Assignment 2 : Report on Scientific Content Enrichment in the Text Retrieval Conference (TREC) Polar Dynamic Domain Dataset

Team : 18 (Siddharth Bhayani [sbhayani@usc.edu], Nimesh Jain[nimeshja@usc.edu] , Suhas Suresh[suhassur@usc.edu])

USEFUL FEATURES IN POLAR DATASET

Following were the most useful features :

1. Geo Locations information extracted from the GeoParser.
2. Measurements information extracted using NER technique.
3. Data related to publications extracted from GROBID parser.

TAG-RATIO ALGORITHM : IMPLEMENTATION & OBSERVATION / INFERENCES

- We implemented TagRatio algorithm as a separate parser in Tika. Also before parsing any document we parsed the incoming document with Tika's AutoDetectParser and used it's XHTML Output as input to our TagRatio Parser.
- Due to this, we were parse document with any mimetype and extract information from it.
- Since TagRatio parsed output is only the actual content , it was easy and quick to extract measurement from it.
- We used https://github.com/ldidry/lstu to generate the unique id (tiny url) for each document.

SWEET CONCEPTS EXTRACTION:

We extracted only weather related SWEET concepts. Extracted around 422 concepts and applied Tika NER Regex over the polar dataset to map each file entity to the sweet concept.

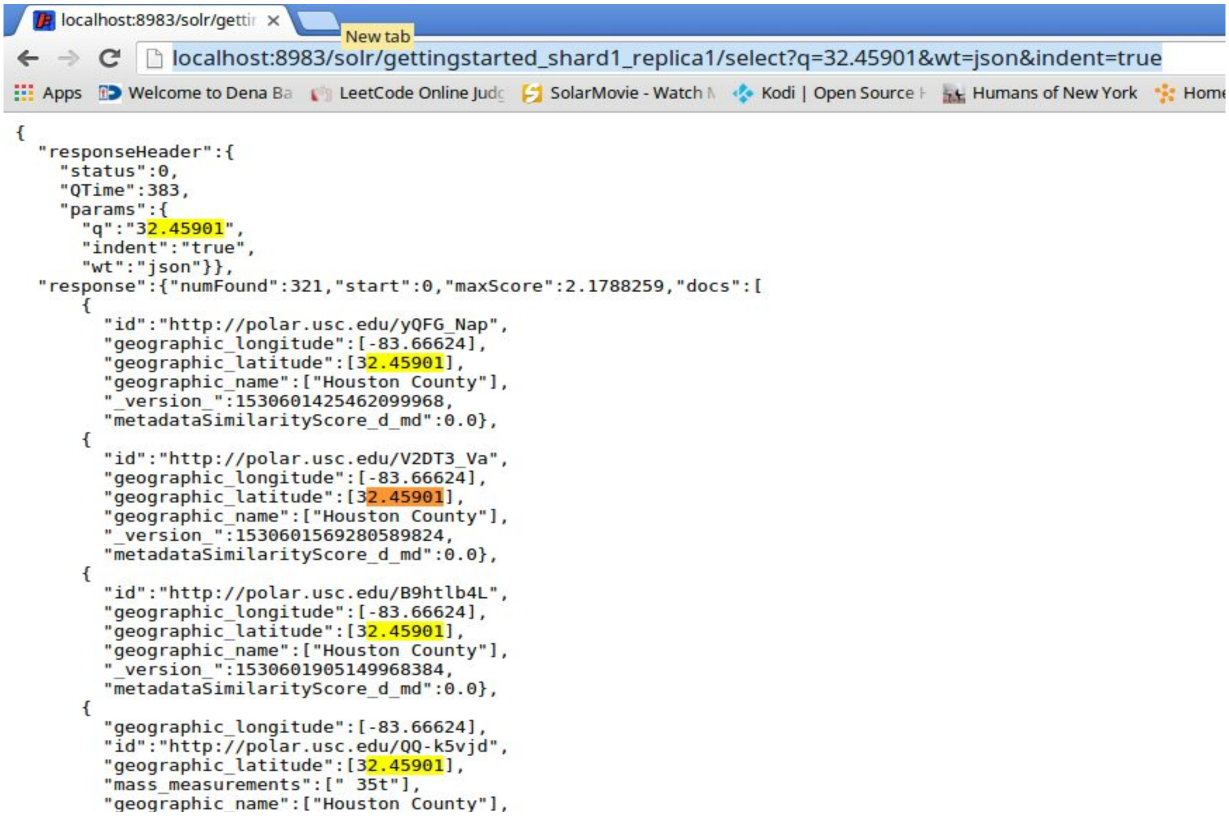
```
        [
        {
                "id":"http://polar.usc.edu/Jm94xwYp",
                "sweet_concepts":[
                        "drop",
                        "lead",
                        "rime"
                ]
        }
]
```

