Homework: Scientific Content Enrichment in the Text Retrieval Conference (TREC) Polar Dynamic Domain Dataset
Due: April 5th, 2016, 12pm PT

1. Overview

You have spent a great deal of time analyzing and learning from the TREC Dynamic Domain Polar dataset. In the first assignment, you learned how to leverage the knowledge from the MIME taxonomy lecture; from the MIME detection lecture, and from the Deduplication lecture to better identify the ~10% of content from that dataset that was previously identified as application/octet-stream. You have taken those lessons learned and developed amazing D3 visualizations and automated fingerprint identification and ContentBasedMIME identification capabilities in Tika. Great work!

In this assignment, we will expand on your understanding of the Polar dataset by directly leveraging the knowledge you've gained from several lectures in course: Content Extraction; Metadata; Information Clustering and Similarity; and Named Entity Recognition – to scientifically enrich the Polar dataset and to make it something that we can begin to pipe into http://polar.usc.edu a new website being created to provide Polar Data Insights to the scientific community. You will extract measurements from the 1.7 million files and URLs captured in the Polar Deep Data search. Measurement extraction is the process of identifying numbers in the data, and their associated units e.g., "temperature is 7 degrees Celsius" in which the measurement is *7 degrees Celsius* as extracted from the text and in which the measurement number is *7* and the measurement units is *degrees Celsius*. To do so you will build upon both Content Extraction and Tag Ratio techniques that we discussed and use those techniques to isolate and identify the text in the data files and by adding these capabilities to enrich Tika's XHTML annotation based extraction.

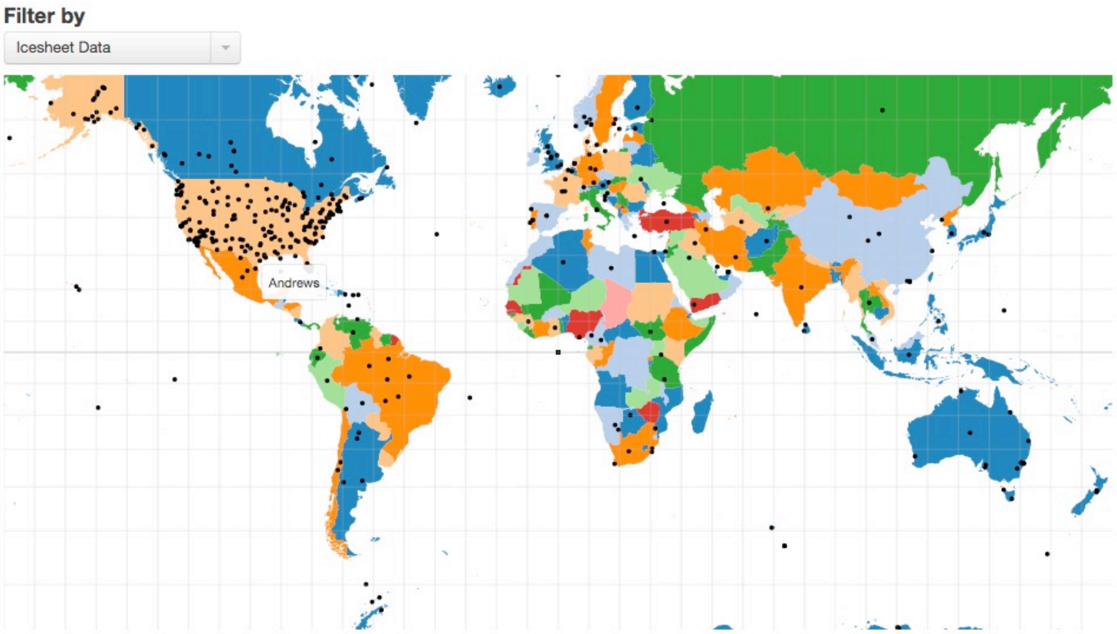
You will also leverage lessons learned from our Metadata lectures to add TEI metadata to any and all scientific publications collected in the dataset, and also bring in external publications through the use of the Google Scholar API to tie scientific publications and associated records to the TREC Polar DD dataset. You will enrich the data by extracting location features from text and mentions using the GeoTopicParser in Tika. You will build upon extraction techniques to classify and annotate the dataset using terminology from the Semantic Web for Earth and Environmental Terminology (SWEET) ontology. You will make the data identifiable using Digital Object Identifiers (DOI), and finally you will construct a metadata evaluation/quality score and annotate each record with the richness of its metadata using heuristics that we discussed during the Metadata lecture.

