ASSIGNMENT: 2

CSCI599: CONTENT DETECTION AND ANALYSIS FOR BIG DATA

TEAM 4

In this assignment, we have worked on enriching the polar dataset by running various parsers and tools to extract more meaningful information from the dataset. We have performed a series of algorithms ranging from Tag ratio to extract the measurements, Grobid parser to extract the publication details, GeoTopic parser to extract geolocation information, discovered the SWEET ontologies being referred to, calculated metadata quality score, created Solr index, cluster the data based on different criteria's and express them in the form of D3 visualizations.

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

While performing NER technique on the dataset we came across the many SWEET ontologies which were being referred to in the various files in the full crawl. We also came across various other NER entities but they were not all relevant.

With the help of Geotopic parser, we could extract the various different geographical locations mentioned in the files across the dataset with their exact latitude and longitude co-ordinates along with their geographical names.

We also found co-relation between the files present in the dataset based on the "related" feature with the help of Google scholar API.

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

Yes, it helped. If we try to read the whole file to isolate the measurement data, we see that there are many downfalls such as increase in execution time and encountering irrelevant data. Tag ratio focuses on relevant data thereby reducing computing time.

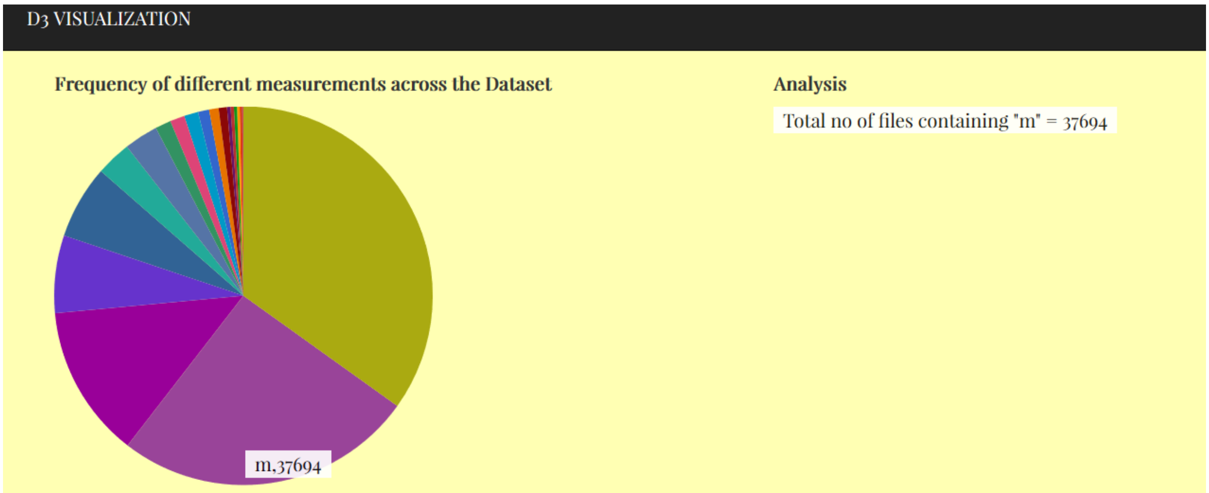
Did NER and SWEET terminology mapping work well – was the NER unable to identify SWEET categories and concepts?

We extracted the concepts from the SWEET ontology by the OWL language parser. There are around 5700 concepts of different classes and subclasses in the SWEET ontology. We then executed NER on the polar dataset for around 2, 00, 000, files. According to our analysis, only a small fraction(10%) of NER entities were part of the SWEET concepts and at the same time there were many SWEET concepts which were not recognized by NER. We also found that there were many NER entities which were not SWEET concepts. Hence, we can say that NER did not map perfectly with SWEET terminology.

Did the D3 interactive visualizations help you understand the data?