INGESTING METADATA :

We used Apache Solr to index the data . We indexed following fields :

Field	**Description**
id	Unique Identifier for each file.
geographic_name	Geographic Name identified by GeoTopic Parser.
geographic_latitude	Latitude corresponding to geographic_name
geographic_longitude	Longitude corresponding to geographic_name
optional_name1	Optional Geographic Name Identified by GeoTopic Parser
optional_latitude1	Latitude corresponding to optional_name1
optional_longitude1	Longitude corresponding to optional_name1
optional_name2	Another Optional Geographic Name Identified by GeoTopic Parser
optional_latitude2	Latitude corresponding to optional_name2
optional_longitude2	Longitude corresponding to optional_name2
sweet_concepts	An array of string holding extracted SWEET Concepts.
temperature_measurements	An array of string holding extracted temperature measures. Eg. 5°F
mass_measurements	An array of string holding extracted mass measures. Eg. 50kg
time_measurements	An array of String holding extracted temporal information. Eg. 1998 year
length_measurements	An array of Strings holding extracted length measurements. Eg : 150 miles, 20 km.
author	Name of the Author
citations	Corresponding Citation of Author
journal_url	Url of Journal
journal_cluster	Cluster ID

Below is sample query output snapshot :


localhost:8983/solr/gettir ×
New tab
localhost:8983/solr/gettingstarted_shard1_replica1/select?q=32.45901&wt=json&indent=true
Apps
Welcome to Dena Ba
LeetCode Online Judg
SolarMovie - Watch M
Kodi | Open Source H
Humans of New York
Home

```
{
  "responseHeader":{
    "status":0,
    "QTime":383,
    "params":{
      "q":"32.45901",
      "indent":"true",
      "wt":"json"}},
  "response":{"numFound":321,"start":0,"maxScore":2.1788259,"docs":[
      {
        "id":"http://polar.usc.edu/yQFG_Nap",
        "geographic_longitude":[-83.66624],
        "geographic_latitude":[32.45901],
        "geographic_name":["Houston County"],
        "_version_":1530601425462099968,
        "metadataSimilarityScore_d_md":0.0},
      {
        "id":"http://polar.usc.edu/V2DT3_Va",
        "geographic_longitude":[-83.66624],
        "geographic_latitude":[32.45901],
        "geographic_name":["Houston County"],
        "_version_":1530601569280589824,
        "metadataSimilarityScore_d_md":0.0},
      {
        "id":"http://polar.usc.edu/B9htlb4L",
        "geographic_longitude":[-83.66624],
        "geographic_latitude":[32.45901],
        "geographic_name":["Houston County"],
        "_version_":1530601905149968384,
        "metadataSimilarityScore_d_md":0.0},
      {
        "geographic_longitude":[-83.66624],
        "id":"http://polar.usc.edu/QQ-k5vjd",
        "geographic_latitude":[32.45901],
        "mass_measurements":[" 35t"],
        "geographic_name":["Houston County"],
```

SIMILARITY and CLUSTERING :

Tika Similarity uses metadata to compare two files and get the similarity score. However, this would have been a wrong measure for us if we ran Tika Similarity as it is because the input to the similarity are the json metadata files. Since each file has same structure and extension, their metadata values would also be same and similarity scores will be misguiding in deriving inferences. Thus, the clustering and distance metric is done on the content rather than metadata. We actually modified the Tika Similarity code to compare files on content rather than metadata.

