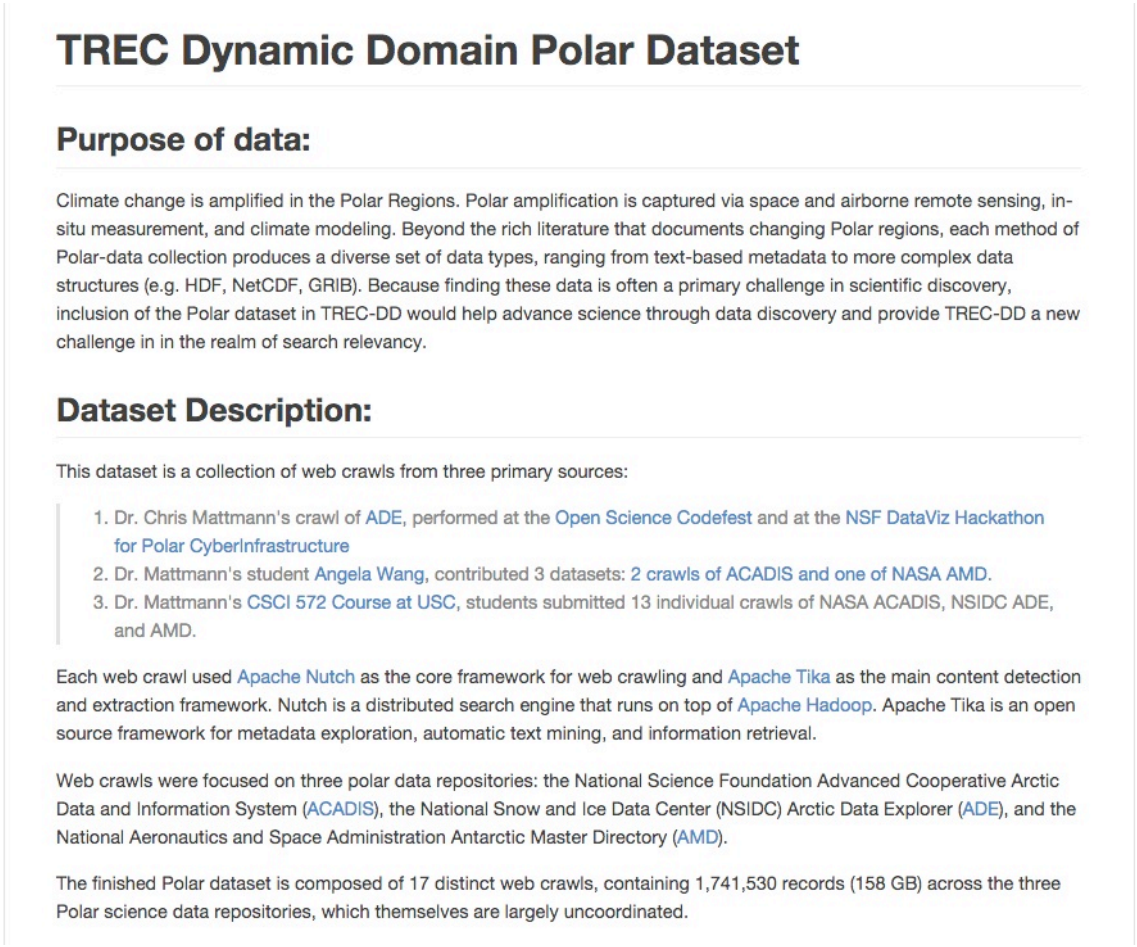
Homework: MIME Diversity in the Text Retrieval Conference (TREC)
Polar Dynamic Domain Dataset
Due: Tuesday, March 1, 2016 12pm PT

---

1. Overview



TREC Dynamic Domain Polar Dataset

Purpose of data:

Climate change is amplified in the Polar Regions. Polar amplification is captured via space and airborne remote sensing, in-situ measurement, and climate modeling. Beyond the rich literature that documents changing Polar regions, each method of Polar-data collection produces a diverse set of data types, ranging from text-based metadata to more complex data structures (e.g. HDF, NetCDF, GRIB). Because finding these data is often a primary challenge in scientific discovery, inclusion of the Polar dataset in TREC-DD would help advance science through data discovery and provide TREC-DD a new challenge in in the realm of search relevancy.

