



Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

Evaluating Text Extraction At Scale: A Case Study from Apache Tika

Tim Allison, Ph.D.

Data Scientist/Relevance Engineer

Artificial Intelligence, Analytics and Innovative
Development Organization(1740)

ITSD

The research was carried out at the NASA (National Aeronautics and Space Administration) Jet Propulsion Laboratory, California Institute of Technology under a contract with the Defense Advanced Research Projects Agency (DARPA) SafeDocs program. © 2020 California Institute of Technology. Government sponsorship acknowledged.



Jet Propulsion Laboratory
California Institute of Technology

About me

- Data scientist (files and search) Jet Propulsion Laboratory, California Institute of Technology
- Chair/V.P. Apache Tika
- Committer Apache PDFBox, POI, Lucene/Solr, OpenNLP
- Member Apache Software Foundation

The research was carried out at the NASA (National Aeronautics and Space Administration) Jet Propulsion Laboratory, California Institute of Technology under a contract with the Defense Advanced Research Projects Agency (DARPA) SafeDocs program.

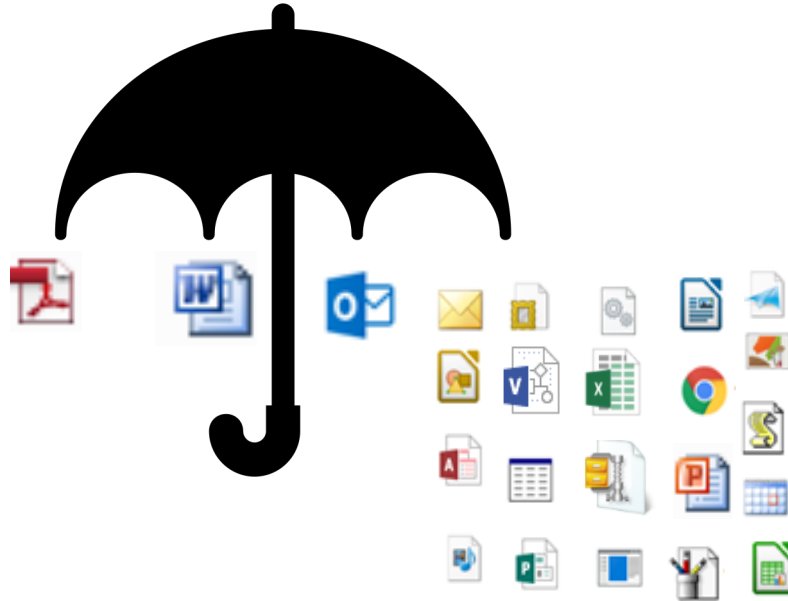
© 2020 California Institute of Technology. Government sponsorship acknowledged.

Outline

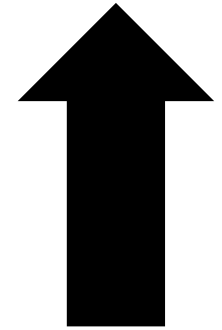
- Apache Tika, an overview
- What can possibly go wrong?
- tika-eval module
- Evaluation at scale and public corpora

Apache Tika, an overview

Framework for file
type detection,
parsing and uniform
output for ~75
parsers, ~100+
formats