Yes, the D3 interactive visualizations was helpful to understand the data to some extent. We found out about the different entities which were used and being referred to in the data as it helped us know more about the percentage of files based on a particular entity such as geological locations, measurements, publications, authors, ontologies etc. This gave us the ability to access certain features of the data and comprehend it in a much better way.

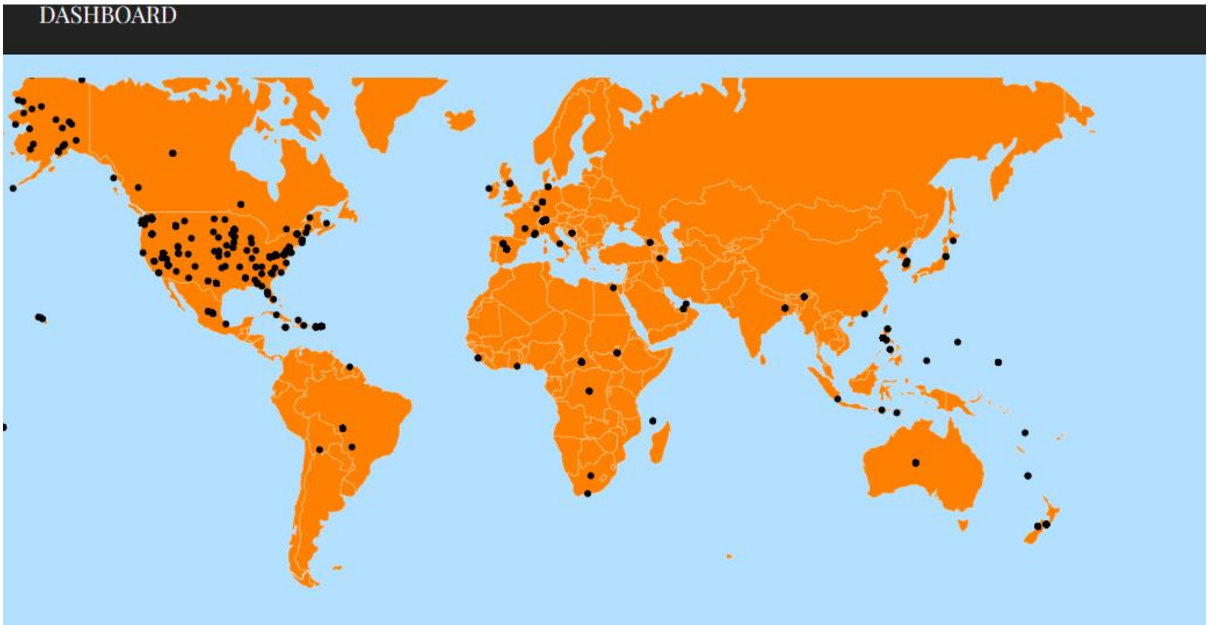
We created the following D3 visualisations:



