**Report Assignment 2 – Team 11**

**Subject: CSCI 599 - Content Detection and Big Data Analysis**

**Indexing technology used:** Solr

1. **What features did you find most useful in exploring the Polar data?**

We used Tika to parse through the dataset and extract the features that were required for each question and also for general analysis. The features that were most useful were the *title, description, file size, the duration of the media files, author, and publication year.*

1. **Were you able to take advantage of Tag Ratios to isolate the measurement data?**

Yes, we were able to use Tag ratios to extract measurement data. We first ran Tag Ratios and identified the lines which had text which could be read to extract data. Then we ran OpenNLP NER to find out any measurement data that was present.

1. **Did NER and SWEET terminology mapping work well – was the NER unable to identify SWEET categories and concepts?**

Yes, after extracting concepts from SWEET ontology and extracting named entities using

OpenNLP, we find lots of overlaps between these two sets of values which leads us to the

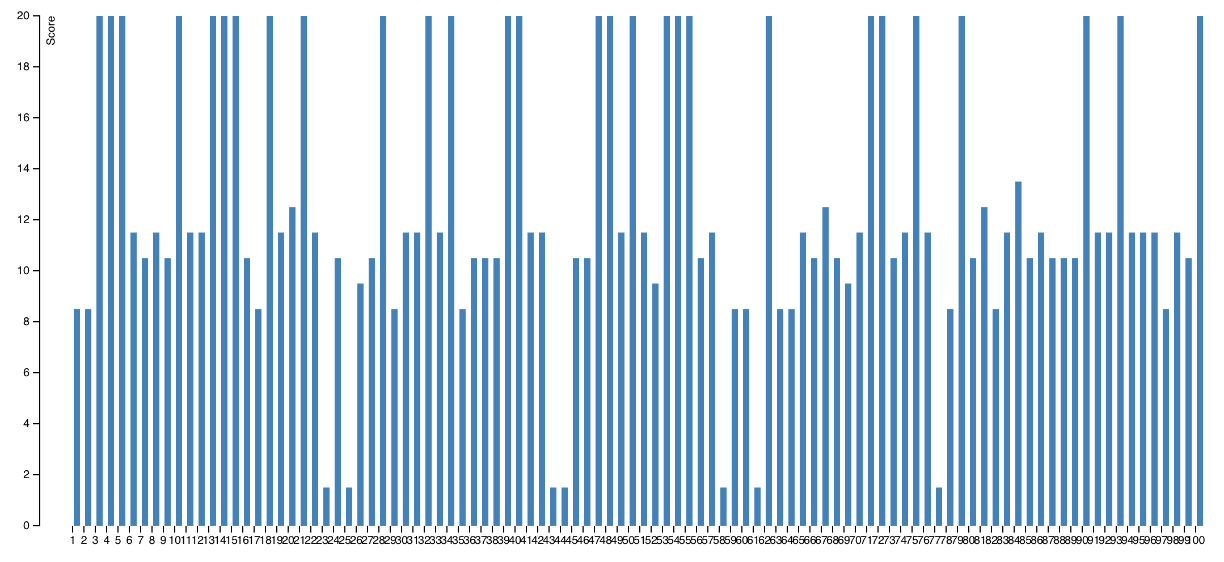
result as an intersection of these two sets.

1. **Did the D3 interactive visualizations help you understand the data?**

Yes, the D3 interactive visualizations helped us understand the behavior of a feature over a range of files. We were able to generate the 8 D3s. The understanding of each is given below:

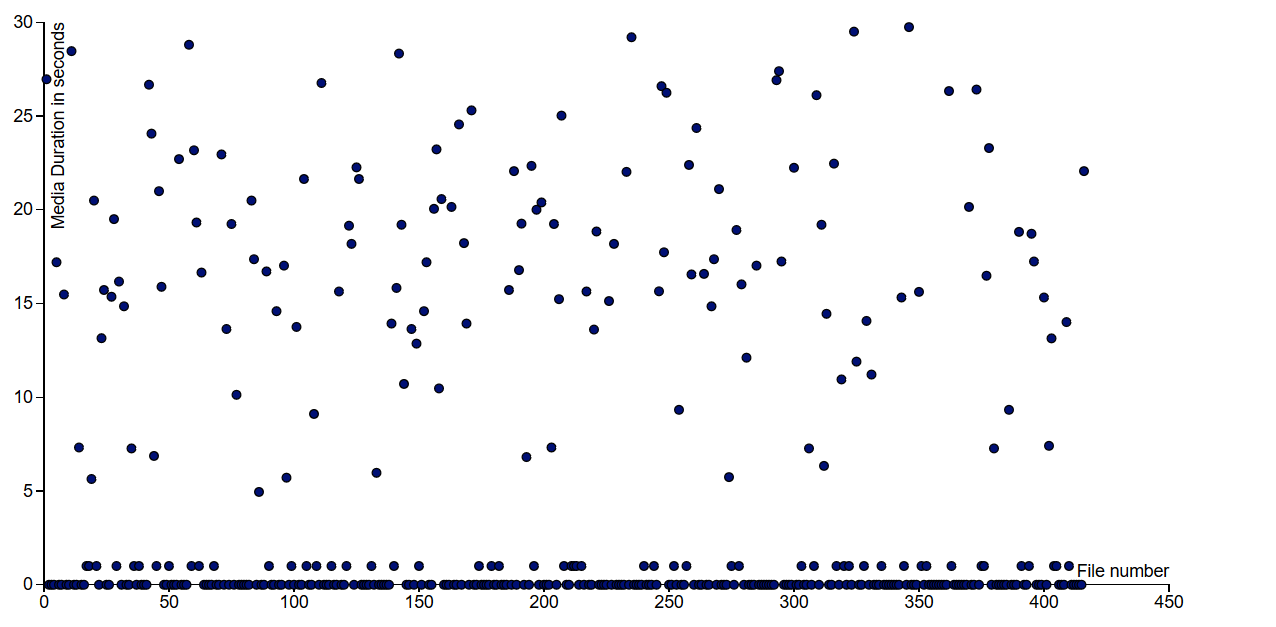
* + **Metadata score**

The bar chart gives the visual representation of metadata score for each file that is present in the solr index. Is is observed that most of the score is between 8 and 12.



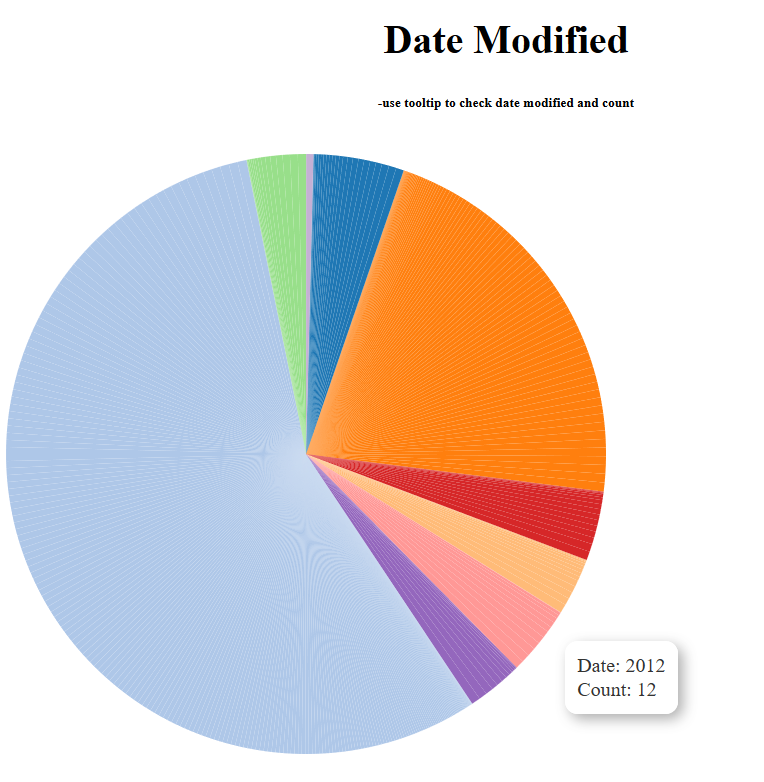
* + **Media duration**

The scatterplot gives a visual representation of the duration of the media file in seconds. This data was generated using the EXIFTool. The media duration ranges from 0 to 30 seconds and a lot of files have a duration of 0 seconds.