We applied below four distance measuring and clustering algorithms:

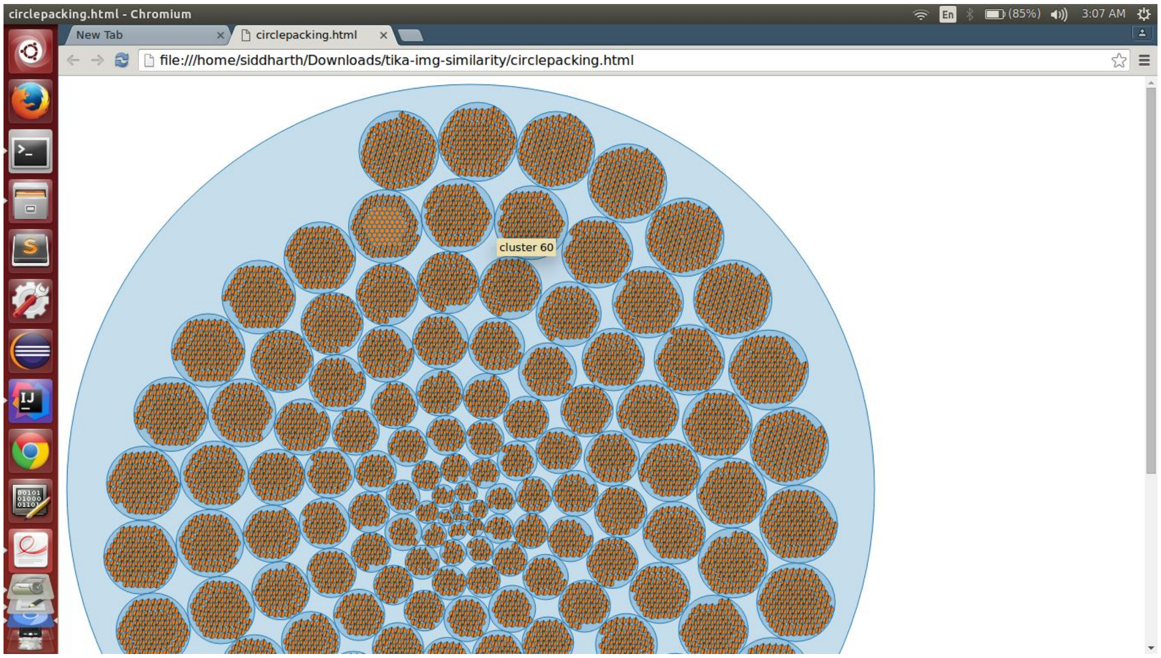
1) Edit Distance
2) Jaccard
3) Cosine
4) K means

