Evaluating Text Extraction: Developing a Toolkit for Apache Tika

ApacheCon NA 2015

Tim Allison
Paul M. Herceg
The MITRE Corporation
Overview

• Opening Notes of Gratitude
• Quick Overview on Tika – Tika on the Stack
• Motivation
• Exploratory Study: Tika 1.5 vs. Tika 1.7-SNAPSHOT
• Outcomes for Tika and the larger community
• Next steps and need help
• Thank you, Rackspace!

Public Service Announcements
Thank you, James Ahlborn, Brian O’Neill and others on Jackcess!

Recently, James Ahlborn, the current maintainer of Jackcess, kindly agreed to relicense Jackcess to Apache 2.0. Brian O’Neill, the CTO at Health Market Science, a LexisNexis® Company, also agreed with this relicensing and led the charge to obtain all necessary corporate approval to deliver a CCLA for Jackcess to Apache. As anyone who has tried to get corporate approval for anything knows, this can sometimes require not a small bit of effort.

If I may speak on behalf of Tika and the larger Apache community, I offer a sincere thanks to James, Brian and the other developers and contributors to Jackcess!!!
Thank you, Tyler Palsulich!
Quick Overview: Tika On the Stack
Overview of Tika

File Type Identification

PDF  MSOffice  HTML  JPEG  MP3  WAV  ...

Metadata and Text Content

Uniform Metadata and Text Content for text processing and/or ingestion into search engine
Overview of Tika

Welcome to Apache Tika version 1.6-SNAPSHOT!

To see what Tika can do, just drop a file or a URL to this window.
Use the View menu to switch views.

Informal, Quick and Technical Introduction to Lucene


Document

Fields User defines fields appropriate for a document (e.g. isbn, author, title, publication date) Store (yes/no) Index (analyzed, not analyzed)
Overview of Tika

geo:lat: 38.974833
geo:long: -77.018333
Altitude: 96 metres
exif:DateTimeOriginal: 2013-08-15T10:58:08
High Level Components of a Search Stack

TIKA ->

Content/Metadata Extraction

Indexer/Search System

User Interface
What the User Sees

User Interface

Indexer/Search System

Content/Metadata Extraction
Motivation

When Things Go Wrong with Text Extraction
When Things Go Wrong with a Foundation

W. Lloyd MacKenzie, via Flickr @http://www.flickr.com/photos/saffron_blaze/
When Things Go Wrong with Text Extraction

Taking a close look at the forest or open meadows reveals that there are often subtle differences in plant species across a wide landscape. Unique micro-climates, exposure to the sun, soil types, moisture availability, and a variety of other factors influence the types of plant species present in any given location. Changes in any of these factors will cause changes to

BGQOTM G IRUYK RUUQ GZ ZNK LUXKYZ UX UVKT SKGJU]Y XK\KGRY ZNGZ ZNKXK GXK ULZKT Y[HZRK JOLLKXKTIKY OT VRGTZ YVKIOKY GIXUYY G]OJK RGTJYGVK% CTOW[K SOIXU- IROSGZKY$ K^VUY[XK ZU ZNK Y[T$ YUOR Z_VKY$ SUOYZ][XK G\GORGHOROZ_ GTJ G \GXOKZ_ UL UZNKX LGIZUXY OTLR[KTIK ZNK Z_VKY UL VRGTZ YVKIOKY VXKYKTZ OT GT_MO\KT RUIGZOUT% 4NGTMKY OT GT_UL ZNKYK LGIZUXY JORR IG[YK INGTMKY ZU
When Things Go Wrong with Text Extraction

Statement

Seasoned professional with a skilled ability to connect co-workers and clients with the information, products and services they are seeking by utilizing professional experiences, organizational and client skills both as a team and an individual.