You will then take the extracted content and metadata and create an inverted index / search index of the JSON data using ElasticSearch and/or Apache Solr. This index will serve as a basis to connect Named Entity Recognition (NER) tools such as the MEMEX GeoParser technology that will allow you to create interactive zoomable maps that allow exploration of the data and its location mentions. Finally, you will also bring in the knowledge gained in information clustering to create metadata and knowledge driven clusters and relationships of the Polar data using SWEET classifications; related publications and authors, and also group together objects with similar metadata quality and enrichment.

You will explore this knowledge by selecting no less than six gallery visualizations from the Data

Driven Documents library to explore and interact with your enriched Polar data. Imagine exploring a dendrogram of citations and authors of related publications to a dataset. Imagine a binned histogram showing number of datasets that measure a particular climate parameter, such as temperature, sea level rise, and other parameters. A geo-map with density information overlaying textual mentions of ice sheets in the Polar dataset – such as the one shown in Figure 1.

In short you will *significantly* enrich the content of the Polar dataset and make the enrichment clickable, interactive and something anyone can explore. The ultimate goal is for each team to contribute their enrichment to the http://polar.usc.edu/ website significantly enriching the Polar insights available from this national dataset.



Filter by

Icesheet Data

Figure 1: Map of Ice Sheet Data extracted from the TREC Polar Dataset generated using D3 from previous classes.

2. Objective

The objective of the assignment is to significantly enrich the metadata, and automatically extracted text and entities from the TREC Polar Dataset, and to make the dataset easily to relate to and to interact with. To do so, you will apply and leverage knowledge gained from context extraction, metadata, information similarity and clustering, and from the named entity recognition lectures.

To do this, the assignment is broken down into several phases, which more or less can be mapped to:

1. Context Extraction Enrichment – you will apply the Tag Ratios algorithm to identify text, and you will construct a Tika parser to extract Measurement

mentions from text automatically.

2. Metadata Enrichment – you will apply the GROBID journal parser with Tika, and extract TEI metadata, and also scientific publication metadata using the Google Scholar API to develop a network of related scientific publications to your Polar dataset, and to map publications to the data. In addition, you will classify the data using a common Earth science domain model, ontology, called SWEET, for Semantic Web for Earth and Environmental Terminology (http://sweet.jpl.nasa.gov/). You will also create Digital Object Identifiers (DOIs) for your data.
3. Information Similarity and Clustering – you will create clusters of your Polar data using the enriched measurements extracted, and using the enriched metadata, and browse and expose your information using Data-Driven-Documents visualizations after ingesting data into Apache Solr and/or ElasticSearch.
4. Named Entity Recognition (NER) – you will apply geospatial NER using the GeoTopicParser in Apache Tika and using the MEMEX GeoParser tools.

You will create a Github repository in your group to store the code and to separate and break up your project into functional tools along the above lines.

The assignment specific tasks will be specified in the following section.

3. Tasks

1. Download and install Apache Tika using same instructions from assignment #1.
2. Baseline off the Polar Full Dump and Common Crawl data from assignment #1.
3. Create a TagRatio Tika Parser as an approach to isolating the text found within any file and content type
    a. Use the TagRatio algorithm to determine areas of text within extracted content.
    b. Apply your TagRatio parser to the Polar dataset.
    c. After identifying relevant portions of each file in which the text is identified, apply NamedEntity recognition using Tika's NER capabilities to extract relevant measurement mentions from the text. You may use any of the NER tools listed on: http://wiki.apache.org/tika/TikaANDNER and/or http://wiki.apache.org/tika/TikaAndNLTK
    d. The output of your parser should be the extracted measurements from any of the files within the polar dataset.
4. Create a DOI generation ContentHandler or Parser in Tika
    a. The DOI generation can use as a model https://pkp.sfu.ca/wiki/index.php?title=DOIPluginsDocumentation however you are not required to officially register your DOIs.
    b. You can also use URL shortener libraries such as any found on Github that have a permissive license e.g., https://github.com/ldidry/lstu, or https://github.com/YOURLS/YOURLS
        i. If you use a URL shortener the URL prefix should be polar.usc.edu/<short%20url>