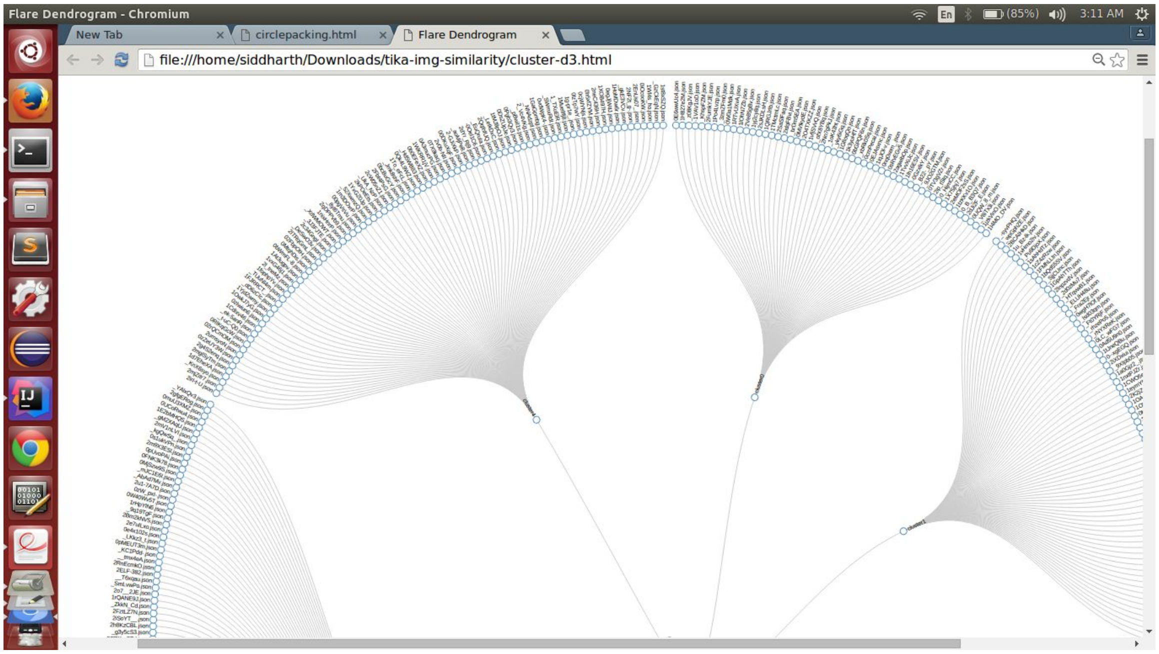
We ran each algorithm on a small subset of 500 documents and used D3 graphs generated to understand the data.

Our views/observations on tika-similarity D3 and inferences :

1. From the kmeans , we inferred that clusters produced with measurement were more meaningful than clusters produced with geolocations.
2. D3 visualizations produced for cosine and jaccard similarities are very large and it's very difficult to infer something from them.

Here are the snapshots of visualizations for clustering & similarity :


circlepacking.html - Chromium
(85%)
3:07 AM
New Tab
circlepacking.html
file:///home/siddharth/Downloads/tika-img-similarity/circlepacking.html
cluster 60


Flare Dendrogram - Chromium
(85%)
3:11 AM
New Tab
circlepacking.html
Flare Dendrogram
file:///home/siddharth/Downloads/tika-img-similarity/cluster-d3.html

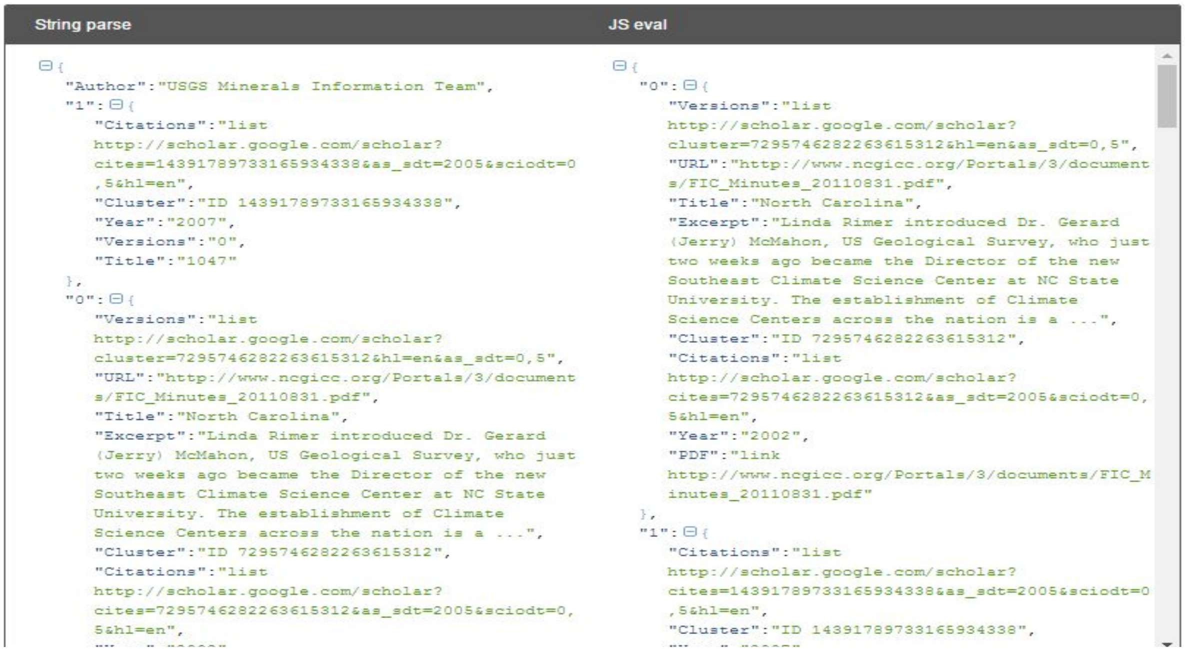
OBSERVATION REGARDING GROBID and GOOGLE SCHOLAR API :

