Chris A. Mattmann, NASA JPL, USC & the ASF

@chrismattmann
mattmann@apache.org
Content Extraction from Images and Video in Tika
Background: Apache Tika
Outline

- Text
- The Information Landscape
- The Importance of Content Detection and Analysis
- Intro to Apache Tika
The Information Landscape
Proliferation of Content Types

• By some accounts, 16K to 51K content types*
  • What to do with content types?
    • Parse them, but How?
    • Extract their text and structure
    • Index their metadata
      • In an indexing technology like Lucene, Solr, ElasticSearch
  • Identify what language they belong to
    • Ngrams

* [http://fileext.com](http://fileext.com)
Importance: Content Types

Language Identification

About: 0.620,000 results (0.25 seconds)

Translation Wizard > Language Identification

Aug 25, 2003... Identify the language. If nothing, this page will be

Language Identification

Language identification is the process of automatically determining the language of a text, speech or image. There are different approaches to language identification, such as statistical methods, rule-based methods, and machine learning.

Language Identification in Practice

Language identification is used in various applications, such as machine translation, speech recognition, and image captioning.

Language Identification Tools

There are several tools available for language identification, such as Google Translate, Microsoft Translator, and IBM Watson.

PDF Language Identification

This PDF contains language identification tools and resources that can be used for further study in this area.

CARS.gov - Car Allowance Rebate System (CAR) - Formerly the Rebate System for Alternative Fuel Vehicles

Feb 22, 2010 - The official website of the CARS or Car Allowance Rebate System.

Disney/Pixar Cars - The Official Site

The latest Cars news and movie clips, character biographies, games, photos, toys, and downloads from the Disney/Pixar movie Cars.

Images for cars - View images

Top Prices on New Cars

Find your dream car at the top price! Get the latest car deals and discounts at your local dealer.

California - Cars

Looking for cars in California? Find deals on new and used cars in your area.

Learn more about language identification in this setting and its importance for various applications.
Importance: Content Types
IANA MIME Registry

- Identify and classify file types
  - MIME detection
  - Glob pattern
    - *.txt
    - *.pdf
  - URL
    - http://...pdf
    - ftp://myfile.txt
- Magic bytes
- Combination of the above means
- Classification means reaction can be targeted
Many Custom Applications

• You need these apps to parse these files
  • …and that’s what Tika exploits
Most of the custom applications come with software libraries and tools to read/write these files. Rather than re-invent the wheel, figure out a way to take advantage of them. Parsing text and structure is a difficult problem. Not all libraries parse text in equivalent manners. Some are faster than others. Some are more reliable than others.
Extraction of Metadata

• Important to follow common Metadata models
  • Dublin Core
  • Word Metadata
  • XMP
  • **EXIF**
• Lots of standards and models out there
  • The use and extraction of common models allows for content intercomparison
• All standardizes mechanisms for searching
• You always know for X file type that field Y is there and of type String or Int or Date
Lang. Identification/Translation

• Hard to parse out text and metadata from different languages
  • French document: J’aime la classe de CS 572!
    • Metadata:
      • Publisher: L’Universitaire de Californie en Etas-Unis de Sud
  • English document: I love the CS 572 class!
    • Metadata:
      • Publisher: University of Southern California

• How to compare these 2 extracted texts and sets of metadata when they are in different languages?
• How to translate them?
Apache Tika

- A content analysis and detection toolkit
- A set of Java APIs providing MIME type detection, language identification, integration of various parsing libraries
- A rich Metadata API for representing different Metadata models
- A command line interface to the underlying Java code
- A GUI interface to the Java code
- Translation API
- REST server
- Ports to NodeJS, Python, PHP, etc.

http://tika.apache.org/
Tika’s History

• Original idea for Tika came from Chris Mattmann and Jerome Charron in 2006
• Proposed as Lucene sub-project
• Others interested, didn’t gain much traction
• Went the Incubator route in 2007 when Jukka Zitting found that there was a need for Tika capabilities in Apache Jackrabbit
• A Content Management System
• Graduated from the Incubator to Lucene sub-project in 2008
• Graduated to Apache TLP in 2010
• Many releases since then, currently VOTE’ing on 1.8
Images and Video
The Dark Web