5. Perform Content Extraction and NER using the Grobid Journal Parser as documented on the Tika wiki
    a. Examine the parser here: http://wiki.apache.org/tika/GrobidJournalParser/
    b. Ensure that TEI annotations are extracted from the scientific publications in your Polar dataset
    c. For each Polar dataset record, pull in at least 20 related publications using the Google Scholar API. Focus on the text/html, application/x-html, application/pdf MIME types.
        i. Generate a Tika Parser that calls the Google Scholar API. Since the API isn't officially documented you can use: https://github.com/ckreibich/scholar.py
    d. Extract and identify publication Authors, Publication Year, Affiliations and other information from the related publications.
6. Perform GeoTopicParsing using the Tika GeoTopicParser
    a. See: http://wiki.apache.org/tika/GeoTopicParser/
    b. Perform geotopic parsing on the entire Polar dataset – develop a program that performs this extraction across the entire dataset.
7. Develop a Tika Parser and NER technique that extracts concepts from the Semantic Web for Earth and Environmental Terminology (SWEET)
    a. See SWEET described at: http://sweet.jpl.nasa.gov/
    b. Hint: take a look at Tika's NER integration with e.g., NLTK, CoreNLP/NER and OpenNLP. Consider doing an NER pass first to identify concepts or entities and then comparing the entities against SWEET's ontology concepts.
8. Create a Metadata quality score that takes into account so-called "Good Metadata" practices we discussed in lecture
    a. Appropriate to the materials, and users in the collection and describe the object's intended use – measure this e.g., by whether or not the metadata for the object has a description field, and a title, and a version, etc.
    b. Supports interoperability (common representation, and abilities to resolve naming conflicts) – measure this e.g., by whether or not there are other known names, and/or aliases for the object.
    c. Uses standard controlled vocabularies – measure this by identifying the number of metadata models and fields available after extracting metadata from the object.
    d. Identifies terms and conditions of use – measure this by e.g., if the object has metadata that identifies its license.
    e. Is an object itself and allows for unique identification and preservation – measure this by using a constant score since you are adding DOIs to the object.
    f. Allows for long term management of objects in collections – measure this by using a constant score as you are creating an index of these objects in the collection.
9. Ingest the extracted text/metadata and features developed in your parsers above into Apache Solr and/or ElasticSearch
    a. Identify which index technology you are using.

   b. Develop a program to iterate through the Polar data and ingest the extracted data from your parsers into the index.
   c. Develop a schema for Solr and/or ElasticSearch to represent your data. You should deliver and identify this schema as part of your report and think carefully about what fields are important to visualize, and what are important to search and to find the data.

10. Deploy and stand up the MEMEX GeoParser, available here:
    a. https://github.com/MBoustani/GeoParser
    b. Run GeoParser NER extractions across the Polar data index you generated in step 9 and generate the location map
11. Deploy and apply the tika-similarity library to connect to your Solr and/or ElasticSearch index with extracted NER features, and text
    a. Cluster the data according to Measurement extractions.
    b. Cluster the data according to related publications and authors.
    c. Cluster the data according to extracted locations.
    d. Cluster the data according to SWEET features.
    e. Develop a program that connects to your Solr and/or ElasticSearch and that performs either the k-means and/or hierarchical clustering technique using Jaccard similarity, edit distance, and/or cosine similarity and that produces output D3 visualizations. Please choose carefully and describe your thought process behind your distance metric and clustering decision.
12. Create a simple set of web pages with at least 6 D3 visualizations of your Polar data. You must choose the 6 visualizations from Mike Bostock's library here:
    a. https://github.com/mbostock/d3/wiki/Gallery
    b. The visualizations and websites should connect to your Solr and/or ElasticSearch index
    c. Contribute your web pages as pull requests to polar.usc.edu, by contributing them here: https://github.com/USCDataScience/polar.usc.edu
13. (**EXTRA CREDIT**) Connect the GeoParser application to the actual data in your Solr index
    a. Create a pop up in GeoParser that displays the metadata record
    b. In the search box functionality for GeoParser, allow for both the locations and documents to be returned in a search list.
14. (**EXTRA CREDIT**) Run any of the feature and/or content extractions listed on the Tika wiki page http://wiki.apache.org/tika/ that haven't already been run on the Polar data
    a. Why did you chose the Content Extractions?
    b. What additional knowledge did you gain from the features?

4. Assignment Setup

4.1 Group Formation

Please keep the same groups as for your assignment #1. If you have any questions please contact:

Divydeep Agarwal
divydeea@usc.edu

Salonee Rege
saloneer@usc.edu

Chandrashekar Chimbili
chimbili@usc.edu

Use subject: CS 599: Team Details

4.2 TREC-DD-Polar Dataset

Access to the Amazon S3 buckets containing the TREC-DD-Polar dataset has already been made. Please use both the NSF common crawl data and the Polar full dump data.