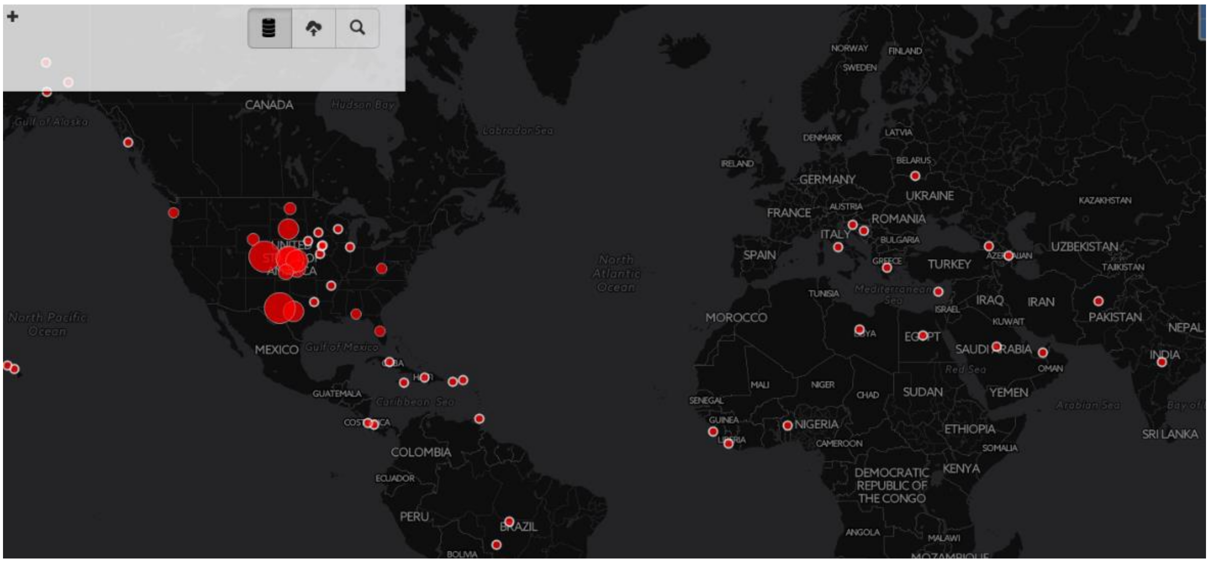
Analysis

Total no of files containing "m" = 37694

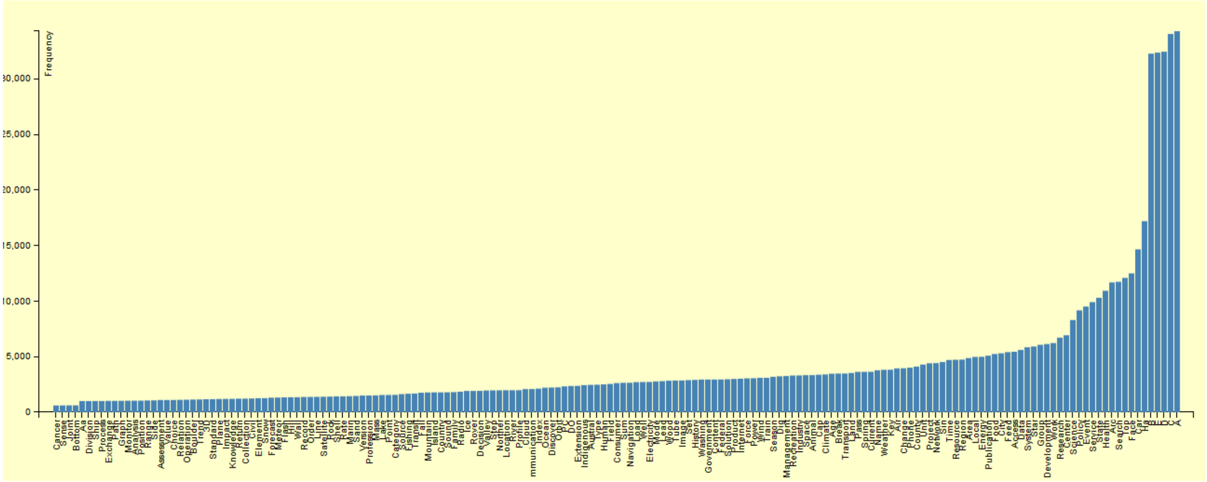
We ran Tag Ratio algorithms and NER technique on the dataset to find the different measurement values used in the files. We found out that measurements such as centimetre, pound, seconds, hours, minutes, metres, inches etc. were used extensively across the dataset.