Dataset Description:

This dataset is a collection of web crawls from three primary sources:

1. Dr. Chris Mattmann's crawl of ADE, performed at the Open Science Codefest and at the NSF DataViz Hackathon for Polar CyberInfrastructure
2. Dr. Mattmann's student Angela Wang, contributed 3 datasets: 2 crawls of ACADIS and one of NASA AMD.
3. Dr. Mattmann's CSCI 572 Course at USC, students submitted 13 individual crawls of NASA ACADIS, NSIDC ADE, and AMD.

Each web crawl used Apache Nutch as the core framework for web crawling and Apache Tika as the main content detection and extraction framework. Nutch is a distributed search engine that runs on top of Apache Hadoop. Apache Tika is an open source framework for metadata exploration, automatic text mining, and information retrieval.

Web crawls were focused on three polar data repositories: the National Science Foundation Advanced Cooperative Arctic Data and Information System (ACADIS), the National Snow and Ice Data Center (NSIDC) Arctic Data Explorer (ADE), and the National Aeronautics and Space Administration Antarctic Master Directory (AMD).

The finished Polar dataset is composed of 17 distinct web crawls, containing 1,741,530 records (158 GB) across the three Polar science data repositories, which themselves are largely uncoordinated.

Figure 1: The TREC Dynamic Domain Polar Dataset
http://github.com/chrismattmann/trec-dd-polar/

In the first assignment you are going to explore the domain diversity and MIME diversity of the National Institutes of Standards and Technology (NIST) Text Retrieval Conference (TREC) Polar Dynamic Domain dataset. The TREC-Polar-DD dataset (as it will be referred to from here on in the assignment) was collected over the past few years across various CSCI 572 courses here at the University of Southern California (USC) and in collaboration with the NSF Polar CyberInfrastructure program, and the DARPA Memex program and its TREC Dynamic Domain track. TREC is a national set of working groups that includes researchers interested in information retrieval, and in content detection and analysis.

TREC-Polar-DD is a diverse dataset. It was collected with the Apache Tika, Nutch and Solr software systems, and includes over 158GB in total of data comprising 1.7 million web pages obtained during 17 distinct web crawls from the National Science Foundation Advanced Cooperative Arctic Data and Information System (ACADIS), the National Snow and Ice Data Center (NSIDC) Arctic Data Explorer (ADE), and the National Aeronautics and Space Administration Antarctic Master Directory (AMD) data archives. The dataset has many or most of the characteristics of Big Data and Deep, Dark Data that we have discussed in class, some of which will be enumerated in the ensuing paragraph.

The ACADIS, AMD and ADE sites each had their own crawling difficulties. ACADIS required a login in order to get at many of the datasets, crawling it required having to register and crawl behind forms. Much of the ADE site was paginated Ajax and JavaScript so the crawler had to click on Ajax links, and grab the content. Once it got to the content, the data was more than simply HTML files, it also included scientific data files like Gridded Binary (GRIB) data, and Hierarchical Data Format (HDF) and Network Common Data Format (NetCDF) files that store complex array data and measurement multi-dimensional data in space and time.

There are of course in this dataset of 1.7 million records a number of interesting MIME formats, as we learned about in the MIME taxonomy lectures in class. For example, there are ~700, 000 application/* files (~400,000 of which are XHTML); 1000 audio files; 200, 000 images, hundreds of email messages, 800, 000 text/* files, and 100s of videos. Apache Tika generated these classifications using mechanisms for MIME detection which we have discussed in class e.g., using filename extensions ("glob patterns"); MIME magic and precedence, and XML namespace detection, amongst other approaches.