- The web behind forms
- The web behind Ajax/Javascript
- The web behind heterogeneous content types

Examples
- Human and Arms Trafficking
- Tor Network
- Polar Sciences
- Cryosphere data in archives
- DARPA Memex / NSF Polar Cyber Infrastructure

http://www.popsci.com/dark-web-revealed
DARPA Memex Project

• Crawl, analyze, reason, and decide about the dark web
• 17+ performers
• JPL is a performer based on the Apache stack of Search Engines technologies
  • Apache Tika, Nutch, Solr

How is it done today, and what are the limitations?
Search has progressed through several stages due to the increasing size of the Web. While in
initially estimated as hundreds of thousands of pages in 1998, today it has grown to over 20 billion
pages [1]. Search engines first focused on text and its rate of occurrence; then focused on the no-	ion of link analysis and citation [2] [3]; then on interactivity and guided search; and now on the
use of social media – who we interact with, what we comment on, and who we follow (and who
follows us). The next stage, referred to as “deep search,” requires solutions that can bring togeth-
er text, images, video, importance, interactivity, and social media to solve this challenging pro-
blem. Current software and approaches do not effectively integrate these data, hence new software
platforms are required to address the scale and scope of the emerging Web.

The Apache Nutch project [4] provides an open framework for large
targeted, ve
erical
search with capabilities to support all past and potenti-
al future search engine foci. Nutch is a
flexible infrastructure allowing open access to ranking; URL selection and filtering approaches,
to the link graph generated from search, and Nutch has spawned entire sub communities inclu-
ding Apache Hadoop [5] and Apache Tika [1]. It addresses many current needs with the capability
to support new technologies such as image and video search.

What is the proposed work attempting to accomplish or do?
We propose to create specific extensions (Figure 1) to Nutch that w
ill directly improve its
overall technological superiority for search and that will directly allow us to address complex
search problems including human trafficking – we will spend particular time discussing human
trafficking as a motivating example in Section 2.1. We will integrate state-
of-the-art algorithms
developed by Kitware for IARPA Aladdin combined with work by Harvard to provide image and
video understanding support allowing automatic detection of people and things and
massive deployment via Nutch (Section 3.4). We will expand Apache Tika (Nutch’s content system) for
scene understanding, object/person detection and classification in images and video.

We will deliver an interactive and visual
interface for initiating Nutch crawls. The interface
will use the Blaze platform [107] to expose Nutch data and to provide a domain specific language
for crawls. With the Bokeh visualization library [226], the interface will deliver simple interactive
crawl visualization and plotting techniques for exploring crawled information (Section 3.5). The
platform will classify, identify, and thwart predators, help to find victims
and to identify buyers
in human trafficking and will deliver technological superiority in search engines for DARPA.

Who or what will be affected and what will be the impact if the work is successful?
The improvements to media oriented search and the
user interface (UI) and domain specific
language for search will unleash deep search”activities enabling easy use by analysts, and quick
DARPA Memex Project

• 60 Minutes (February 8, 2015)
  • DARPA: Nobody’s Safe On The Internet News:
    • http://www.cbsnews.com/videos/darpa-nobodys-safe-on-the-internet
  • 60 Minutes Overtime (February 8, 2015)
    • New Search Engine Exposes The “Dark Web”

• Scientific American (February 8, 2015)
  • Human Traffickers Caught on Hidden Internet
  • Scientific American Exclusive: DARPA Memex Data Maps
NSF Polar CyberInfrastructure

• 2 specific projects
  • http://www.nsf.gov/awardsearch/showAward?
    AWD_ID=1348450&HistoricalAwards=false
  • http://www.nsf.gov/awardsearch/showAward?
    AWD_ID=1445624&HistoricalAwards=false