Experience

OLS: Office Liquidations Solutions

May 2010 – May 2013

Statement

OLS: Office Liquidations Solutions May 2010 – May 2013

Experience

Bialek Healthcare Environments June 2001 – May 2010

Bialek Healthcare Environments
Design Associate, Client Services Coordinator

Furniture bid package review, quotation, response and presentation. Small office design, space planning, and procurement, presentation and quoting for commercial systems and furniture.
When Things Go Wrong with Text Extraction

You don’t know what you can’t find
What Can Go Wrong

• Catastrophic failures
  • Out of Memory Errors
  • Infinite Hangs
  • Memory Leaks
• Exceptions: Null Pointer, etc.
• Extraction with loss of fidelity
  • Missing text/metadata/attachments
  • Garbled text
Public Service Announcement #1

Tika will break catastrophically. Very rarely, but it will.

- Out of Memory Errors
- Permanent Hangs
- Memory Leaks

Catastrophic problems happen very rarely. We fix problems when they’re identified; but they will happen. The only way to avoid these problems is to isolate Tika at the process level!!!
Unit Testing

• Tika-app (and dependencies): ~45 MB jar
• Tika: ~70k lines of code
• PDFBox: ~120k lines of code
• POI: ~312k lines of code

• Tika: ~400 test files for unit tests
• PDFBox: ~75 files
• POI: ~950 files
The Straw

```xml
--- tika/trunk/tika-parsers/pom.xml  2014/02/04 15:12:41  1564334
+++ tika/trunk/tika-parsers/pom.xml  2014/02/04 15:13:08  1564335
@@ -100,7 +100,7 @@
    <dependency>
        <groupId>org.apache.pdfbox</groupId>
        <artifactId>pdfbox</artifactId>
-       <version>1.8.3</version>
+       <version>1.8.4</version>
    </dependency>

<!-- TIKA-370: PDFBox declares the Bouncy Castle dependencies as optional, but we prefer to have them always to avoid -->

<table>
<thead>
<tr>
<th>jukka</th>
<th>818405</th>
</tr>
</thead>
<tbody>
<tr>
<td>jukka</td>
<td>769404</td>
</tr>
<tr>
<td>tallison</td>
<td>1564335</td>
</tr>
<tr>
<td>jukka</td>
<td>769404</td>
</tr>
</tbody>
</table>
```

```xml
    <dependency>
        <groupId>org.apache.pdfbox</groupId>
        <artifactId>pdfbox</artifactId>
        <version>1.8.4</version>
    </dependency>
```
The Straw and the Camel

Luis Filipe Nassif added a comment - 19/Feb/14 20:07 - edited

I also got this with Tika 1.5 on ~1500 pdf files from my base of 8500 pdf files, did not with 1.4.
When Things Go Wrong with Text Extraction (Coda)

- Search engines
- Email filters
- Parental control filters
- Accessibility software for the blind
- Smart phone hands-free applications
- Summarization tools
- Entity Extraction/Resolution
- Machine Translation

See: Herceg (2009), Herceg and Ball (2010) and Herceg and Ball (2011)
Comparing Tika 1.5 with Tika 1.7-SNAPSHOT
(vintage October, 2014)

Exploratory Study with govdocs1
Related Work

- Peter May’s **Batch File Id** – small batch file identification
- Lynn Marwood’s **File Type Id** – govdocs1 and different versions of Tika
- William Palmer’s **Tika to Ride** – challenges of running Tika within Hadoop
- William Palmer’s **github** site and PDFBOX-1757 – identifying extraction issues with govdocs1 docs

Research Evaluations (Metrics):
- Optical Character Recognition – character/word error rate
- Information retrieval – precision, recall, F-measure
- Machine Translation – **Bleu** (string similarity)
The Plan

• Find a corpus
• Run both versions of Tika against the corpus
• Compare the output of the two versions

• Goals
  • Discover and define an evaluation methodology
  • Identify and fix potential issues before the Tika 1.7 release
Find a Corpus: **govdocs1**