text and metadata



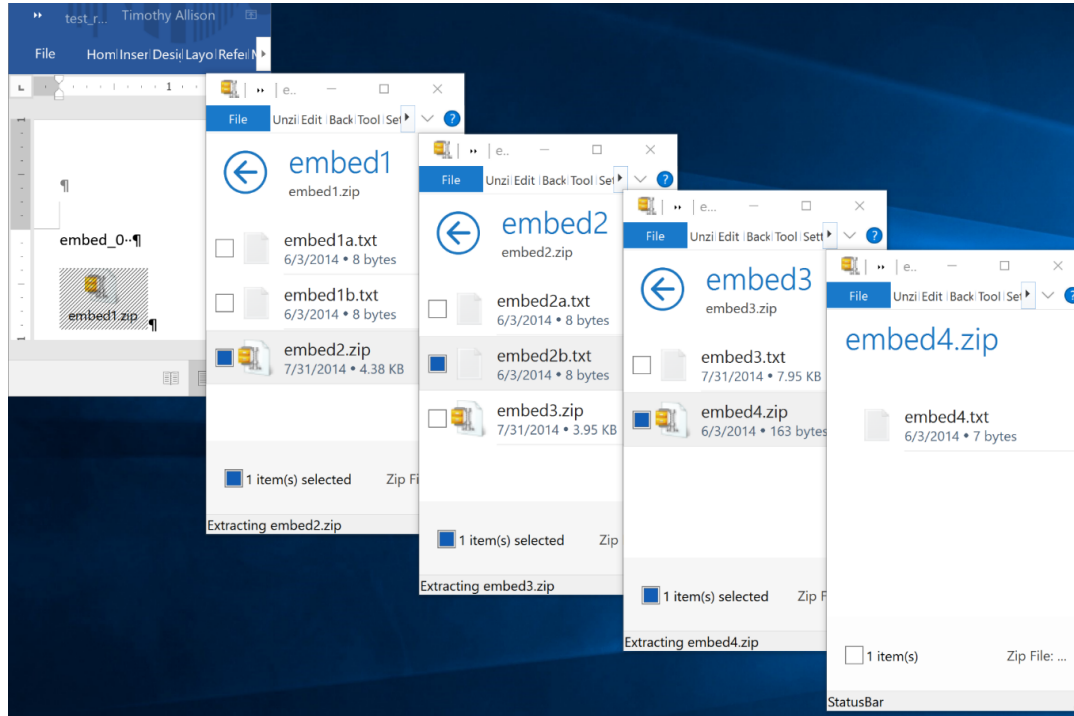
bytes

<https://tika.apache.org/>

Apache Tika, features

- Easy to add new file types for detection
- Easy to add new parsers
- Works recursively with embedded files/attachments
- Integration with tesseract-ocr

Embedded files/attachments



Embedded files extracted

```
[ {
  "Application-Name": "Microsoft Office Word",
  "Content-Length": "27082",
  "Content-Type": "application/....wordprocessingml.document",
  "X-TIKA:content": "embed_0 ",
  ... },
{
  "Content-Type": "text/plain; charset=ISO-8859-1",
  "Last-Modified": "2014-06-04T04:08:28Z",
  "X-TIKA:content": "embed_1a",
  "X-TIKA:embedded_resource_path": "/embed1.zip/embed1a.txt",
  ... },
{
  "Content-Type": "application/zip",
  "Last-Modified": "2014-06-04T04:09:40Z",
  "X-TIKA:content": "embed4.txt",
  "X-TIKA:embedded_resource_path":
"/embed1.zip/embed2.zip/embed3.zip/embed4.zip"
  ...}, ...]
```

From PDF to ...

Apache Tika - Apache Tika

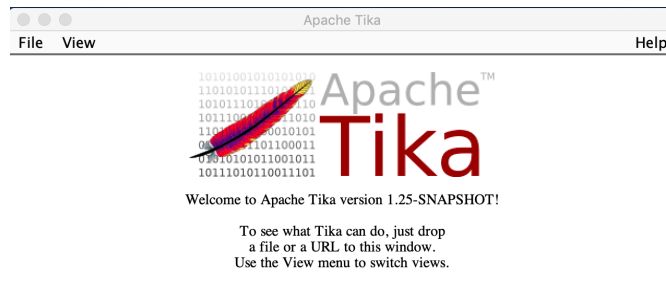
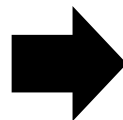
<http://incubator.apache.org/tika/>

Tika - Content Analysis Toolkit

Apache Tika is a toolkit for detecting and extracting metadata and structured text content from various documents using existing parser libraries.

Apache Tika is an effort undergoing [incubation](#) at [The Apache Software Foundation \(ASF\)](#), sponsored by the [Apache Lucene PMC](#). Incubation is required of all newly accepted projects until a further review indicates that the infrastructure, communications, and decision making process have stabilized in a manner consistent with other successful ASF projects. While incubation status is not necessarily a reflection of the completeness or stability of the code, it does indicate that the project has yet to be fully endorsed by the ASF.