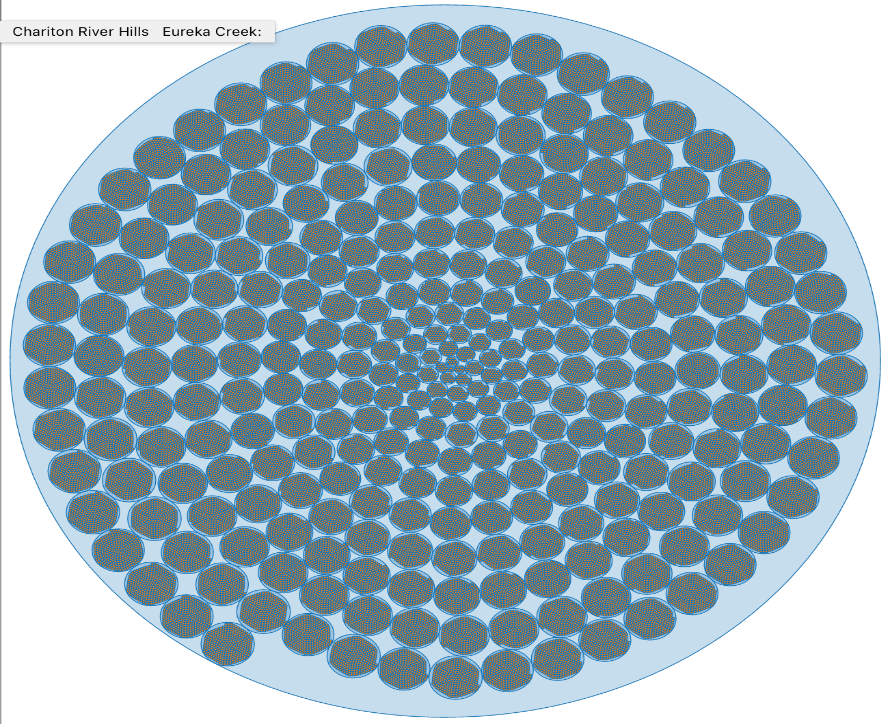
* + **Date modified**

The scatterplot gives a visual representation of the years that the files were modified. This data was generated using the EXIFTool. It is observed that most of the files were modified in 2014.



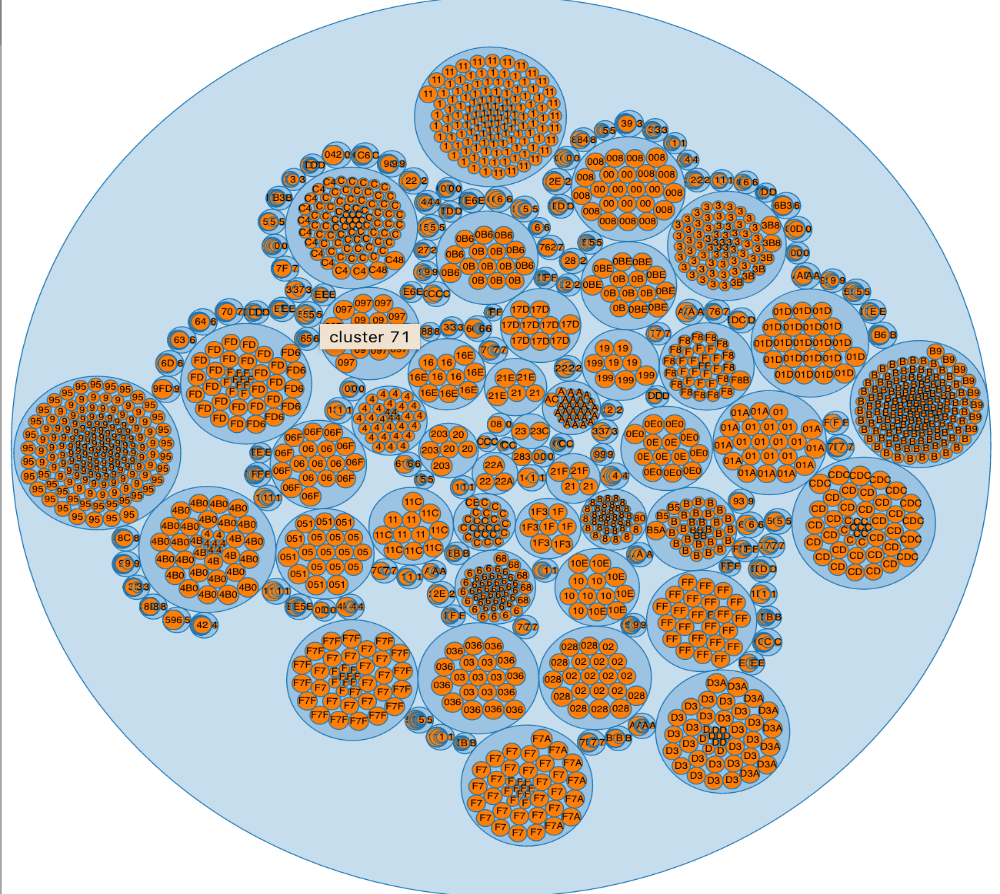
* + **Geo-location**

The visual representation for geolocation was generated by running tika similarity on the indexed dataset.