• I call this my “Polar Memex”
• Crawling NSF ACADIS, Arctic Data Explorer
  and NASA AMD
• Exposing geospatial and temporal content types
  (ISO 19115; GCMD DIF; GeoTopic Identification; GDAL)
• Exposing Images and Video

http://nsf-polar-cyberinfrastructure.github.io/datavis-hackathon/
Specific improvements

• Tika doesn’t natively handle images and video even though it’s used in crawling the web
  • Improve two specific areas
    • Optical Character Recognition (OCR)
    • EXIF metadata extraction
      • Why are these important for images and video?
  • Geospatial parsing
    • Geo reference data that isn’t geo referenced (will talk about this later)
OCR and EXIF

• Many dark web images include text as part of the image caption
  • Sometimes the text in the image is all we have to search for since an accompanying description is not provided
• Image text can relate previously unlinkable images with features
• Some challenges: Imagine running this at the scale of 40+Million images
  • Will explain a method for solving this issue
• EXIF metadata
  • Allows feature relationships to be made between e.g., camera properties (model number; make; date/time; geo location; RGB space, etc.)
Enter Tesseract

• [https://code.google.com/p/tesseract-ocr/](https://code.google.com/p/tesseract-ocr/)

• Great and Accurate Toolkit, Apache License, version 2 ("ALv2")
• Many recent improvements by Google and Support for Multiple Languages

• Integrate this with Tika!
  • [http://issues.apache.org/jira/browse/TIKA-93](http://issues.apache.org/jira/browse/TIKA-93)
• Thank you to Grant Ingersoll (original patch) and Tyler Palsulich for taking the work the rest of the way to get it contributed
Tika + Tesseract In Action

- https://wiki.apache.org/tika/TikaOCR
- brew install tesseract --all-languages
- tika -t /path/to/tiff/file.tiff
  - Yes it’s that simple
  - Tika will automatically discern whether you have Tesseract installed or not
- Yes, this is very cool.
- Try it from the Tika REST server!
  - In another window, start Tika server
    - java -jar /path/to/tika-server-1.7-SNAPSHOT.jar
  - In another window, issue a cURL request
    - curl -T /path/to/tiff/image.tiff http://localhost:9998/tika --header "Content-type: image/tiff"
Tesseract – Try it out
EXIF metadata

• Example EXIF metadata
  • Camera Settings; Scene Capture Type; White Balance Mode; Flash; 
    Fnumber (Fstop); File Source; Exposure Mode; Xresolution; 
    Yresolution; Recommended EXIF interoperability Rules, Thumbnail 
    compression; Image Height; Image Width; Flash Output; AF Area 
    Height; Model; Model Serial Number; Shooting Mode; Exposure 
    Compensation..
  • AND MANY MORE

• These represent a “feature space” that can be used to relate images,
  *even without looking directly at the image*
  • Will speak about this over the next few slides
What are web duplicates?

• One example is the same page, referenced by different URLs:

  http://espn.go.com

  http://www.espn.com

• How can two URLs differ yet still point to the same page?
  • the URL’s host name can be distinct (virtual hosts),
  • the URL’s protocol can be distinct (http, https),
  • the URL’s path and/or page name can be distinct
What are web duplicates?

• Another example is two web pages whose content differs slightly:
  - Two copies of www.nytimes.com snapshot within a few seconds of each other;
  - The pages are essentially identical except for the ads to the left and right of the banner line that says The New York Times;
Solving (near) Duplicates

• Duplicate: Exact match;
  • Solution: compute fingerprints or use cryptographic hashing
  • SHA-1 and MD5 are the two most popular cryptographic hashing methods