Installed the Grobid and executed with tika-app and tika-server and verified the results. Execute the Grobid parser for PDF files in the polar data set extracted the “Author” from the json output and queried the Google Scholar API with 20 related document of that Author.
Observations: Several files around 1000 had related files in polar dataset and the results by Google Scholar. There were some Authors extracted from Polar Dataset which did not have any publication as well while there were some Authors in Polar which had less than 20 publication.

Executed on Approximately : **15000** PDFs and found the related pdfs of around 960 related publications. There were issues that the google blocked the IP address of the machine for the repeated queries which caused some publications in Polar to not match the publications found out by Google Scholar.

There were several publications in Polar Set which did not have Author while the TEI Annotations had. Also the Tika Parser did not retrieve the publications field for majority of the json output it extracted for pdfs.

Json Output:



String parse

```
⊖{
   "Author":"USGS Minerals Information Team",
   "1":⊖{
      "Citations":"list
      http://scholar.google.com/scholar?
      cites=14391789733165934338&as_sdt=2005&sciodt=0
      ,5&hl=en",
      "Cluster":"ID 14391789733165934338",
      "Year":"2007",
      "Versions":"0",
      "Title":"1047"
   },
   "0":⊖{
      "Versions":"list
      http://scholar.google.com/scholar?
      cluster=7295746282263615312&hl=en&as_sdt=0,5",
      "URL":"http://www.ncgicc.org/Portals/3/document
      s/FIC_Minutes_20110831.pdf",
      "Title":"North Carolina",
      "Excerpt":"Linda Rimer introduced Dr. Gerard
      (Jerry) McMahon, US Geological Survey, who just
      two weeks ago became the Director of the new
      Southeast Climate Science Center at NC State
      University. The establishment of Climate
      Science Centers across the nation is a ...",
      "Cluster":"ID 7295746282263615312",
      "Citations":"list
      http://scholar.google.com/scholar?
      cites=7295746282263615312&as_sdt=2005&sciodt=0,
      5&hl=en",
```

