Evaluating Text Extraction: Developing a Toolkit for Apache Tika

ApacheCon NA 2015

Tim Allison
Paul M. Herceg
The MITRE Corporation

MITRE

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!

The Apache Software Foundation
http://www.apache.org/

Dashboards Projects More Create

Tika / TIKA-1601

Integrate Jackcess to handle MSAccess files

Edit Comment Assign More Resolve Issue Close Issue

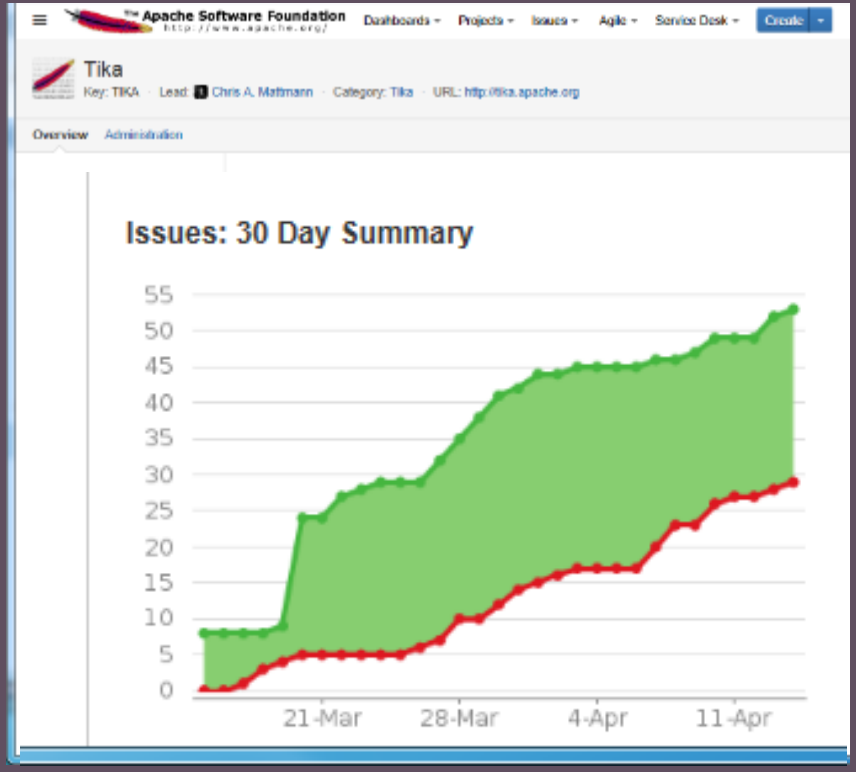
Details

Description

Recently, James Ahlborn, the current maintainer of Jackcess, kindly agreed to relicense Jackcess to Apache 2.0. Brian ONeill, 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.