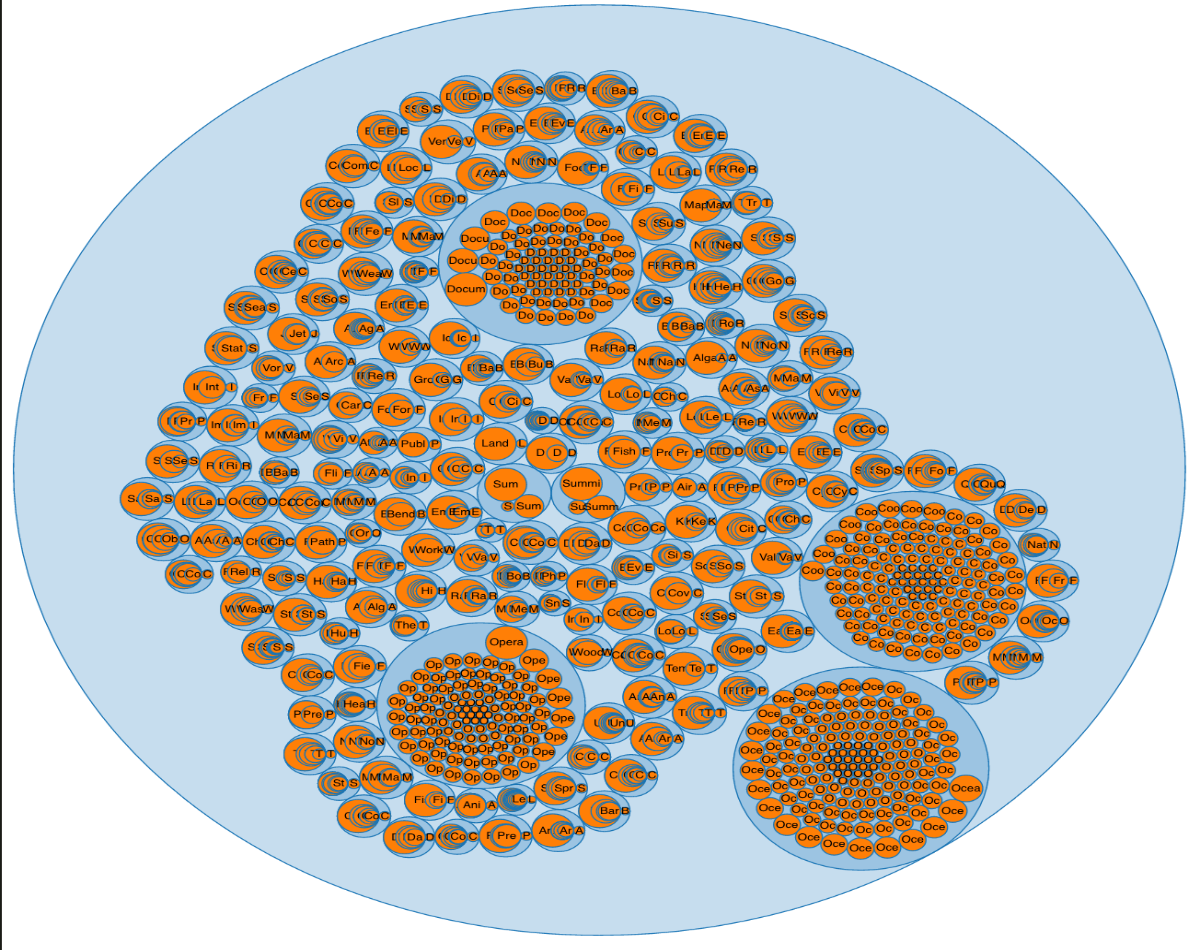
* + **Measurement extraction**

The measurements that were extracted from the dataset after running tag ratios and NER is also represented visually using the circle packing implementation.