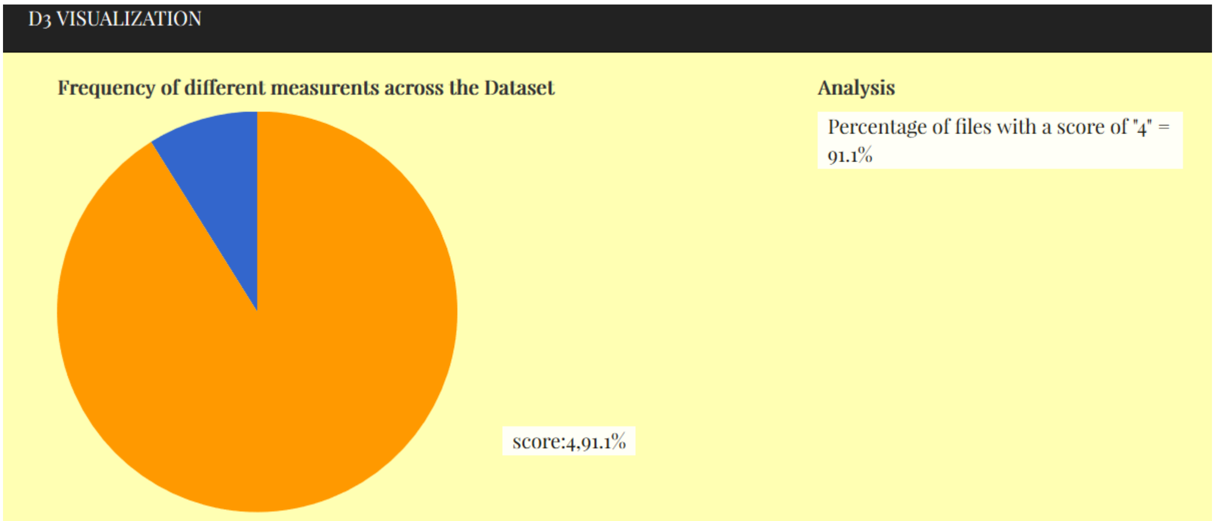
The above D3 shows the different geological locations which were extracted using the GeoTopic parser across the dataset. The parser gave us the top two most frequently used co-ordinates.



This is the visualisation provided by the MEMEX geoparser highlighting the geographical locations present in the files.



We compared the dataset against the SWEET ontologies and found out the different ontologies used throughout the dataset. Here is a snippet showing the frequencies of the different ontologies used across the dataset.



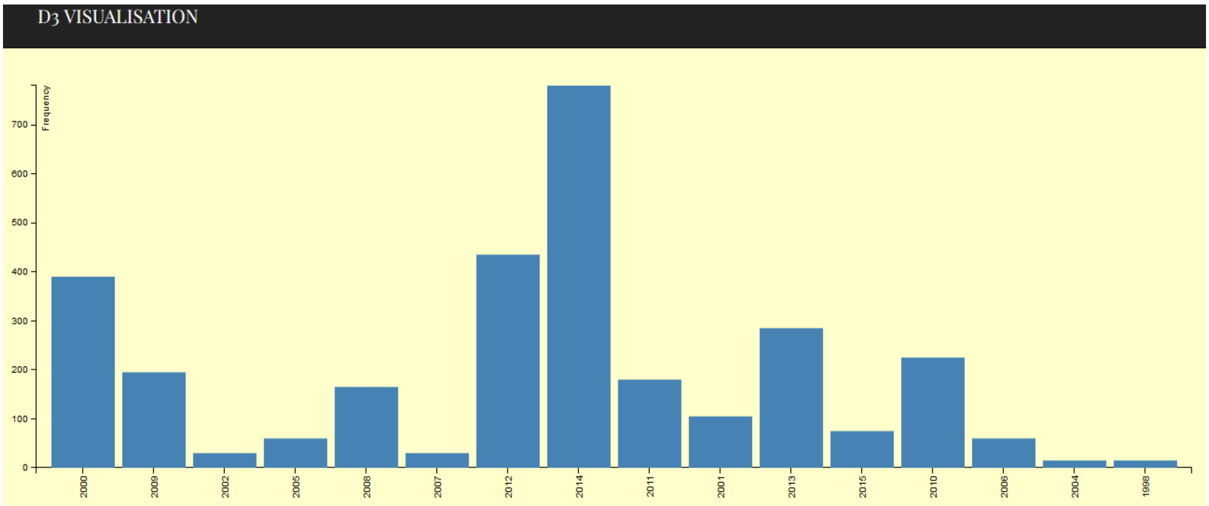
D3 VISUALIZATION

Frequency of different measurents across the Dataset

Analysis

Percentage of files with a score of "4" = 91.1%

We calculated a Metadata quality score for every file based on certain criteria's to determine the richness of the data. We found 90% of the data is very rich with its metadata content overall.