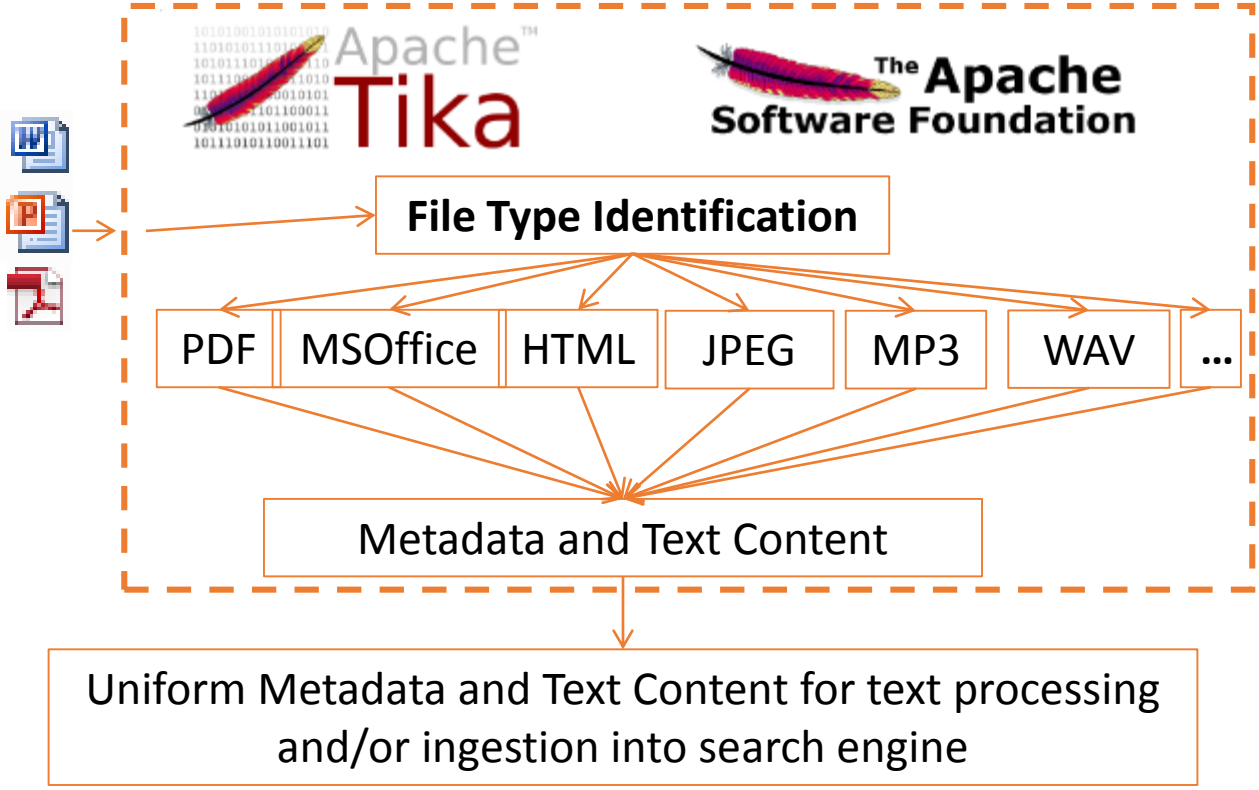
If I may speak on behalf of Tika and the larger Apache community, I offer a sincere thanks to James, Brian and the other developers and contributors to Jackcess!!!

Thank you, Tyler Palsulich!