JS eval

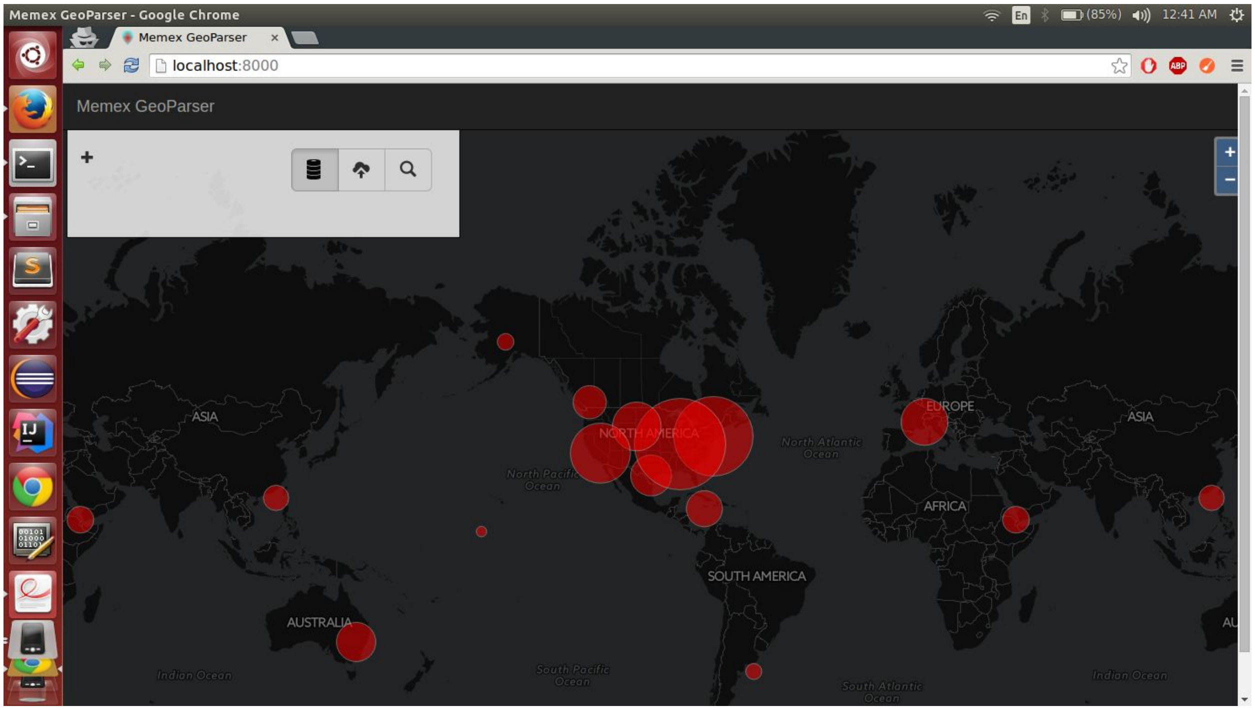
```
⊖{
   "0":⊖{
      "Versions":"list
      http://scholar.google.com/scholar?
      cluster=7295746282263615312&hl=en&as_sdt=0,5",
      "URL":"http://www.ncgicc.org/Portals/3/document
      s/FIC_Minutes_20110831.pdf",
      "Title":"North Carolina",
      "Excerpt":"Linda Rimer introduced Dr. Gerard
      (Jerry) McMahon, US Geological Survey, who just
      two weeks ago became the Director of the new
      Southeast Climate Science Center at NC State
      University. The establishment of Climate
      Science Centers across the nation is a ...",
      "Cluster":"ID 7295746282263615312",
      "Citations":"list
      http://scholar.google.com/scholar?
      cites=7295746282263615312&as_sdt=2005&sciodt=0,
      5&hl=en",
      "Year":"2002",
      "PDF":"link
      http://www.ncgicc.org/Portals/3/documents/FIC_M
      inutes_20110831.pdf"
   },
   "1":⊖{
      "Citations":"list
      http://scholar.google.com/scholar?
      cites=14391789733165934338&as_sdt=2005&sciodt=0
      ,5&hl=en",
      "Cluster":"ID 14391789733165934338",
```

MEMEX GEOPARSER :

We indexed the json with geo-locations and measurements together. We indexed the jsons with journal output and SWEET concepts separately. This was done because we didn’t wanted to waste time and we were going very slow on following :

1. We were blocked by google many times for days.
2. We didn’t knew clearly what was expected from the SWEET task.

Below is the snapshot of Map generated by MEMEX GeoParser.



EXTRA CREDIT :

Getting Tika up and running EXIFTool

- EXIT provides the technical metadata of the of images, sound file formats handled by digital cameras.
- The metadata that we fetch from this is a great deal of help to photography. It provides certain metadata which is very useful. For Example : The photo manipulation software might modify certain data but the embedded data extracted from EXIF will still persist providing true sense of what the image or the audio is.