D3 VISUALISATION

We used the Grobid Journal Parser to perform content extraction and found out the different publications present in the dataset. Here we have the visualisation of the no. of publications in a given year.

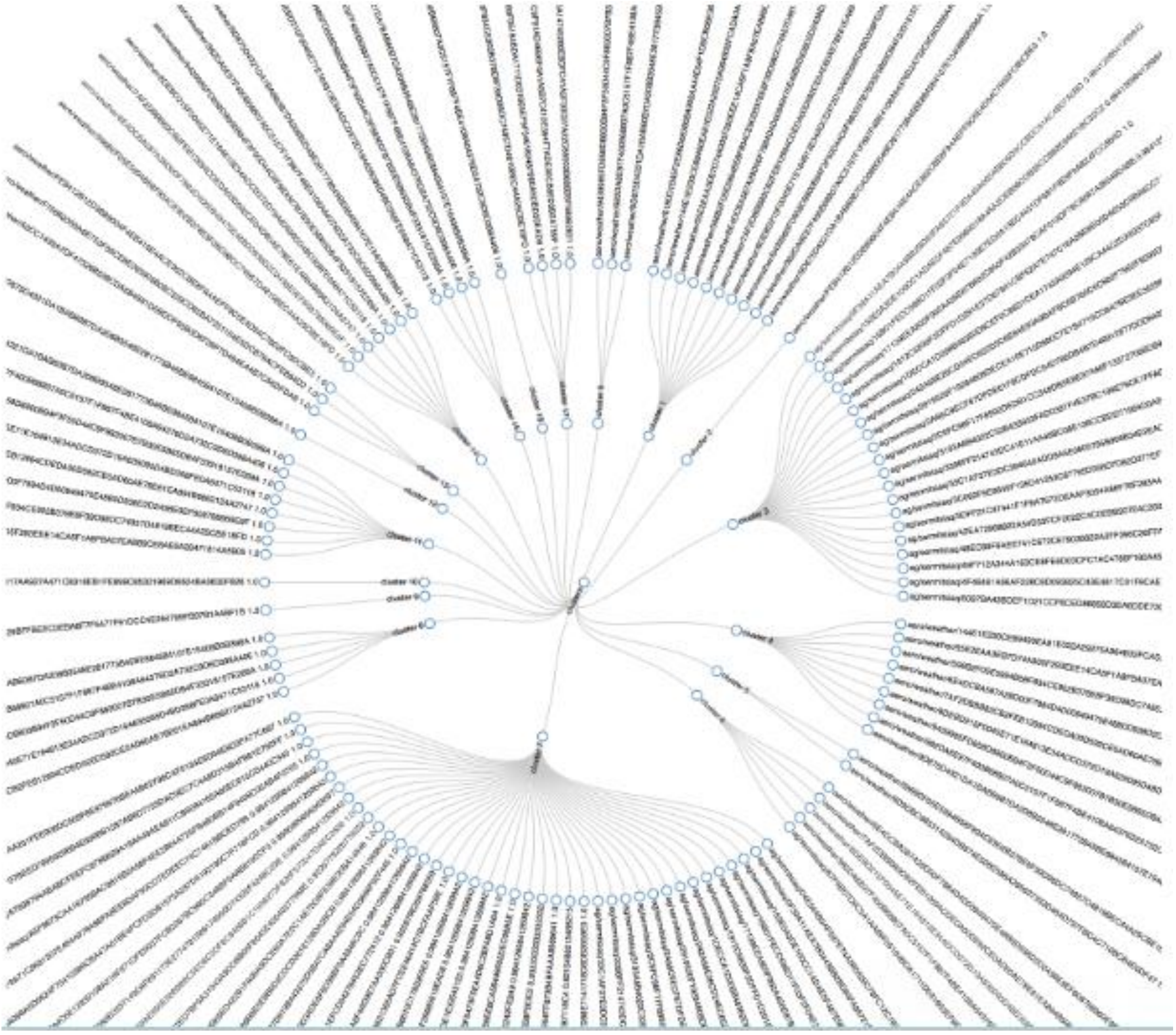
Were particular features that you extracted such as the geo-locations more effective in producing clusters?