See the [Apache Tika Incubation Status](#) page for the current incubation status.

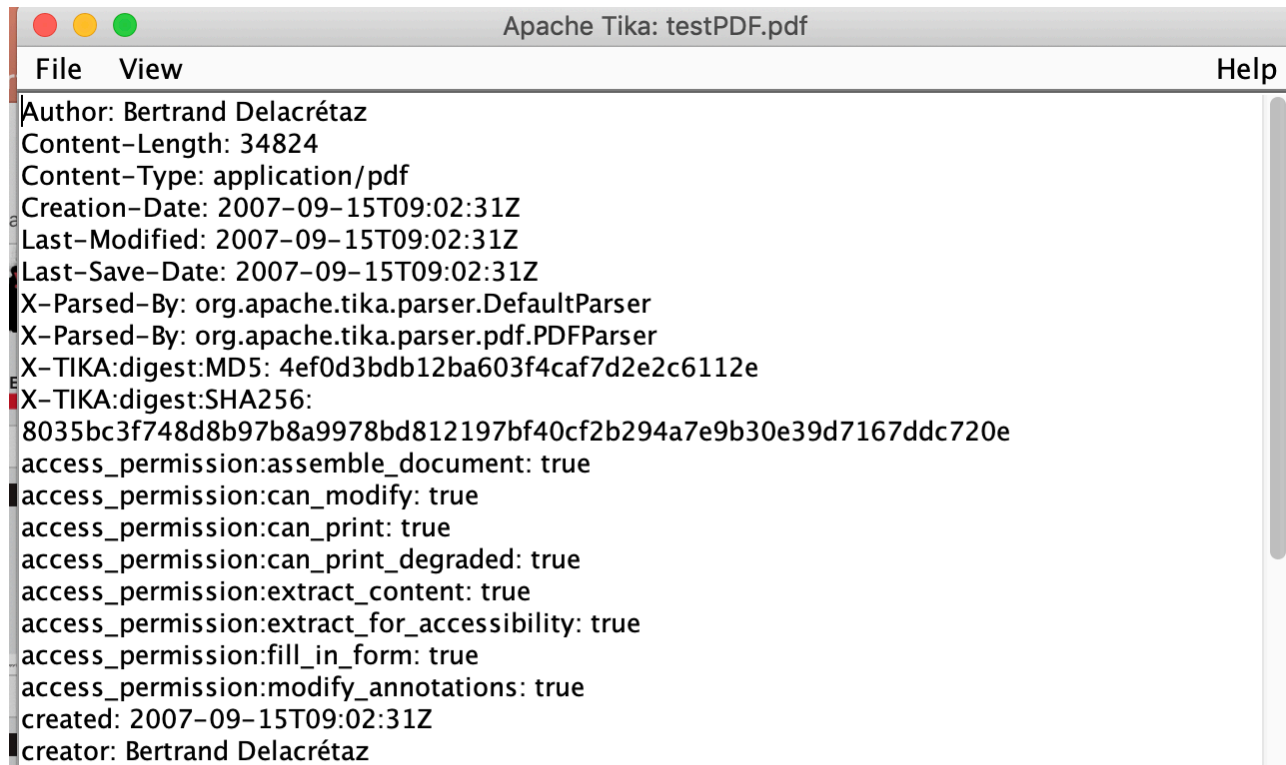


Latest News

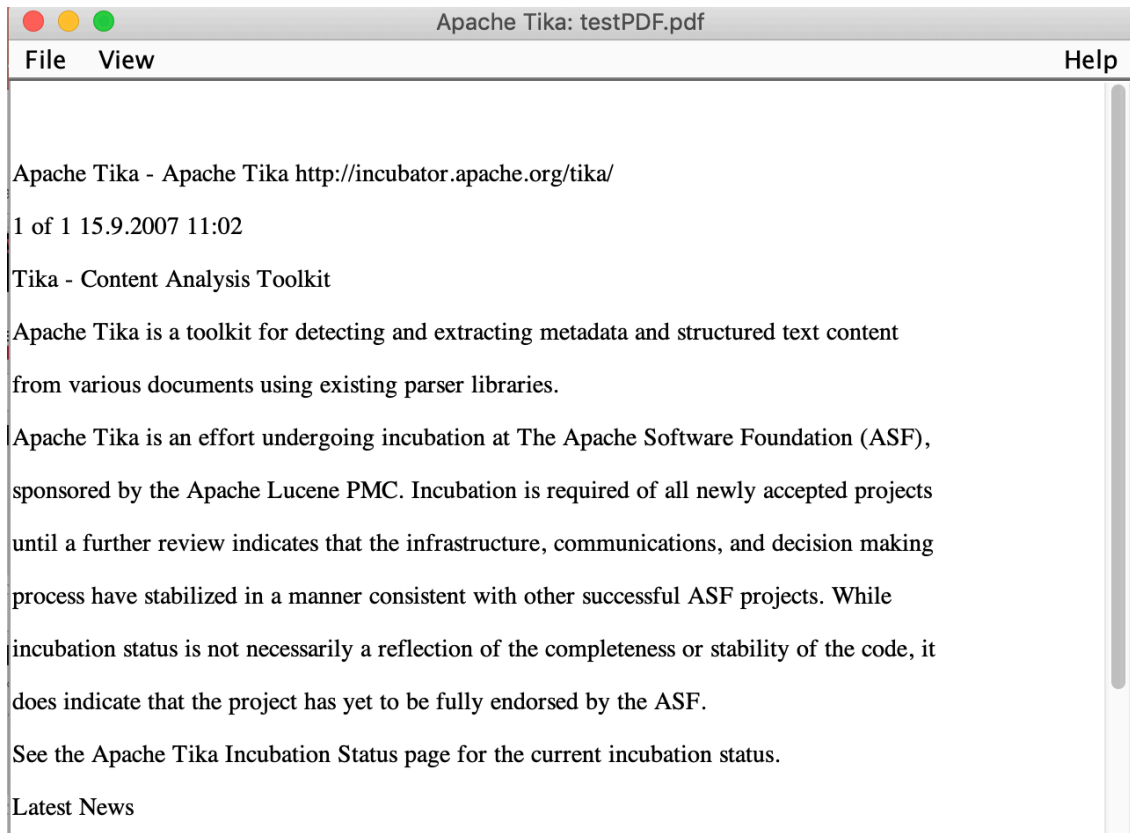
March 22nd, 2007: Apache Tika project started

The Apache Tika project was formally started when the [Tika proposal](#) was [accepted](#) by the [Apache Incubator PMC](#).