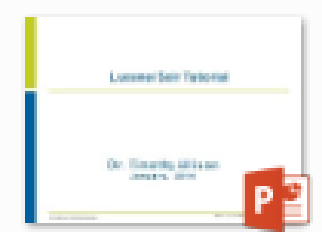


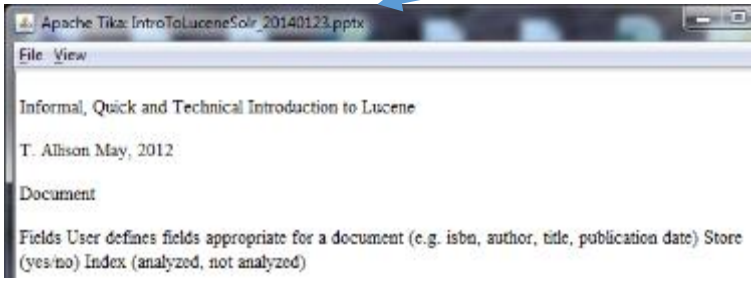
Quick Overview: Tika On the Stack

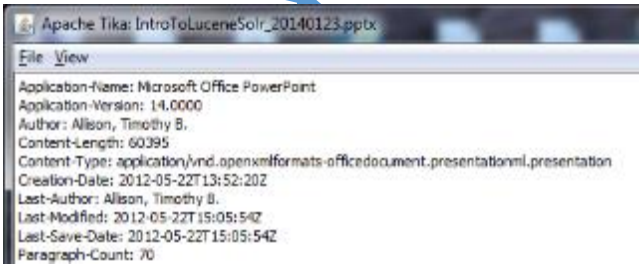
Overview of Tika



Overview of Tika



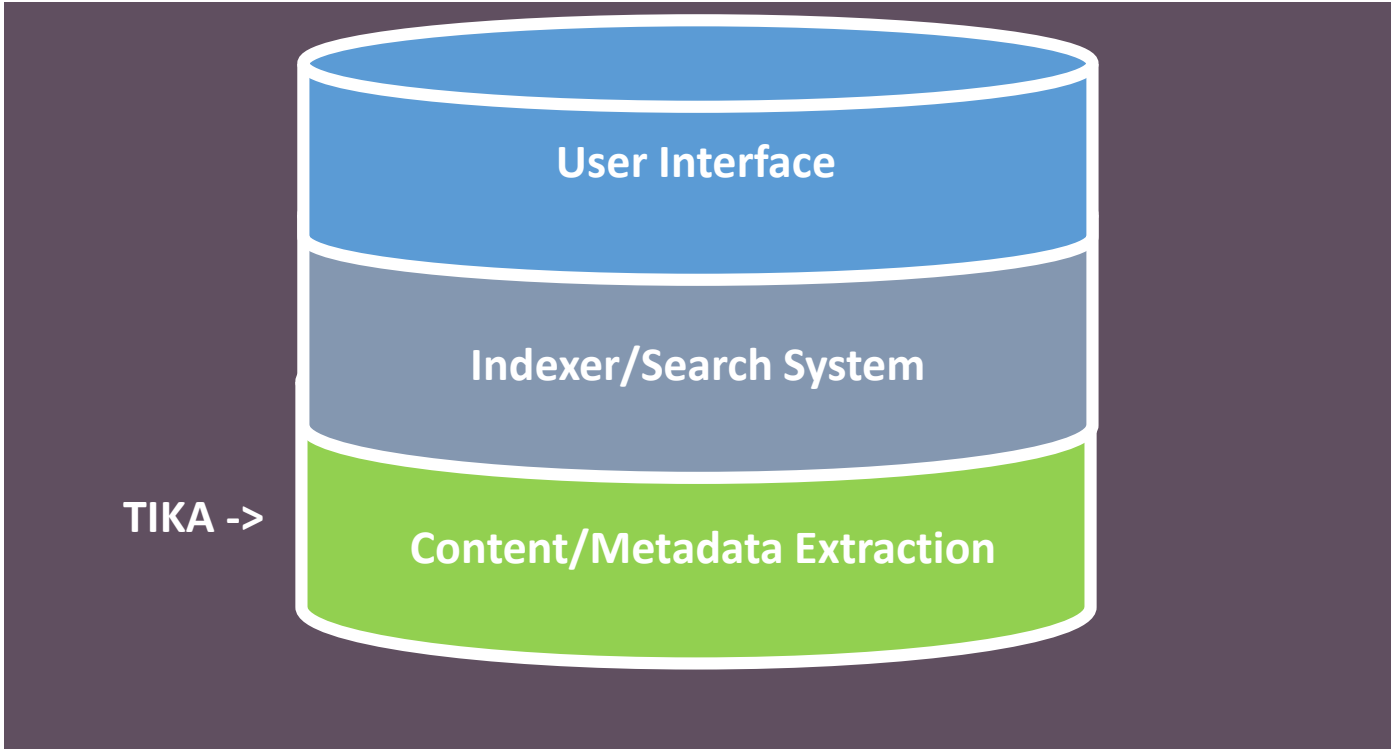




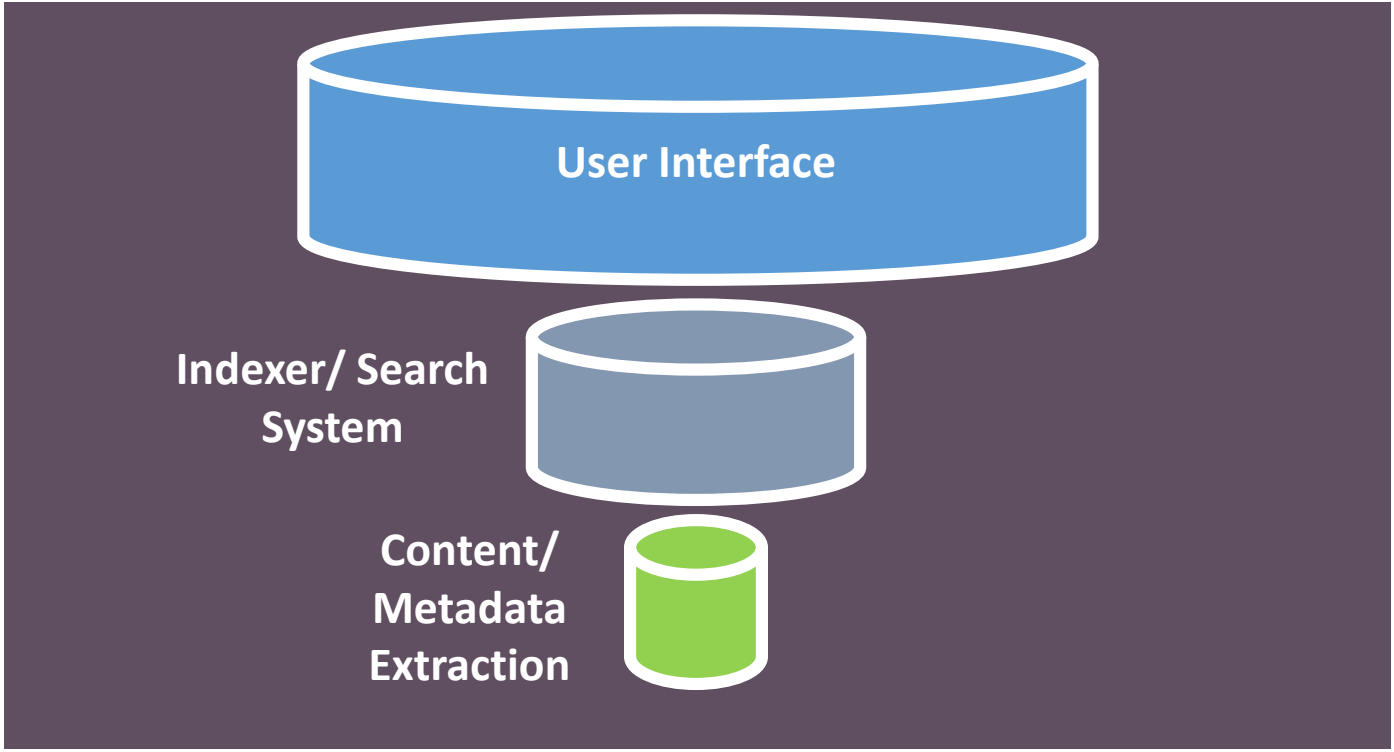
Overview of Tika

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



What the User Sees



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 RGTJYIGVK% 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]ORR 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

Experience

OLS: Office Liquidations Solutions May 2010 – May 2013

Bialek Healthcare Environments June 2001 – May 2010

Bialek Healthcare Environments June 2001 – May 2010

Design Associate, Client Services Coordinator

Furniture bid package review, quotation, response and presentation. Small office design, space

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

```
--- 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
```

		`<dependency>`
jukka	818405	`<groupId>org.apache.pdfbox</groupId>`
jukka	769404	`<artifactId>pdfbox</artifactId>`
tallison	1564335	`<version>1.8.4</version>`
jukka	769404	`</dependency>`

The Straw and the Camel

Tika / TIKA-1233

PDFBox can throw StringIndexOutOfBoundsException on some dates

Comment | Agile Board | More | Reopen Issue

Details

Type:	Bug	Status:	CLOSED
Priority:	Trivial	Resolution:	Fixed
Affects Version/s:	1.5	Fix Version/s:	1.6

People

Assignee: Unassigned

Reporter:

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.