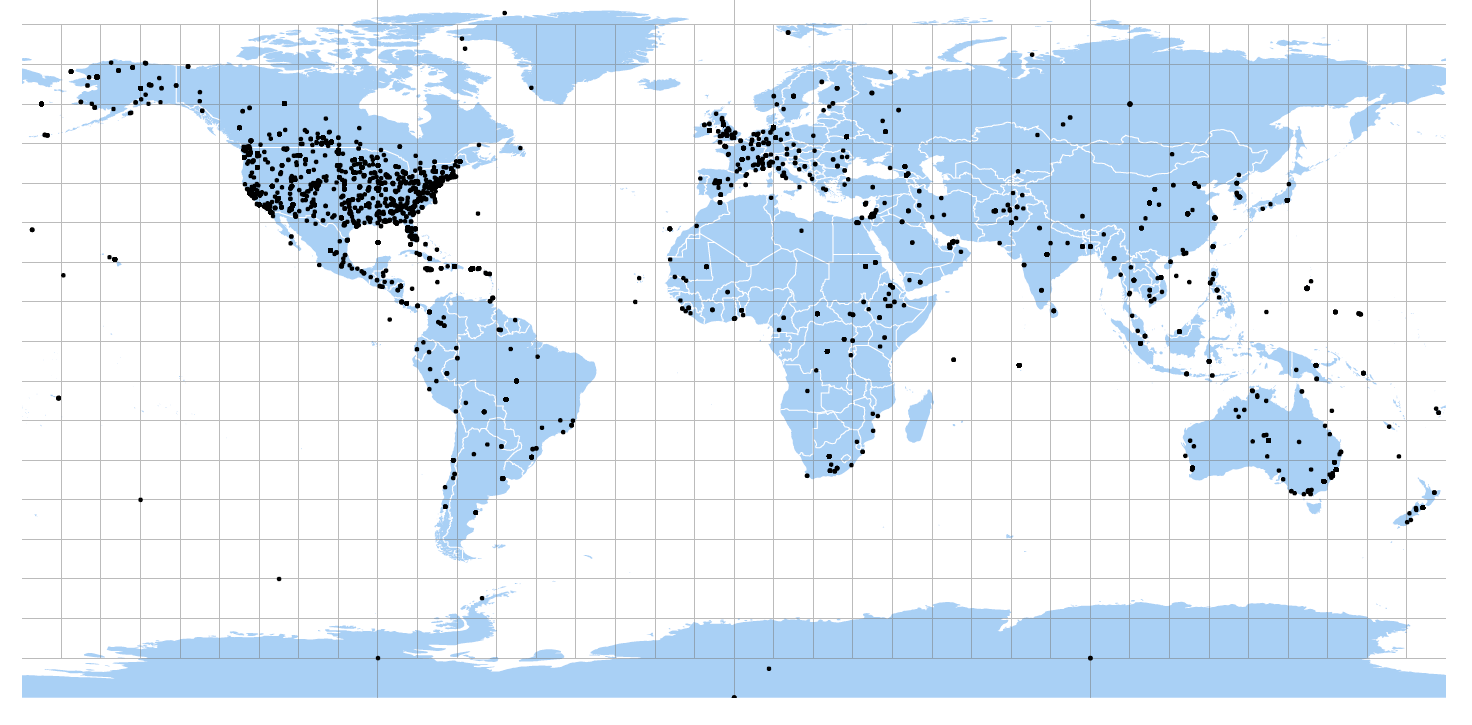
* + **SWEET ontology**

A d3 for the concepts that were extracted after running and a circle packing visualization was created.