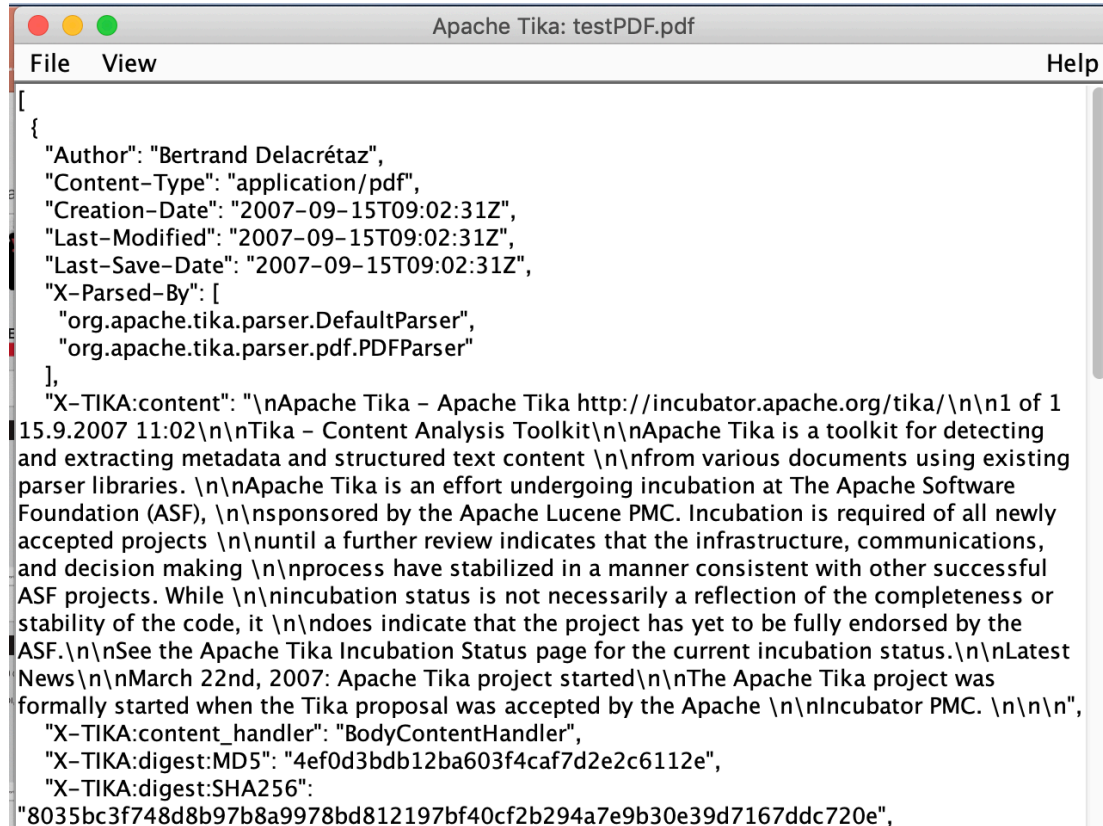
Metadata



Text



Metadata and Content in JSON



The screenshot shows a web browser window titled "Apache Tika: testPDF.pdf". The browser has a menu bar with "File", "View", and "Help". The main content area displays a JSON object representing the metadata of a PDF file. The metadata includes fields for Author, Content-Type, Creation-Date, Last-Modified, Last-Save-Date, and X-Parsed-By. The X-TIKA:content field contains the full text of the document, which is a news article about the Apache Tika project. The X-TIKA:content_handler, X-TIKA:digest:MD5, and X-TIKA:digest:SHA256 fields are also present.

```
[
  {
    "Author": "Bertrand Delacr  taz",
    "Content-Type": "application/pdf",
    "Creation-Date": "2007-09-15T09:02:31Z",
    "Last-Modified": "2007-09-15T09:02:31Z",
    "Last-Save-Date": "2007-09-15T09:02:31Z",
    "X-Parsed-By": [
      "org.apache.tika.parser.DefaultParser",
      "org.apache.tika.parser.pdf.PDFParser"
    ],
    "X-TIKA:content": "\nApache Tika - Apache Tika http://incubator.apache.org/tika/\n\n1 of 1
15.9.2007 11:02\n\nTika - Content Analysis Toolkit\n\nApache Tika is a toolkit for detecting
and extracting metadata and structured text content \n\nfrom various documents using existing
parser libraries. \n\nApache Tika is an effort undergoing incubation at The Apache Software
Foundation (ASF), \n\nsponsored by the Apache Lucene PMC. Incubation is required of all newly
accepted projects \n\nuntil a further review indicates that the infrastructure, communications,
and decision making \n\nprocess have stabilized in a manner consistent with other successful
ASF projects. While \n\nincubation status is not necessarily a reflection of the completeness or
stability of the code, it \n\ndoes indicate that the project has yet to be fully endorsed by the
ASF.\n\nSee the Apache Tika Incubation Status page for the current incubation status.\n\nLatest
News\n\nMarch 22nd, 2007: Apache Tika project started\n\nThe Apache Tika project was
formally started when the Tika proposal was accepted by the Apache \n\nIncubator PMC. \n\n\n",
    "X-TIKA:content_handler": "BodyContentHandler",
    "X-TIKA:digest:MD5": "4ef0d3bdb12ba603f4caf7d2e2c6112e",
    "X-TIKA:digest:SHA256":
"8035bc3f748d8b97b8a9978bd812197bf40cf2b294a7e9b30e39d7167ddc720e",
```

What could possibly go wrong?

- Basic
 - Thrown Exceptions – corrupt files, bad parser?
 - Image only – need to have mechanism to determine when to run OCR
 - Password protected files
- Catastrophic
 - Infinite loops/out of memory/seg fault
 - Security compromises

What could subtly go wrong?