• Nearly 1 million documents gathered from *.gov in 2009

<table>
<thead>
<tr>
<th>File Extension</th>
<th>Number of Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>pdf</td>
<td>231,009</td>
</tr>
<tr>
<td>html</td>
<td>214,264</td>
</tr>
<tr>
<td>jpg</td>
<td>109,094</td>
</tr>
<tr>
<td>txt</td>
<td>78,178</td>
</tr>
<tr>
<td>doc</td>
<td>76,507</td>
</tr>
<tr>
<td>xls</td>
<td>62,577</td>
</tr>
<tr>
<td>ppt</td>
<td>49,600</td>
</tr>
<tr>
<td>gif</td>
<td>36,279</td>
</tr>
<tr>
<td>xml</td>
<td>33,451</td>
</tr>
<tr>
<td>ps</td>
<td>22,012</td>
</tr>
</tbody>
</table>

Find a Corpus: **govdocs1**

- Nearly 1 million documents gathered from *.gov in 2009

<table>
<thead>
<tr>
<th>File Extension</th>
<th>Number of Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>pdf</td>
<td>231,009</td>
</tr>
<tr>
<td>html</td>
<td>214,264</td>
</tr>
<tr>
<td>jpg</td>
<td>109,094</td>
</tr>
<tr>
<td>txt</td>
<td>78,178</td>
</tr>
<tr>
<td>doc</td>
<td>76,507</td>
</tr>
<tr>
<td>xls</td>
<td>62,577</td>
</tr>
<tr>
<td>ppt</td>
<td>49,600</td>
</tr>
<tr>
<td>gif</td>
<td>36,279</td>
</tr>
<tr>
<td>xml</td>
<td>33,451</td>
</tr>
<tr>
<td>ps</td>
<td>22,012</td>
</tr>
</tbody>
</table>

**Known Limitations**

- Mostly monolingual
- Aging – 215 pptx, 163 docx and 37 xlsx

Can always use more formats...

But this is a great resource!

---

Run Both Versions of Tika against the Corpus

• What we found/knew
  • Single threaded not so fast
  • Rare but catastrophic failures
  • Out-of-the-box Tika formats didn’t maintain metadata from embedded documents
Standard Legacy Output

<?xml version="1.0" encoding="UTF-8"?>
<meta .../>
...
<div class="package-entry">
<h1>embed4.txt</h1>
<p>embed_4</p>
</div>
...

Keep the Embedded Metadata (app’s -J or server’s /rmeta)

```json
{
"Content-Length": "7",
"Content-Type": "text/plain; charset=ISO-8859-1",
"Last-Save-Date": "2014-06-04T01:09:10Z",
"X-Parsed-By": [
"org.apache.tika.parser.DefaultParser",
"org.apache.tika.parser.txt.TXTParser"
],
"X-TIKA:content": "embed_4
",
"X-TIKA:embedded_resource_path":
  "embedded-1/embed1.zip/
   embed2.zip/embed3.zip/embed4.zip/embed4.txt",
}
```
Compare the Output of the Two Versions: Basic

• Counts (by file type) and pairwise comparisons
  • Catastrophic errors
  • Exceptions
  • Attachments
  • Metadata values
  • Unique tokens
  • Total tokens

• Identified file type

• Language id
## Comparison of Exception Counts

<table>
<thead>
<tr>
<th>File Extension</th>
<th>Tika 1.5</th>
<th>Tika 1.7-SNAPSHOT</th>
<th>Percentage of Exceptions by File Type with Tika 1.7-SNAPSHOT</th>
</tr>
</thead>
<tbody>
<tr>
<td>xls</td>
<td>2,824</td>
<td>2,828</td>
<td>4.52%</td>
</tr>
<tr>
<td>log</td>
<td>1,253</td>
<td>1,253</td>
<td>12.56%</td>
</tr>
<tr>
<td>ppt</td>
<td>2,195</td>
<td>1,191</td>
<td>2.40%</td>
</tr>
<tr>
<td>doc</td>
<td>847</td>
<td>795</td>
<td>1.04%</td>
</tr>
<tr>
<td>pdf</td>
<td>644</td>
<td>123</td>
<td>0.05%</td>
</tr>
<tr>
<td>xml</td>
<td>417</td>
<td>417</td>
<td>1.25%</td>
</tr>
<tr>
<td>html</td>
<td>161</td>
<td>161</td>
<td>0.08%</td>
</tr>
<tr>
<td>pps</td>
<td>28</td>
<td>8</td>
<td>0.49%</td>
</tr>
<tr>
<td>unk</td>
<td>20</td>
<td>18</td>
<td>0.35%</td>
</tr>
<tr>
<td>kml</td>
<td>19</td>
<td>19</td>
<td>1.91%</td>
</tr>
</tbody>
</table>
Limits to Simple Exception Counting