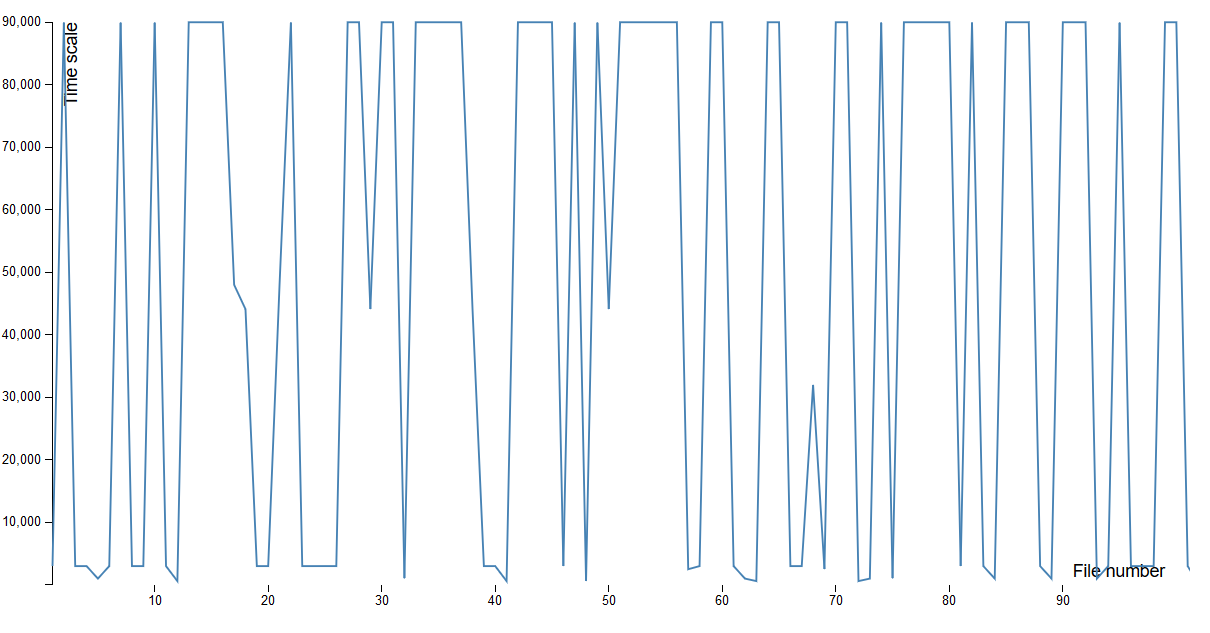
* + **World Map**

With the locations that were extracted we created a world map visualization showing the points of extracted geolocations.



* + **Time scale**

We created a line chart showing the time scale across various files. This data was obtained by running the EXIFTool.



1. **Were particular features that you extracted such as the geo-locations more effective in producing clusters?**

Yes. Longitude and latitude are the most effective features. The extracted locations could be grouped into clusters since they described the actual locations in real world.

1. **Were particular cluster techniques e.g., k-means, more meaningful than hierarchical clustering?**

In general, k-means clustering provides more meaningful insight about the data since the data does not have any particular hierarchical structure. This in turn makes hierarchical clustering less effective.

1. **What about distance metrics – which ones were more effective (Jaccard, Edit Distance, etc.) Why?**

For measurement data and SWEET data, edit distance is more effective than Jaccard. With measurement data, when the values have similar character, they seem to be have some relationship (e.g km and m). This is also applied SWEET data. Therefore, by using edit distance we can capture these characteristic.

For extracted locations and publications, cosine similarity is the most effective measure because we have many fields’ different values and trying to capture the similarity across field by cosine similarity is more effective.

1. **Was your metadata quality score something that you could leverage to find richly curated records and ultimately is it something that could be leveraged to point users to the more meaningful polar data?**

Yes. According to the D3 graph generated by metadata quality score, we can identify high value ranges of score which are usually located in specific kinds of data. We believe that the files with higher metadata score will have more information than the ones with less score. So priority can be given to these files first.

1. **Were you able to find related scientific publications, and did the authors you found both inside the dataset and using Google Scholar have a high degree of overlap with the existing Polar dataset?**

Yes. But we were not able to find it for all the authors because we got banned few times even after putting a 10 second time delay. For each publication of author, we check its other authors are there in our authors.txt file. For some publication, we found all its author in our authors list (co-authors). But for the most part, we were not able to find overlapping authors between polar-dataset and google scholar journal result.

1. **Why did you chose the Content Extractions?**

We chose the EXIFTool to do the Content Extractions. This tool gave us better detail about the media files and we were able to understand the different features like file size, time scale, media duration in seconds, date the file was modified etc.

We chose the EXIFTool because after trying out the other tools, this was the tool that we were able to install and run successfully.

1. **What additional knowledge did you gain from the features?**

We noticed that for some of the files there were as many as 66 metadata features while the others had as less as 10 metadata features. We developed 2 D3s using the data generated by this tool. The scatterplot gave details about the media duration of the file in seconds. We noticed that the duration ranged from 0 to 30 seconds and there were a lot of files with duration of 0 seconds. The pie chart gave details about the year that the files were modified. We noticed that a major portion of the files were modified in 2014 and there were very few files that were modified in 2007.