- Missing text
- Missing attachments
- Garbled text

Missing Text

Jane Coady

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
Sales and Project Administrator

Sales support and sales. Lead generation and follow up. Developed solutions for individual projects. Determine price schedules, budgets and profit margins. Created and streamlined forms and procedures. Located project specific furniture. Project Management. Plan and coordinate work schedules and duties for employees, freight companies and customers. Space planning/placement of systems furniture inventories into client's AutoCAD drawings with Giso. Coordinate project details and schedules with General Contractors, Building Engineers and Property Managers. Attend company meetings to exchange product information and coordinate work activities with other departments. Keep records and create reports regarding purchases, sales, bids and installation schedules. Coordinate marketing campaigns by compiling lists, marketing pieces to promote inventories. Inventory management. Resolve customer questions regarding sales, service and installations.

Bialek Healthcare Environments June 2001 – May 2010
Design Associate, Client Services Coordinator

Furniture bid package review, quotation, response and presentation. Small office design, space planning, need assessment, presentation and quotation for commercial systems and freestanding furniture. Maintenance of client accounts including need assessment, quotation, order processing, purchasing, job costing, tracking and invoicing. Created streamlined procedures to reduce redundancies. Employee Training. Member of various committees including Process Streamlining, Marketing, and Fun.

Rhosymedre Design Group August 1998 – April 2001
Office Manager

Processing and maintenance of accounts receivable, payable and payroll with Business Works Accounting System and QuickBooks Pro. Maintenance of client accounts including estimating, job costing, purchasing, tracking, and invoicing and project management. Establish and maintain vendor relations. Research new residential products.

Education University of Nebraska August 1984 – May 1987
Bachelors of Science with a focus in Textiles, Clothing and Interior Design, with a minor in Business
Honors: Gold Key Honorary Jan 1986, Sigma Phi Upsilon Honorary Officer – October 1985

Jane Coady

Statement

OLS: Office Liquidations Solutions May 2010 – May 2013

Experience

Bialek Healthcare Environments June 2001 – May 2010

University of Nebraska August 1984 – May 1987

Education

Bachelors of Science with a focus in Textiles, Clothing and Interior Design, with a minor in Business

JC

2

Skills

File publicly available: <https://issues.apache.org/jira/browse/TIKA-1130>

Garbled Text

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

Upgrade from
Apache PDFBox
1.8.6 to 1.8.7



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\GORGHORZ_\$GTJ G \GXOKZ_ UL UZNXK 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

tika-eval

- Library and commandline tool to calculate these stats
- Profile a single batch of text extracts
- Compare two batches of extracts
 - Different tools
 - Different settings
- Coming soon – multi-compare

<https://cwiki.apache.org/confluence/display/TIKA/TikaEval>

Out of vocabulary (OOV) – Same file, different Tika

	Tika 1.14	Tika 1.15-SNAPSHOT
Unique Tokens	786	156
Total Tokens	1603	272
LangId	zh-ch	de
Common Words	0	116
Alphabetic Tokens	1603	250
Top N Tokens	拙故: 18 獐档: 14 略獐: 14 m: 11 柿溪: 11 瑶拙: 11 畚柿: 11 档溪: 10 捌敦: 9 敲沫: 9	die: 11 und: 8 von: 8 deutschen: 7 deutsche: 6 1: 5 das: 5 der: 5 finanzministerium: 5 oder: 5
OOV%	$1-(0/1603) = 100\%$	$1-(116/250) = 54\%$

Overlap: 0%

Increase in Common Words: 116

Out of vocabulary (OOV) – Same file, different Tika

	Tika 1.14	Tika 1.15-SNAPSHOT
Unique Tokens	1916	1995
Total Tokens	14187	14302
LangId	en	en
Common Words	7498	7409
Alphabetic Tokens	13472	13587
Top 10 Unique Tokens	applicant's: 8 1.69: 1 arbitrary: 1 collecting: 1 constitution: 1 e112: 1 ei.b: 1 equating: 1 magnetically: 1 o: 1	ss: 106 applicantis: 8 ssss: 7 iactsi: 4 ithe: 4 imeansi: 3 iprocessi: 3 calculations.i: 2 iabstract: 2 idata: 2
OOV%	$1-(7498/13472) = 44\%$	$1-(7409/13587) = 45\%$