• Near-Duplicate: Approximate match
  • Solution: compute the syntactic similarity with an edit-distance measure, and
  • Use a similarity threshold to detect near-duplicates
    • e.g., Similarity > 80% => Documents are “near duplicates”
• Compare character by character two documents to see if they are identical.
  • However, this could be very time consuming if we must test every possible pair.
• We might hash just the first few characters and compare only those documents that hash to the same bucket.
  • But what about web pages where every page begins with `<HTML>`?
• Another approach would be to use a hash function that examines the entire document.
  • But this requires lots of buckets.
• A better approach is to pick some fixed random positions for all documents and make the hash function depend only on these.
  • This avoids the problem of a common prefix for all or most documents, yet we need not examine entire documents unless they fall into a bucket with another document.
• But we still need a lot of buckets.
General Paradigm: Similarity

• Define a function f that captures the contents of each document in a number
  • E.g. hash function, signature, fingerprint
• Create the pair <f(doc_i), ID of doc_i> for all doc_i
• Sort the pairs
• Documents that have the same f value or an f value within a small threshold are believed to be duplicates
Distance Measures

• Distance measure must satisfy 4 properties
  • No negative distances
  • D(x,y) = 0 iff x=y
  • D(x,y) = d(y,x) symmetric
  • D(x,y) <= d(x,z) + d(z,y) triangle inequality

• There are several distance measures that can play a role in locating duplicate and near-duplicate documents
  • Euclidean distance – \(d([x_1 \ldots x_n], [y_1, \ldots, y_n]) = \sqrt{\sum (x_i - y_i)^2} \) for \(i = 1 \ldots n\)
  • Jaccard distance – \(d(x,y) = 1 - \text{SIM}(x,y)\) or 1 minus the ratio of the sizes of the intersection and union of sets \(x\) and \(y\)
  • Cosine distance – the cosine distance between two points (two \(n\) element vectors) is the angle that the vectors to those points make; in the range 0 to 180 degrees
  • Edit distance – the distance between two strings is the smallest number of insertions and deletions of single characters that will convert one string into the other
  • Hamming distance – between two vectors is the number of components in which they differ (usually used on boolean vectors)
Jaccard Similarity

• Similarity Measures
• Resemblance(A,B) is defined as
  \[ \frac{\text{size of } (S(A,w) \text{ intersect } S(B,w))}{\text{size of } (S(A,w) \text{ union } S(B,w))} \]
• Containment(A,B) is defined as
  \[ \frac{\text{size of } (S(A,w) \text{ intersect } S(B,w))}{\text{size of } (S(A,w))} \]
• 0 ≤ Resemblance ≤ 1
• 0 ≤ Containment ≤ 1
• EXIF metadata can be treated as “FEATURES” that you can compute containment and resemblance.
Tika Image Similarity


• First pass it a directory e.g., of Images
  • For each file (image) in the directory
    • Run Tika, extract EXIF features
    • Add all unique features to “golden feature set”
  • Loop again
    • Use extracted EXIF metadata for file, compute size of feature set, and names, compute containment which is a “distance” of each document to the golden feature set
  • Set a threshold on distance, then you have clusters
Tika Image Similarity

• Results are extremely promising
Wait, DataViz??!

- [http://d3js.org/](http://d3js.org/)
- Invented by Mike Bostock and Vadim Ogievetsky and Jeff Heer [http://vis.stanford.edu/papers/d3](http://vis.stanford.edu/papers/d3)
Wait, DataViz??!

- Creates SVG tied to DOM aspects of the page
- Page loads e.g., data, (JSON or other), controls access via DOM
- Manipulate DOM and bind DOM to SVG elements
- Tons of examples

[Visual Index]

Demo

• Tika Image Similarity
Image Similarity Viz

- Dendogram, flare dendogram
  - Excellent for showing cluster relationships as generated by tika-img-similarity
- Circle packing
  - What metadata features distinguish each cluster?
- Dynamic versions of each allow for interaction
- Future work
  - Integrating into Nutch administration GUI and allowing for Tika-based similarity and clustering
- The power of this approach: doesn’t require Computer Vision
Image Catalog ("ImageCat")

- OCR and EXIF metadata around images
- Can handle similarity measures
- Can allow for search of features in images based on text
- Can relate images based on EXIF properties (all taken with flash on; all taken in same geographic region, etc.)