**Schema for Solr to index the data**

Measurement data:

* Name: String – File name where we extracted the measurement value
* Measurement: String – Value of extracted measurement

Scholar data:

* Name: String – Name of publication
* Author: String – Name of author
* URL: String – URL of publication
* Year: String – Published time of publication

Location data:

* name: Name of the location
* location.latitude: latitude of location
* location.longitude: longitude of location

Metadata score data:

* Name: Name of the file
* Score: metadata score of the file

SWEET data:

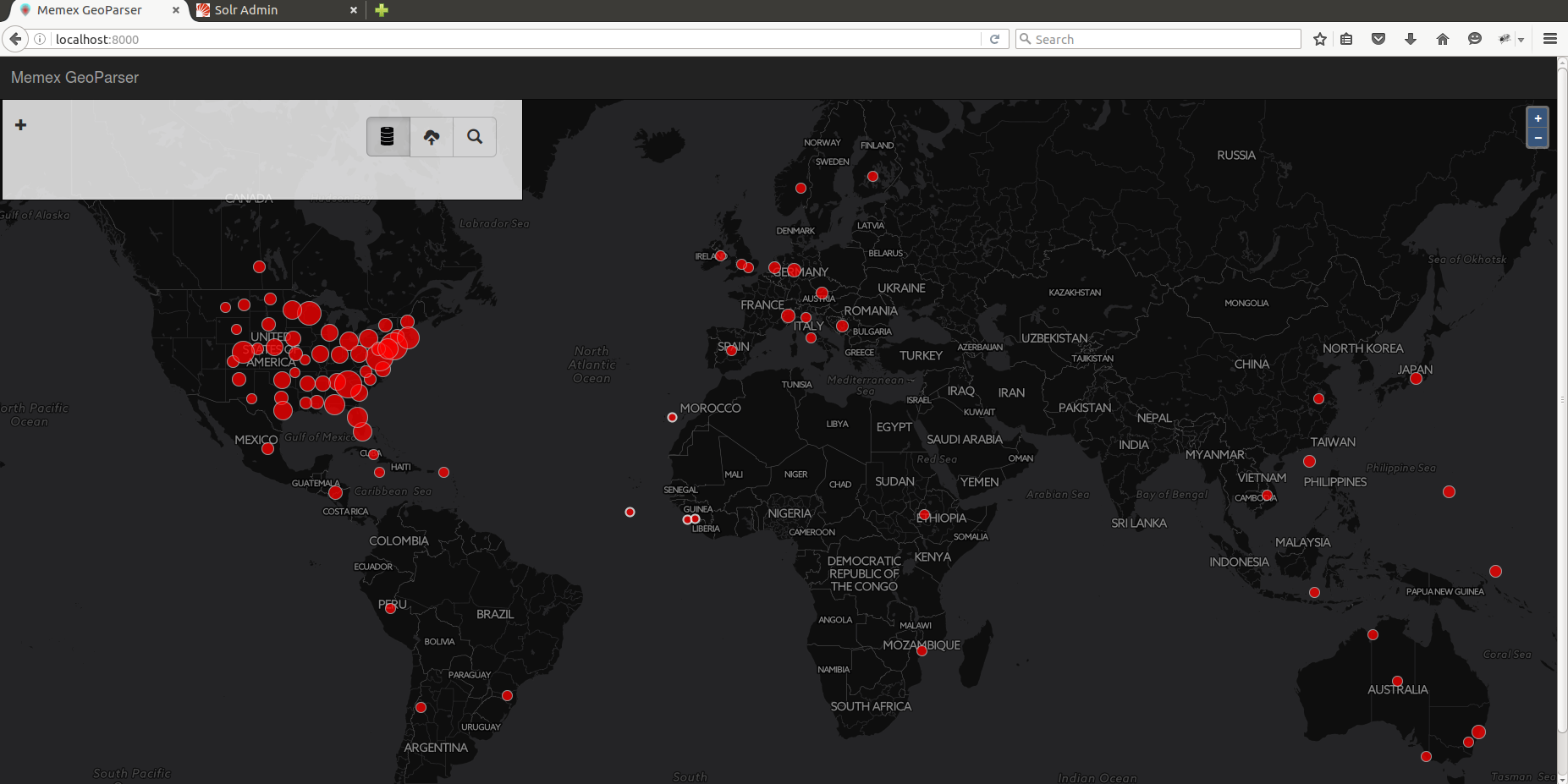
* Concept: Value of the concept