- Unsupported file type/version exceptions
- Encrypted/Access Permission exceptions
- Exception percolation – embedded document exception reported for container file type
- “Fixed” exception could yield junk text

- Still not counting: “Extraction with Loss of Fidelity”
25% of exceptions in our study were from the XML Parser trying to parse non-compliant XML

Consider configuring HtmlParser for XML files.

Currently hard-coded into tika-server, but it is not the default with the regular Tika parser.
Compare the Output of the Two Versions: Content Differences

- **Dice Coefficient** (on unique tokens)
  - Doc A: foo foo foo bar bar bat bat bat bat
  - Doc B: foo bar bar baz
  - \( \frac{2*(\text{foo bar})}{(\text{foo bar bat} + \text{foo bar baz})} = \frac{4}{6} = 66\% \)

- **Dice Coefficient** (on total tokens)
  - \( \frac{2*(\text{foo bar bar})}{4x(\text{foo}) + 4x(\text{bar}) + 4x(\text{bat}) + \text{baz}} = \frac{6}{13} = 46\% \)

- Identified 618 PDFs, 101 XLS files and 95 java files that met the threshold for differences in content

- Manual review of a selection of documents
  - PDFs (worse), XLS (better), java (better)
Lessons Brought Home

• Nick’s theorem on Tika Exceptions:
  “1% of a lot is still a lot…”

• Proposed Theorem #2:
  “A small problem for me could be a large problem for you.”
Outcomes
Outcomes: Modifications to Tika

• Numerous bug fixes and added capability to parse old XLS 3 and XLS 4 files (thank you, Nick Burch!)
• Parser wrapper with JSON serialization to maintain metadata of embedded documents, available in Tika 1.7 (thank you, Jukka Zitting and Nick Burch!)
• tika-batch: new module in Tika 1.8
• tika-eval: module in development
tika-eval: Single Batch

• Stack traces and exception counts
• File id, language id
• Attachment and metadata counts
• Top 10 most common words
• Content length
• Token length statistics
• Token entropy
tika-eval: Compare Two Batches

- The same as Single Batch, side by side
- Content Similarity (unigram tokens)
  - Dice Coefficient (unique tokens, tokens)
- Top 10 most common tokens unique to A vs. B and vice versa
- Top 10 tokens with the greatest difference in counts between A and B
Current State of Eval Reports
Different applications of A vs. B

• A vs. ground truth
• Different configurations
• Different software versions
• Different tools
• Single vs. Multithreaded
• Different files (partial, w/o filenames)
Outcomes: Community Involvement

Thank you, Tilman Hausherr!
Moving Ongoing Evaluation to the Community: **TIKA-1302**

- **Completed**
  - Set up Rackspace vm with govdocs1
  - Staged ~250GB compressed slice of Common Crawl from Julien Nioche
  - Received 3GB of NSF Polar data from Chris Mattmann

- **Planned**
  - (re)Publish corpora/input data
  - Publish extracted content including stack traces
  - Publish results of comparisons
  - Set up fairly regular runs of regression testing

- **Farther down the road**
  - Integrate Tika with monthly **Common Crawl** (?)
Help Needed

• Issues and patches – please keep them coming!
• User interface for tika-eval
• Statistics/methods to help identify junk output (language/format agnostic) ([TIKA-1443](#))
• More data
• More participants in the public evaluations
Thank you!

• Questions?

Contact info:

Tim Allison and Paul M. Herceg
tallison@apache.org