Homework: Evaluating your Content Analysis Text Retrieval Conference (TREC) Polar Dynamic Domain Dataset
Due: May 3rd, 2016 12pm PT

---

1. Overview

Over the semester you have added a great understanding and analysis to the Text Retrieval Conference's (TREC) Polar Dynamic Domain Dataset (TREC-Polar-DD). You started out improving its MIME diversity analysis and putting to rest any doubt of what the 250K application/octet-stream files were, helping to improve Tika's MIME repository along the way and to make it more robust and comprehensive. In addition, you've now spent a large assignment and effort working to enrich the content from TREC-DD-Polar with improved extraction, named entity recognition, metadata analysis, information similarity and clustering, and scientific enrichment. It's amazing what you have all accomplished and I'm extremely proud of you all!

In your final assignment, you will continue with your analysis of the TREC-DD-Polar Dataset, and you will complete your analysis by *evaluating your content detection for Big Data*. As we have discussed in class, evaluating the efficacy, utility, and overall contribution of your Content detection approach is an extremely important and difficult challenge. Questions such as Is my MIME detection good? Are my parsers extracting the right text? Are we selecting the right parser? Is my Metadata appropriate? What's missing? How well is my language detection performing? Are there mixed languages? How well is my Machine Translation? Do my Named Entities make sense?

Figure 1. NER evaluation for Planetary Science.

The overall content detection evaluation process can be divided into several steps:

1. Selection of Content – how do you acquire your data, and what process do you follow?
2. Units for Content – in other words, what are the default content objects you are evaluating?
3. Preparing Content for Coding (Text Processing) – making text and metadata uniformly processed for enrichment.
4. Coding the Content (Enrichment) – Performing annotation and extraction from the

uniform text and metadata.
5. Counting and Weighting – Evaluating and assessing your results.
6. Drawing Conclusions – Deriving meaning and insights from your results.

The answers to these questions specifically for TREC-DD-Polar are what you are going to explore in this assignment.

2. Objective

The objective of this assignment is to undergo a full evaluation of your content detection and analysis process following steps 1-6 in the preceding section for the TREC-DD-Polar dataset. To do so we will first identify the answers to the questions above and frame where we're headed in this assignment.

1. Your content is the TREC-DD-Polar dataset https://github.com/chrismattmann/trec-dd-polar. In particular, with the Common Crawl Architecture (CCA) data format version of TREC-DD-Polar you have full information about the request and response for each content request.
2. The content object units that you will be evaluating are:
    a. Your enriched JSON extracted from HW #2 and inserted into either Apache Solr or ElasticSearch
    b. The upstream CCA TREC-DD-Polar data that you have been using all semester.

    The text, metadata, and language information from 2a and 2b are your source content unit for the evaluation.
3. Your tokenized, analyzed text that you inserted into Apache Solr and/or ElasticSearch is your processed text for the analysis. You have performed TagRatios on the text, and put it into a search index.
4. Your enrichment process that produced NER, Geo topic information, associative publications and other derived metadata and entities will be evaluated in this step. You will compare amongst several sources of ground truth and look for agreement – such as that shown in Figure 1.
5. You will compute several metrics that you can use in step 6 to evaluate your content evaluation process and to draw insights as specified in Section 3.5.
6. From the analysis conducted, draw insights from the data as will be specified in Section 3.

As with the prior assignments 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. You will need Apache Tika installed.
2. You will need your Apache Solr index, and/or Elasticsearch.
3. Baseline off the Polar Common Crawl data.
4. Identify the classification path from request to content – what categories of pages were part of the request, and what named entities were present on the arrived at page?
    a. Produce a D3 visualization showing this iteration.
    b. Did the crawler find the most relevant pages? Why, why not?
5. Derive File size diversity of CCA dataset by MIME type.
    a. Compute size ratio of Solr index and/or ElasticSearch to file size original diversity.
    b. Produce a D3 visualization of these metrics.
6. Develop a program and associated D3 visualization that demonstrates Parser call chain and how much text and metadata was actually obtained
    a. Plot Parser Hierarchy versus
        i. Amount of Text retrieved per file size per MIME type (use information from 5)
        ii. Amount of Metadata retrieved per file size per MIME type (use information from 5)
7. Compute Language identification and diversity across the dataset.
    a. Produce D3 visualizations of the language diversity.
8. Produce a Word Cloud D3 visualization of your text, metadatea, and language to find maximal occurring topics in your dataset. This should include relevant SWEET topics from Assignment. Note some teams already produced a Word Cloud in Assignment #2. If so, skip this step.
9. Download and install the four Named Entity Recognition toolkits
    a. NLTK http://nltk.org/ and NLTK REST
    http://github.com/manalishah/NLTKRest
    http://wiki.apache.org/tika/TikaAndNLTK
    b. CoreNLP/NER – http://wiki.apache.org/tika/TikaAndNER
    c. OpenNLP – http://wiki.apache.org/tika/TikaAndNER
    d. Grobid Quantities - https://github.com/kermitt2/grobid-quantities/
        i. For Grobid Quantities, produce a Tika NER implementation that invokes Grobid Quantities via its REST service.
        ii. Contribute as a pull request to Apache Tika **for extra credit**.
        iii. Update the NER wiki for Tika describing how to use Grobid Quantities **for extra credit.**
    e. Build an algorithm to compute maximal joint agreement between the 3 algorithms and encode the algorithm as a CompositeNERAgreementParser in Tika.
    f. Produce a visualization in D3 for evaluating the maximal joint agreement NER between the most frequently occurring entities.

    g. Analyze whether your new joint agreement produces any update NER for your metadata records, and if so, add new maximal joint agreement NER to the metadata records in Solr and/or ElasticSearch.
10. Identify the Spectrum (range, min/max) of measurements
    a. For each measurement, show the min-max range, and average/mean for the measurements.
    b. Compute for all measurements of a particular type;
    c. Compute for all pages from a particular domain;
    d. Compute for all files of a particular MIME type;
    e. Produce D3 visualizations for a-d.
11. Contribute your D3 visualizations for all steps as a pull request to http://polar.usc.edu/.
12. Submit a video demonstrating a screencast walking through the answers to the above questions using your D3 visualizations and knowledge gained from this assignment. Publish your Video on YouTube and include a link to it in the report for your assignment.

4. Assignment Setup

4.1 Group Formation

Please keep the same groups as for your assignment #2. 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.

5. Report

Write a short 4 page report describing your observations, and answer the following questions, using your visualizations, and knowledge gained through the semester:

1. Is your MIME detection good? Define "good".
2. Are your parsers extracting the right text? Define "right".
3. Are we selecting the right parser? Define "right".
4. Is your Metadata appropriate? What's missing? You can use your Metadata score generated from assignment #2 here, and also your results from this assignment.
5. How well is my language detection performing? Comment based on the diversity of the languages derived in this assignment. Are there mixed languages? Did it affect your accuracy?
6. Do your Named Entities make sense?

In addition to your report you will submit a video demonstrating a screencast walking through the answers to the above questions using your D3 visualizations and knowledge gained from this assignment. Publish your Video on YouTube.

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: EVAL 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: EVAL 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. If you submit your code as pull requests for extra credit to e.g., Tika, then please identify so in your report.
- Teams are asked to file issues in the http://polar.usc.edu/ Github repository to help augment the site with your Content evaluation. 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.
- If you use any libraries other than Tika, you should include those jar files/requirements.txt/etc. 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_ EVAL.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_ EVAL.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