- How do you do this at the scale of the Internet
  - "Deep Web" as defined by DARPA in domain of e.g., human trafficking ~60M web pages, 40M images
- You use ImageCatalog, of course! 😊
Image Catalog (“ImageCat”)

- Apache OODT – ETL, Map Reduce over LONG list of files
  - Partition files into 50k chunks
  - Ingest into Solr Extracting RequestHandler
- Apache Solr / Extracting RequestHandler
  - Augmented Tika + Tesseract OCR
  - Tika + EXIF metadata

[Diagram of Image Catalog]

- https://github.com/chrismattmann/imagecat/
ImageSpace

With ImageCat you can build... Image Space
https://github.com/memex-explorer/image_space/

Connect to ImageCat
Search on similar images
EXIF, Jaccard, computer vision based approaches
Continuum Analytics + Kitware, Inc. + JPL
Funded by DARPA Memex
So far, two semesters of projects
Fall 2014 and Spring 2015
Fall 2014
Crawl NASA AMD, NSF ACADIS and NSIDC
ADE
Bayesian MIME detection
Gridded Binary Image Parser

Angela started out exploring Apache OODT and Push Pull as a crawler, along with Apache Nutch. She found that Nutch was more configurable and easier to set up for crawling ACADIS and AMD.

She discovered while crawling and indexing in Solr that AMD and ACADIS had robots.txt file problems that prevented download of science data. One of the science data files present in AMD that Tika didn’t support and wouldn’t index was Grib (Gridded Binary) Files. So Vineet’s project was a Grib file parser in Tika.

She wanted to examine the MIME detector in Tika - he was wondering if the issue with parsing science data was that the MIME detector was incorrectly detecting it.

Prasanth wanted to treat the information provided from the glob file pattern, MIME magic, file name regular expression, and XML root chars as “evidence” in a Bayesian learning algorithm. He came up with the basic algorithm idea, and did some preliminary data gathering.
• Nutch REST API
• Drives crawling and eventually dataviz
• DataViz in D3 for image similarity
• Geo Topic Parser
• ML and NLP with GeoNames
• GCMD DIF, ISO19115 parsers
• Bayesian Detector
• Spark and Tika

Using Angela’s downloaded content from Nutch, and in Solr (shaded orange) and also her configuration in Github, Spring 2015 students have something to start with.

This includes Rishi Verma, who believes that using Apache Spark, an interactive analytic framework, will allow Tika to be run on large Polar data sets. He is investigating a Tika-based learning method for creating resilient distributed datasets.

Several students wanted to develop new parsers identified via crawler efforts. The first is Gautham Gowrishankar who takes on ISO 19139 parsing as it was not supported by Tika. Aakarsh Math took on Global Change Master Directory DIF file parsing. Yun Liu is using topic identification techniques and GeoNames to identify places and locations in text.
DARPA Memex started a new TREC “track” in Dynamic Domains

- http://trec-dd.org/
- Memex contributions + polar contributions
- Polar

- 1.7m URLs
- 158Gb, lots of images and data to work on

http://github.com/chrismattmann/trec-dd-polar/
Cop out: What about videos?

- I know I know
- Working on FFmpeg and Tika support for metadata
  - https://issues.apache.org/jira/browse/TIKA-1510
- Have someone working on similarity and deduplication methods for video (Michael Ryoo)
  - Pooled Motion for First Person Videos
  - Streaming video parser
    - https://issues.apache.org/jira/browse/TIKA-1598
Tie back to NASA / JPL

- Transition into Physical Oceanographic Distributed Active Archive Center (PO.DAAC) for images from satellites
- PolarCyberInfrastructure community
- Science images and videos for Mars
Thank you!

• Chris Mattmann
• @chrismattmann
• mattmann@apache.org

• http://memex.jpl.nasa.gov/
• http://trec-dd.org/

• http://nsf-polar-cyberinfrastructure.github.io/datavis-hackathon/
Chris A. Mattmann, NASA JPL, USC & the ASF

@chrismattmann
mattmann@apache.org