4.3 Installing and Building Apache Solr

You will need to build Apache Solr from the 4_10 branch of lucene-solr to take advantage of a fix that the Professor provided:

http://svn.apache.org/repos/asf/lucene/dev/branches/lucene_solr_4_10/

You can find more information here:
https://issues.apache.org/jira/browse/SOLR-7137
https://issues.apache.org/jira/browse/SOLR-7139

Apache Solr comes with a web application server (Jetty), or you can also deploy and configure Solr with Apache Tomcat. Either way will work fine for this assignment and the instructions are provided here:

https://cwiki.apache.org/confluence/display/solr/Running+Solr+on+Jetty
https://cwiki.apache.org/confluence/display/solr/Running+Solr+on+Tomcat

You should also review the basic installation instructions:

http://wiki.apache.org/solr/SolrInstall

Once installed, you will need to configure Solr to accept your data model.

For ElasticSearch, see the relevant documentation on the ElasticSearch site here:
https://www.elastic.co/guide/en/elasticsearch/guide/1.x/_installing_elasticsearch.html

4.4 Some hints and helpful URLs

This is an extremely long and challenging assignment. Please start early. Please use the relevant Tika mailing lists for questions related to content extractions, metadata, clustering, etc. Please also use Github for raising issues and tag the professor and/or TAs in the issues using Github's tagging feature.

5. Report

Write a short 4 page report describing your observations, i.e. what features did you find most useful in exploring the Polar data? Were you able to take advantage of Tag Ratios to isolate the measurement data? Did NER and SWEET terminology mapping work well – was the NER unable to identify SWEET categories and concepts? Did the D3 interactive visualizations help you understand the data? Were particular features that you extracted such as the geo-locations more effective in producing clusters? Were particular cluster techniques e.g., k-means, more meaningful than hierarchical clustering? What about distance metrics – which ones were more effective (Jaccard, Edit Distance, etc.) Why? 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? 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?

6. Submission Guidelines

This assignment is to be submitted ***electronically, by 12pm PT*** on the specified due date, via Gmail csci599spring2016@gmail.com. Use the subject line: CSCI 599: Mattmann: Spring 2016: CE_MET_NER Homework: Team XX. So if your team was team 15, you would submit an email to csci599spring2016@gmail.com with the subject "CSCI 599: Mattmann: Spring 2016: CE_MET_NER Homework: Team 15" (no quotes). **Please note only one submission per team**.

- All source code is expected to be commented, to compile, and to run. You have several programs that you are developing, so please be concise and provide instructions on how to run your programs. Please also identify a command line interface and provide documentation on the parameters. Do **not** submit *.class files. We will compile your program from submitted source. If your modifications include interpreted scripts, we will run those.
- Please consider submitting your Tika extractors directly to Apache Tika using: http://github.com/apache/tika/#contribuing instructions.
- Teams are asked to file issues in the http://polar.usc.edu/ Githb repository to help augment the site with both the extractions, new visualizations and features that come from this project. Contributions also will be used to refine the TREC dataset, and also be disseminated to DARPA and NSF.
- Also prepare a readme.txt containing any notes you'd like to submit.
- Do **not** include tika-app-1.11.jar in your submission. We already have this.
- However, if you have used any external libraries other than Tika, you should include those jar files in your submission, and include in your readme.txt a detailed explanation of how to use these libraries when compiling and executing your program.
- Save your report as a PDF file (Lastname_Firstname_ CE_MET_NER.pdf) and include it in your submission.
- Compress all of the above into a single zip archive and name it according to the following filename convention:

  <lastname>_<firstname>_CSCI599_HW_ CE_MET_NER.zip

  Use only standard zip format. Do **not** use other formats such as zipx, rar, ace, etc.
- If your homework submission exceeds the Gmail's 25MB limit, upload the zip file to Google drive and share it with csci599spring2016@gmail.com.

Important Note:

- Make sure that you have attached the file the when submitting. Failure to do so will be treated as non-submission.

- Successful submission will be indicated in the assignment’s submission history. We advise that you check to verify the timestamp, download and double check your zip file for good measure.
- Again, please note, only **one submission per team**. Designate someone to submit.

6.1 Late Assignment Policy

- -10% if submitted within the first 24 hours
- -15% for each additional 24 hours or part thereof