Welcome to the JPL/Caltech Virtual Summer on Big Data Analytics!

The exercises for the Content Detection and Analysis for Big Data module are described in this document.

Exercises

Pre-requisites

1. Download and Install Apache Tika (http://tika.apache.org/)
    a. http://tika.apache.org/download.html
    b. Grab the tika-app-1.5.jar file
    c. Install it into /usr/local/tika on Unix or to somewhere in the system path on Windows.
    d. Add an alias for "tika" to /usr/local/tika/tika-app-1.5.jar on Unix, e.g., by
        i. (BASH) alias tika="/usr/local/tika/tika-app-1.5.jar"
        ii. (CSH) alias tika "/usr/local/tika/tika-app-1.5.jar"
    e. Test out Tika, e.g., tika –t Exercises-ContentDetection.pdf

Suggested Issues to Tackle

Exercises are centered around Tika related issues described in the Apache Tika issue tracker, here: https://issues.apache.org/jira/browse/TIKA

Issues are selected based on their difficulty and are expected to be completed within a 2-3 hour time frame. Successful completion of the issue involves submitting and attaching a patch on the issue itself (for which you will be credited if the patch gets committed to the Apache Tika project). For more information on contributing to Apache projects, see: http://wiki.apache.org/nutch/Becoming_A_Nutch_Developer

1. Figure out why Tika won't parse HDF-5 files from the NOAA/NASA JPSS GRAVITE project. See: https://issues.apache.org/jira/browse/TIKA-862 for more details. Comment back the results and a potential patch on the issue.
2. Expand and finish a working version of the Geographic Data Abstraction Library (GDAL) Tika parser originally created by Professor Mattmann. See initial patch here: https://issues.apache.org/jira/browse/TIKA-605 (note you will need to install GDAL on your computer to work on this issue)
3. Update Tika's MIME detection XML file to properly detect X12 files, as described here: https://issues.apache.org/jira/browse/TIKA-627
4. Test and evaluate Tika's OCR support provided via Tesseract, as described here: https://issues.apache.org/jira/browse/TIKA-93 (note you will need to install Tesseract to work on this issue.)