Looking more broadly throughout the domain diversity – there are some curiosities with this data, though. For example, within the application/* type, more than 200, 000 files were unclassified (or generally classified as "application/octet-stream"). Were these scientific data files that were not detectable? Are they possible other file types? Other researchers on our team including a former postdoc and now leader at the Earth Science Information Partners (ESIP) organization (Annie Bryant Burgess) created the following chart that demonstrates some recent improvements to Tika's MIME classification for scientific data formats as that may aid in detecting traditionally "unknown" types from this set:

Data types found in Polar data repositories.

	Atmosphere	Agriculture	Biological	Biosphere	Climate	Cryosphere	Human Dim.	Land Surface	Oceans	Paleoclimate	Solid Earth	Hydrosphere	%	
arcinfo							1						0.1%	
binary			4	1		2			4	1			0.6%	
geotiff				2		5	1	1					0.5%	
html	3			2	1	5	1	2	2				0.8%	
jpeg	8			21		70			2			52	8.0%	
matlab						9			29			1	2.1%	*
access			2	2	1	2	2	2	1				0.6%	
excel	55	7	10	176	11	156	4	115	82	39	7	87	39.4%	
word						11		11	3	9		2	1.9%	
netcdf	35				2	4			20			3	3.4%	*
other ascii	62	2	2	49	13	53		11	85			13	15.3%	
other binary	6	2	4	5	2	13	3	17	12	3	3	2	3.8%	
pdf	2			1		13	1	13	6	12	1	4	2.8%	
png	349						22	1	1	1			19.7%	
shapefile					1	2	1					2	0.3%	
grib	4							1					0.3%	*
unknown									10				0.5%	

Fig 2. MIME diversity adapted from A. Bryant. Apache Tika: Cool Insights into Polar Data. ApacheCon NA 2015, Austin, TX, 2015.
http://events.linuxfoundation.org/sites/events/files/slides/ApacheCon2.pdf

Annie's research suggests that the application/octet-stream unknown types may actually be scientific data files that Tika's MIME detection system did not have sufficient detection abilities recorded for. For example, Matlab files, NetCDF files, and GRIB files.

In this project, you are going to perform a MIME diversity analysis of the TREC-DD-Polar dataset helping to answer the questions above, and providing new and improved mechanisms for MIME detection in Tika for scientific Big Data.

2. Objective

You will use the knowledge you have learned thus far regarding the MIME Taxonomy, data similarity, and regarding learning Byte-based fingerprints of the data via Byte Frequency Analysis (BFA), Byte Frequency Distribution (BFD) Correlation, Byte Frequency Cross-Correlation (BFC), and File Header Trailer (FHT). You are going to implement a set of MIME diversity programs and applications that will help in better understanding these unknown types in a rich scientific domain. You will compute BFA, BFC and FHT of these unknown (and other) Polar data types from the dataset, and you will build a system that allows visual interaction and introspection of the MIME diversity in this dataset. Those classifications will improve Tika's overall ability by suggesting new MIME magic for its database, and improve techniques for MIME detection in the Big Data present in the TREC-DD-Polar dataset.

From the TREC-DD-Polar dataset, you will pick at least 15 MIME types and you must include the application/octet-stream ("unknown") type, and you will perform BFA analysis, and leverage the Data-Driven-Documents (D^3) technology (http://d3js.org/) to

build interactive visualizations that display Byte Histograms using the data classified along those MIME types. From the BFA you will perform BFC and write software that will compute correlation between files of that type and the digital fingerprint you arrive at. You will also examine BFC cross correlation, and determine pair-wise correlation of bytes allowing for an increased precision fingerprint, and also suggesting new MIME magic that may be potentially useful for detecting the file type. You will augment Tika and its MIME database with this information. Finally, you will perform FHT analysis and compute new byte fingerprints that will also potentially better inform Tika and its MIME database for detecting these types. Your BFC cross correlation analysis, and FHT analysis should also include interactive D3 plots allowing for exploration of the MIME diversity of the data.

Once complete, Tika's MIME database should be enriched enough that you can perform a re-analysis and MIME diversity classification on the TREC-DD-Polar data. You and your team will validate this by performing an evaluation and comparison of your improved Tika MIME database with the original MIME classifications produced by Tika before your work.