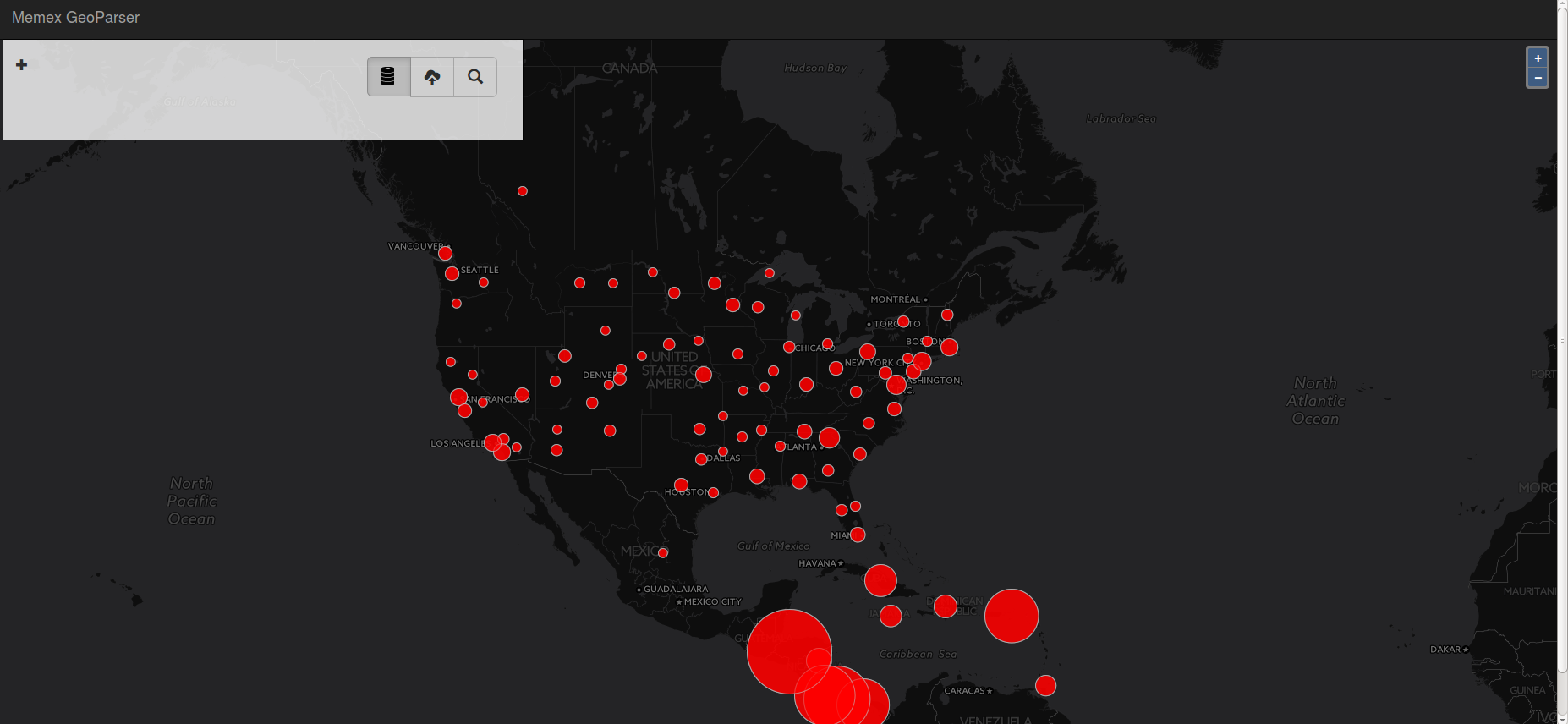
Clustering data (clustering results to be indexed):

* Name: name of the cluster
* Parent: list of parent clusters
* Size: value of similarity
* Type: data type of clusters
  + geo for location data
  + ms for measurement data
  + sweet for SWEET data

**Memex Geoparser**

The output from running the Geoparser is a location map which looks as shown below:





**Tika similarity:**

We modified the clustering code of tika-similarity to cover the bug that the code does not work when csv file shaves the whitespace in values because currently it consider whitespace as delimiters. The code of question 11 contains more details.