Overlap: 95.5%

Increase in Common Words: -89

OOV/Common Tokens corpus wide

Comparing the HTMLDefault encoding detector with Mozilla's Universal Encoding Detector on HTML encoding detection

HTMLDefault	Universal	HTMLDefault Sum Common Tokens	Universal Sum Common Tokens	Difference in Sums
UTF-8	EUC-JP	4,437	481,919	477,482
EUC-JP	Shift_JIS	1,512	391,126	389,614
UTF-16	windows-1252	1,240	368,496	367,256
UTF-16	UTF-8	2,563	321,717	319,154
EUC-JP	UTF-8	764,957	1,047,029	282,072

10/20/20



From analytics to action

Constrained Least Squares Linear Spectral Unmixture by the Hybrid Steepest Descent Method

Nobuhiko Ogura^{*} and Isao Yamada^{**}

1 Introduction

A closed polyhedron is the intersection of finite number of closed half spaces, i.e., the set of points satisfying finite number of linear inequalities, and is widely used as a constraint in various application, for example specifications or constraints in signal processing or estimation problems, resource restrictions in financial applications and feasible sets of probability distributions. By the progress of the convex analysis and the fixed point theory of nonexpansive mapping, a number of convex projection based algorithms are proposed (for example, Bauschke et al, 1997; Combettes, 1993; Yamada et al, 1998–2002).

https://aviris.jpl.nasa.gov/proceedings/workshops/02_docs/2002_Ogura_1_web.pdf

Stored text vs. Optical Character Recognition

Text As Stored in File

```
!"#$%& (') *,+-. ' / 0 1,23 *. 457698;::<=>=75?&@78;ACB  
D(B7E;FHGJICBK5MLNBKOPBKF;B DJD Q R S.TVU9WNXYMY[Z\T]^W_S `badc  
5KICedFgfh5 cji ;edF;A^5KEk<>ImIn;;e[<>EnloedACICe a  
lo<p57Eg5Kqsr;E;<jloe[E 8;O 6hedA5Kq adc 57ItedFk;;B c qslCf;B a
```

Text from Tesseract OCR

Constrained Least Squares Linear Spectral Unmixture by the Hybrid Steepest Descent Method

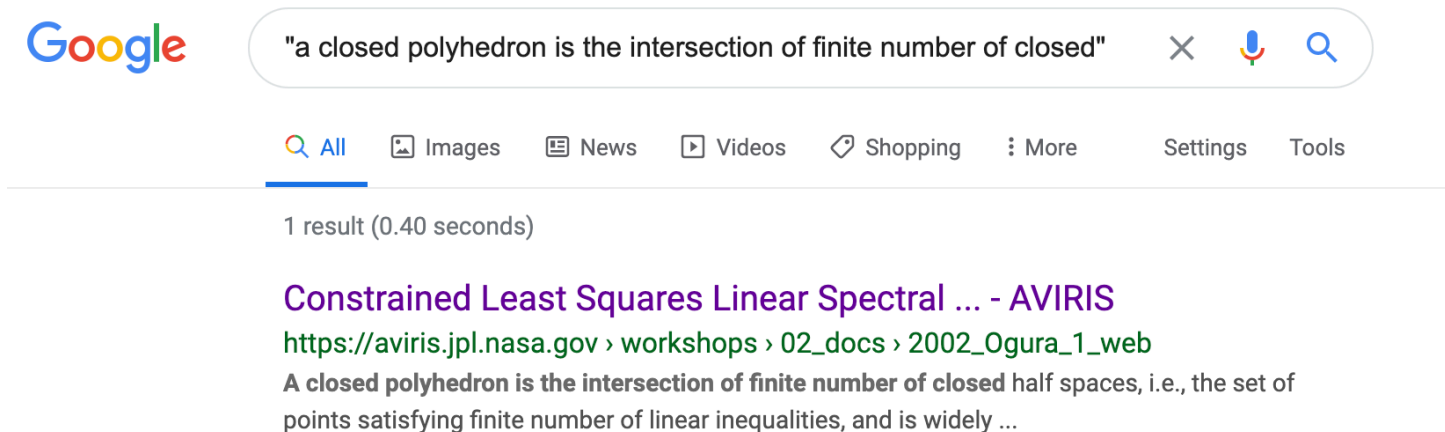
Nobuhiko Ogura' and Isao Yamada"

1 Introduction

A closed polyhedron is the intersection of finite number of closed half spaces, i.e., the set of points satisfying finite number of linear inequalities, and is widely used as a constraint in various applications, for example specifications or constraints in signal processing or estimation problems, resource restrictions in financial applications and feasible sets of

What would Google do?