You should take careful thought to develop your programs to compute BFA/BFC/FHT analysis in such a way that they can be contributed to Apache Tika directly upstream to the source code or as external tools managed on Github with close integration with Tika. You will be assessed in this regard during the assignment.

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

3. Tasks

1. Download and install Apache Tika
    a. Chapter 2 in your book covers some of the basics of building the code, and additionally, see http://tika.apache.org/1.11/gettingstarted.html
    b. If using Tika-Python, you can pip install tika to get started.
2. Download and install D3.js
    a. Visit http://d3js.org/
    b. Review Mike Bostock's Visual Gallery Wiki
        i. https://github.com/mbostock/d3/wiki/Tutorials
3. Download the Amazon S3-based TREC-DD-Polar data
    a. First review http://github.com/chrismattmann/trec-dd-polar/
    b. Create a D3 based Pie Chart of the existing MIME diversity of the TREC-DD-Polar dataset using the existing JSON breakdown from Github
4. Perform Byte Frequency Analysis on at 15 (14+ application/octet-stream) MIME types present in TREC-DD-Polar dataset
    a. Write a program that computes the byte frequency as a signature of sample files from each of the 15 types. For the application/octet-stream type, you must use at least 25% of the dataset (or 50,000 files). For the other MIME types you should attempt to use at least 75% of the data available to you for that type. So if a particular MIME type has 5000 files, we would expect you to use 3500 files to perform BFA on.

   b. Your program should generate a JSON file representing the fingerprint for each of the 15 types. You should use your JSON file to display using D3 the BFA fingerprint for the file type, as a line / bar chart. Your visualization should allow interactive introspection of the fingerprint.
5. Perform Byte Frequency Distribution Correlation and Byte Frequency Cross Correlation on the 15 (14+ application/octet-stream) MIME types present in the TREC-DD-Polar dataset
   a. Write a program that computes the byte frequency correlation between an input file and the generated fingerprint for each of the 15 file types that you created in step 4b. You should perform byte frequency correlation using the remaining 25% of the data that you didn't already test out in step 4a and 4b. For the application/octet-stream type, you can use another 25% of the data (50,000 files). For all other of the 14 types, if a particular MIME type had 5000 files, and you used 3500 of them to generate the BFA fingerprint in 4b, then you should use the remaining 1500 for BFC computation in this step.
   b. Develop a D3 visualization of your BFC analysis, and identify at least 2 areas of high correlation, and two areas of low correlation in the analysis. These should be highlighted using your visualization.
   c. Use the information gleaned from step 5b to in turn perform BFC cross correlation, and generate a cross correlation matrix across your 15 MIME types. Generate a D3 visualization of your BFC matrix.
6. Perform FHT analysis on your 15 (14+ application/octet-stream) MIME types present in the TREC-DD-Polar dataset
   a. Write a program that computes FHT analysis on the first 4, 8 and 16 bytes of a similar data distribution from 4a and 5a of your files present in your chosen 15 MIME types.
   b. Show the FHT sparse matrix for your FHT analysis for each of the types using D3 and create a visualization for each type, or a single visualization that allows you to choose the type and see the resultant FHT sparse matrix.
7. Update Tika's MIME repository based on your BFA, BFC and FHT analyses
   a. Use the knowledge gained from steps 4-6 to identify at least two new MIME magic byte fingerprints (including bytes, offsets and priorities) for all of the 15 types and adapt Tika's mimetypes.xml file with this information
   b. Recompile Tika with this new MIME information or point to a new MIME types file via the classpath
   c. Write a new program for MIME diversity analysis using your new Tika mimetypes.xml file, and compare the results with the initially classified MIME types from TREC-DD-Polar. Identify if any MIME type classifications changed.
   d. Generate a D3 visualization (pie-chart and bar-chart) showing MIME diversity using your updated classifications.
8. (**EXTRA CREDIT**) Download and configure Tika Similarity and run it over your dataset