Until PDFBOX-1803 is resolved, we can add an extra catch to prevent this from causing problems in TIKA

```
@@ -171,6 +171,9 @@
            addMetadata(metadata, TikaCoreProperties.CREATED, info.getCreationDate());
        } catch (IOException e) {
            // Invalid date format, just ignore
```

Dates

Created: 05/Feb/14 12:07

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

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

Garfinkel, S., Farrell P., Roussev, V., and Dinolt, G. (2009). Bringing science to digital forensics with standardized forensic corpora. *Digital Investigation*, 6, S2-S11.

Find a Corpus: govdocs1

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

Known Limitations

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

Can always use more formats…
But this is a great resource!

Garfinkel, S., Farrell P., Roussev, V., and Dinolt, G. (2009). Bringing science to digital forensics with standardized forensic corpora. *Digital Investigation*, 6, S2-S11.

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)


```
[
…
{
"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\n",
    "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

File Extension	Tika 1.5	Tika 1.7-SNAPSHOT	Percentage of Exceptions by File Type with Tika 1.7-SNAPSHOT
xls	2,824	2,828	4.52%
log	1,253	1,253	12.56%
ppt	2,195	1,191	2.40%
doc	847	795	1.04%
pdf	644	123	0.05%
xml	417	417	1.25%
html	161	161	0.08%
pps	28	8	0.49%
unk	20	18	0.35%
kml	19	19	1.91%

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”

Public Service Announcement #2

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
  - 2*(foo bar)/(foo bar bat+foo bar baz) = 4/6 = 66%
- Dice Coefficient (on total tokens)
  - 2*(foo bar bar)/4x(foo)+4x(bar)+4x(bat)+baz = 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

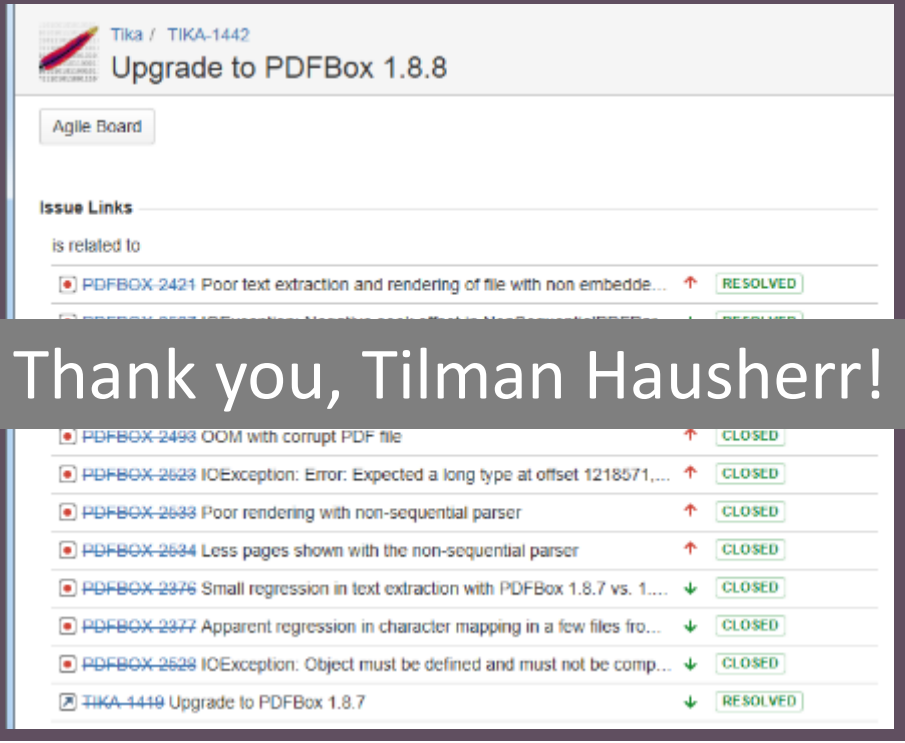
- The same as Single Batch, side by side
- Content Similarity (unigram tokens)
  - Dice Coefficient (unique tokens, tokens)
- Top 10 most common tokens unique to A vs. B and vice versa
- Top 10 tokens with the greatest difference in counts between A and B

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

MITRE