Text in Google's Cache



Popat, Ashok. (2009). A panlingual anomalous text detector. 201-204.
10.1145/1600193.1600237.

Author: Ashok C. Popat Google, Inc., Mountain View, CA, USA

Uses

- Parser developers
- Production – search, Natural Language Processing, etc...
 - Which parsers/extract to use?
 - Which encoding detector to trust?
 - When to run OCR?
 - Do not index/relevance weights

Other indicators

- Low token/page ratio
- Language identification – Catalan in a largely English corpus?!
- Median/mean word length and stdev
- Number of Unicode code blocks
- Alphabetic/non-alphabetic

Limitations in absence of ground truth

- More exceptions – We have a problem! Wait...
 - New parser, we were entirely skipping those file types before
 - Parser was yielding junk before on this file, now it is letting us know there's a problem
- Fewer exceptions – Great! Wait...
 - Mime detection not working – skipping files that we used to parse (theoretical)
 - Now we're getting junk

Limitations in absence of ground truth

- More Common Words – Great! Wait...
 - Serious bug that duplicates worksheets in some xlsx files (TIKA-2356...my fault...ugh!)
- Fewer Common Words – Problem! Wait...
- More attachments, fewer attachments (Your turn!)

Multi-compare example...in the works

A	B	C	D	E	F	G
file	mutool_1_17_0	pdftotext_0_62_0	pdftotext_3_03	tika_1_24_1	tika_1_10_0	TotalSets
file1	set_0	set_1	set_2	set_3	set_0	4
file2	set_0	set_1	set_2	set_3	set_0	4
file3	set_0	set_1	set_1	set_2	set_3	4

file1

- Set 0: 0 bytes
- Set 1: HÖll World!
- Set 2: H'11 World!
- Set 3: HÖllÿ World!

Taking tika-eval public

- Maruan Sahyoun kindly hosts a vm for ongoing evals (TIKA-1302)
- 1 TB (~3 million files) from govdocs1, Common Crawl and bug trackers
- Collaborating with Apache PDFBox and Apache POI to run evals as part of the release process






Taking tika-eval public

- Critical to identifying regressions and building new parsers
- Stacktraces created by public documents are critical for the `hey-I'm-getting-this parse-exception-but-can't-share-the-document-with-you` problem

Corpora

- Govdocs1 (<https://digitalcorpora.org/corpora/files>)
- Common Crawl (<https://commoncrawl.org/>)
- Bug trackers (see Peter Wyatt's article: <https://www.pdfa.org/a-new-stressful-pdf-corpus/>)
- Coming soon(?): Unit-test files from github parser projects

Corpora

<u>Name</u>	<u>Last modified</u>	<u>Size</u>	<u>Description</u>
 Parent Directory		-	
 bug_trackers/	2020-09-04 16:46	-	
 commoncrawl3/	2020-06-23 16:35	-	
 commoncrawl3_refetched/	2020-06-04 14:52	-	
 govdocs1/	2020-06-23 16:35	-	

<https://corpora.tika.apache.org/base/docs/>

Bug-trackers

https://corpora.tika.apache.org/base/docs/bug_trackers/



Please join the fun!

- <https://tika.apache.org/>
- <https://cwiki.apache.org/confluence/display/TIKA/TikaEval>
- corpora-dev@tika.apache.org
- <https://issues.apache.org/jira/projects/TIKA>
- @ApacheTika

Questions?

timothy.b.allison@jpl.nasa.gov
@_tallison



Jet Propulsion Laboratory
California Institute of Technology

jpl.nasa.gov