Yes, Geo-Locations proved to be an effective measure to form clusters as we can clearly visualize the points to which a particular file is pointing to with the help of geotopic parser. GeoTopic parser takes into account various types of files such as text/html, application/x-html, application/pdf and extract geographic measures like 'LATITUDE & LONGITUDE' of most frequent and second most frequent location in the file. Since locations give the exact dimensions clustering becomes more accurate as there is a measure to calculate similarity between documents. While other features like author, publications, ontology give less sound measures to deduce similarity.

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

Yes, based on our analysis we comprehended that hierarchical clustering proved to be more meaningful than particular cluster techniques. This is because, when we clustered the files based on ontologies concept, the classification did not work effectively. On further analysis, we found measurement to be the next metric in the hierarchical clustering. Similarly, authors proved to be the next step in the hierarchy. And finally, we used geographical measurements to give accurate classifications.

Following shows the output of Tika-similarity, which clearly shows the different clusters of files which are similar to each other.



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

The distance metric to calculate the distance between two entities was specific to the entity to be compared for eg, to calculate the distance between two geographical locations we used the Manhattan distance and to calculate distance between two measurements we found the Edit distance between each measurement in the first file against every measurement in the second file.

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?

We calculated individual scores for the files across the polar dataset based on their respective metadata. Different criteria's were considered in the calculation of the score such as

- how well the metadata describes its intended purpose,
- If it uses any standard controlled vocabularies such as,
    - Dublin Core
    - LOM(Linked Open Vocabularies)
    - CP
    - XMP
- If the file is licensed to identify the terms and conditions,
- whether there are any aliases to resolve the naming conflicts,
- If object then if DOI is generated for the same
- and finally if it allows for long term management of objects in collections using indexing.

This score has helped us understand the diversity and richness of the polar data. The higher the quality score the richer the data. According to our analysis, most of the polar data is deemed rich because of its higher score.

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, we were able to find related publications as we gave both "author" and first 12 characters of "Title" to scholar.py and set the number of records to be returned as 20.

We extracted Title, Citations list, Versions list, Year and Excerpt from the publications related to author and phrase. However, there was a small challenge while finding the author of related publications as scholar.py doesn't return the author of the same but we managed to do it by fetching the source code of URL pointed out by "Versions list" and extracted author's name from tag **"<div class=\"gs_a\">(.*)</div>"** .We used User agent to go to required URL as google blocks automated agent. There was still a caveat while running scholar.py because google allows only a fixed number of requests from a given IP. Still the authors we found from related publications and authors of publications from Polar DataSet don't overlap much due to the fact that some author we got are less famous with very few number of publications under them.

EXTRA CREDIT

Why did you chose the Content Extractions?

From the TIKA library of content extractions, we executed FFMPEG content extraction on the full polar dataset. In our analysis of the polar data in this assignment, we were only able to explore the content of text documents and not audio and video. Hence, we chose the FFMPEG content extraction to know more about the audio and video files.

What additional knowledge did you gain from the features?

The FFMPEG tool identified many metadata fields such as creator, artist, album, genre, compressor, release date etc. which was not identified earlier by the Tika parser.