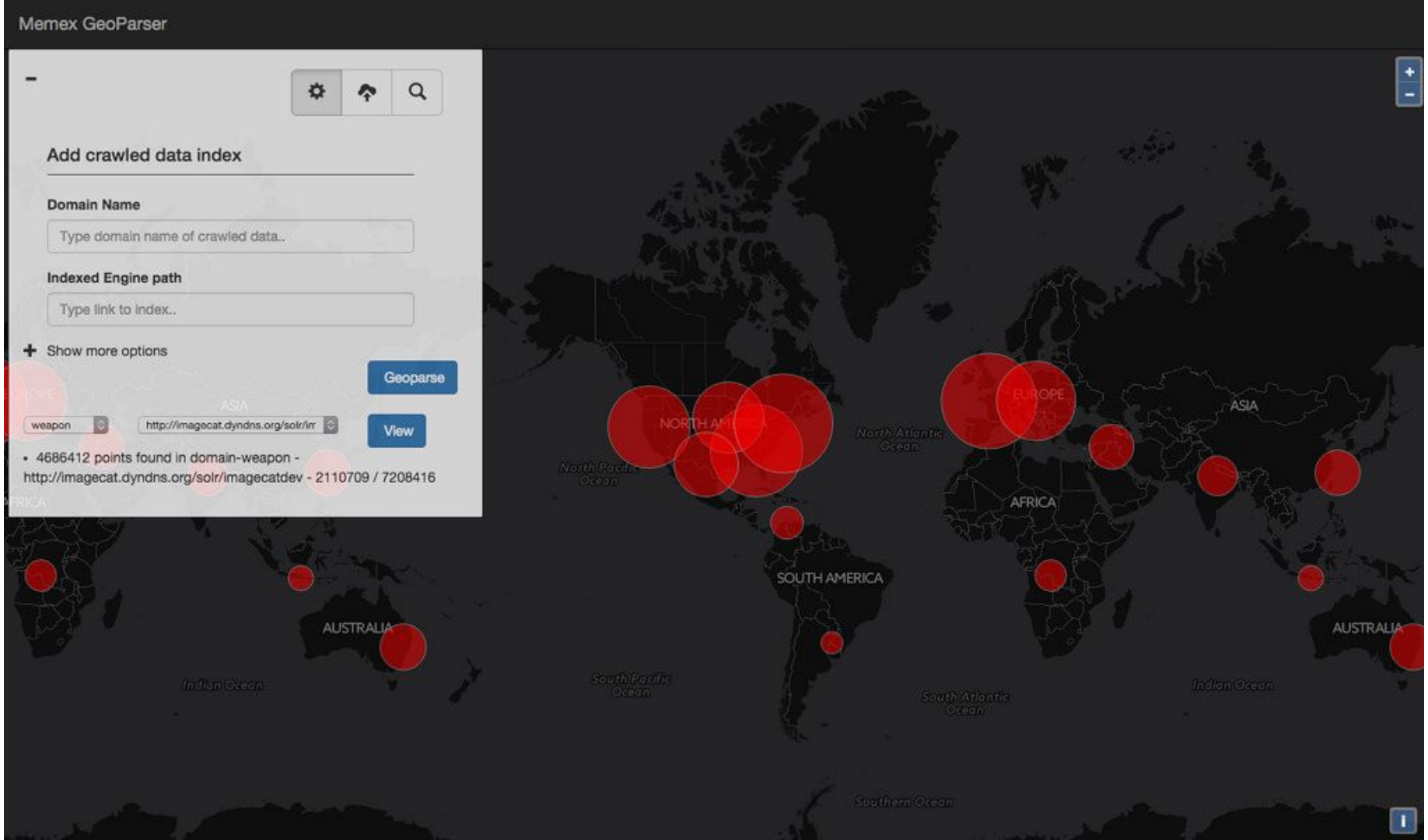
https://github.com/MBoustani/GeoParser

Technologies

- Apache Tika
  - The Apache Tika toolkit detects and extracts metadata and text from over a thousand different file types (such as PPT, XLS, and PDF). All of these file types can be parsed through a single interface, making Tika useful for search engine indexing, content analysis, translation, and much more.
- Khooshe
  - Big GeoSpatial Data Points Visualization Tool by using vector tiles [https://github.com/MBoustani/Khooshe]
- Apache OpenNLP
  - The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text. It supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity recognition, chunking, parsing, and coreference resolution.
- Apache Lucene (Geo Gazetteer)
  - A command line gazetteer built around the Geonames.org dataset, that uses the Apache Lucene library to create a searchable gazetteer.
- Apache Solr
  - Solr is highly reliable, scalable and fault tolerant, providing distributed indexing, replication and load-balanced querying, automated failover and recovery, centralized configuration and more.

GeoParser flow




Memex GeoParser
Add crawled data index
Domain Name
Type domain name of crawled data..
Indexed Engine path
Type link to index..
Show more options
Geoparse
weapon
http://imagecat.dyndns.org/solr/ir
View
4686412 points found in domain-weapon -
http://imagecat.dyndns.org/solr/imagecatdev - 2110709 / 7208416
NORTH AMERICA
EUROPE
ASIA
AFRICA
SOUTH AMERICA
AUSTRALIA
North Atlantic Ocean
North Pacific Ocean
South Pacific Ocean
South Atlantic Ocean
Indian Ocean
Indian Ocean
Southern Ocean

Major Challenges we addressed

- Indexing a solr core of 7 million documents.
  - Speeding up the process to 700 docs per minute.
- Representing 20 million points on a map.
  - Server side clustering using Khooshe
- Finding latitude longitude of a diverse geo locations.
  - Submitted a paper in IRI 2016
  - *“An Automatic Approach for Discovering and Geocoding Locations in Domain-Specific Web Data”*
- Plotting Khooshe layers to OpenLayers 3.

Acknowledgements

- This work was supported by the DARPA XDATA/Memex program.

- NSF Polar Cyberinfrastructure award numbers PLR-1348450 and PLR-144562 funded a portion of the work.

- Effort supported in part by JPL, managed by the California Institute of Technology on behalf of NASA.

References

- GeoParser: https://github.com/MBoustani/GeoParser
- Memex: http://memex.jpl.nasa.gov/
- Apache Tika: https://tika.apache.org/
- Khooshe: https://github.com/MBoustani/Khooshe
- Lucene Geo Gazetteer: https://github.com/chrismattmann/lucene-geo-gazetteer
- Apache OpenNLP: https://opennlp.apache.org/
- Apache Solr: http://lucene.apache.org/solr/