    a. You can find Tika Similarity here (http://github.com/chrismattmann/tika-similarity)
    b. Add cosine distance, and edit distance to Tika-similarity
        i. For Edit distance, look at this repository: https://github.com/harsham05/edit-distance-similarity
        ii. Compare and contrast clusters from Jaccard, Cosine Distance, and Edit Similarity – do you see any differences? Why?
    c. Do the clusters resultant from applying Tika Similarity given you any indication of the MIME type?
9. (**EXTRA CREDIT**) Apply the ContentBased MIME detector in Tika to your TREC-DD-Polar data
    a. See: https://wiki.apache.org/tika/ContentMimeDetection
    b. See: https://issues.apache.org/jira/browse/TIKA-1582
    c. Create a new trained model for one of your identified MIME types you chose above
    d. Apply and evaluate the detector against Tika's default MIME detector – what differences do you see?

4. Assignment Setup

4.1 Group Formation

You can work on this assignment in groups sized 3. Divydeep, the lead grader, will send you instructions on how to formulate groups. If you have questions 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 will be made available by the course producers once your teams have been formed. Take heed, since the dataset itself is ~158Gb (compressed), you may want to distribute the data between your team-mates or only to work with portions of the data at a time. One of the required deliverables in your report is a plan and experience report for how you are going to access the data, and distribute it during your assignment.

4.3 Downloading and Installing Apache Tika

The quickest and best way to get Apache Tika up and running on your machine is to grab the `tika-app.jar` from: http://tika.apache.org/download.html. You should obtain a jar file called `tika-app-1.11.jar`. This jar contains all of the necessary dependencies to get up and running with Tika by calling it your Java program.

Documentation is available on the Apache Tika webpage at http://tika.apache.org/. API documentation can be found at http://tika.apache.org/1.11/api/.

You can also get more information about Tika by checking out the book written by Professor Mattmann called "Tika in Action", available from: http://manning.com/mattmann/.

4.4 Some hints and helpful URLs

Don't forget to study Companding functions such as the A law or u law:
https://en.wikipedia.org/wiki/Companding
You may need this in your byte histogram calculations. You will know as you study the data – if you run into an example similar to the GIF example from lecture, consider using a companding function.

You may want to look into the Tika-Python library as a method for writing your scripts in the assignment:

http://github.com/chrismattmann/tika-python/

More broadly, you can also use any of the Tika libraries present here:
https://wiki.apache.org/tika/API%20Bindings%20for%20Tika

If you do the extra credit, you will need to look at D3:

http://d3js.org/

5. Report

Write a short 4 page report describing your observations, i.e. what you noticed about the dataset as you completed the tasks. Were you able to discern any new MIME types within the 200, 000 application/octet-stream ("unknown") types? How well did BFA work, compared to FHT? Did

the D3 interactive visualizations help you understand the byte frequencies, and to identify patterns? Describe your results from BFA, BFC, and FHT. What MIME types did you pick, and why?

Thinking more broadly, do you have enough information to answer the following:

1. Why Tika's detector was unable to discern the MIME types?
2. Was it lack of byte patterns and specificity in the fingerprint?
3. An error in MIME priority precedence?
4. Lack of sensitivity in the ability to specify competing MIME magic priorities and bytes/offsets?

Also include your thoughts about Apache Tika – what was easy about using it? What wasn't?

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: MIME 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: MIME 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 should have at least a modified mimetypes.xml file and any source files that you added to Tika, if any. Do **not** submit *.class files. We will compile your program from submitted source. If your modifications include interpreted scripts, we will run those.
- Include your program for performing BFA, BFC and FHT analysis.
- Include your D3 HTML and generated JSON code. Also include any program necessary to generate the D3 visualizations.
- Teams will be asked if they would like to contribute their updated MIME classifications code and visualizations to http://polar.usc.edu/ which is currently under development. Contributions also will be used to refine the TREC dataset, and also be disseminatd to DARPA and NSF. If you want to do this, please identify in your report that you would like to do this, and send a message to the professor, and to the course producers.
- 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_MIME.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_MIME.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