Apache Tika

From Wikipedia, the free encyclopedia

Apache Tika is a content detection and analysis framework, written in Java, stewarded at the Apache Software Foundation.[1] It detects and extracts metadata and text from over a thousand different file types, and as well as providing a Java library, has Server and Command Line editions suitable for use from other programming languages.

Tika

Developer(s)	Apache Software Foundation
Stable release	1.13 / May 16, 2016
Development status	Active
Written in	Java
Operating system	Cross-platform
Type	Search and index API
License	Apache License 2.0
Website	tika.apache.org (http://tika.apache.org/)

Contents

History

The project originated as part of the Apache Nutch codebase, to provide content identification and extraction when crawling. In 2007, it was separated out, to make it more extensible and usable by Content management systems, other Web crawlers, and information retrieval systems. The standalone Tika was founded by Jérôme Charron, Chris Mattmann and Jukka Zitting.[2] In 2011 Chris Mattmann and Jukka Zitting released the Manning book "Tika in Action", and the project released version 1.0.

Features

Tika provides capabilities for identification of more than 1400 file types from the Internet Assigned Numbers Authority taxonomy of MIME types. For most of the more common and popular formats,[3] Tika then provides content extraction, metadata extraction and language identification capabilities.

While Tika is written in Java, it is widely used from other languages.[4] The RESTful server and CLI Tool permit non-Java programs to access the Tika functionality.

Notable uses

Tika is used by financial institutions including the Fair Isaac Corporation (FICO),[5] by NASA and academic researchers[6] by major content management systems including Drupal,[7] and Alfresco (software)[8] to analyze large amounts of content, and to make it available in common formats using information retrieval techniques.

On April 4, 2016[9] Forbes published an article identifying Tika as one of the key technologies used by more than 400 journalists to analyze 11.5 million leaked documents that expose an international scandal involving world leaders storing money in offshore Shell corporations. The leaked documents and the project to analyze them is referred to as the Panama Papers.

References

1. "Apache Tika". Retrieved 2016-04-15.
2. "Tika Proposal". Retrieved 2016-04-15.
3. *Apache Tika formats page* http://tika.apache.org/1.12/formats.html. Retrieved 16 April 2016. Missing or empty `|title=` (help)
4. "API Bindings for Tika". Apache Tika. Retrieved 2016-04-17.
5. "FICO to Engage Kaggle's Community of 180,000 Data Scientists to Drive Innovation in the FICO Analytic Cloud | FICO®". *FICO® | Decisions*. Retrieved 2016-04-15.
6. "Studying polar data with the help of Apache Tika". *Opensource.com*. Retrieved 2016-04-15.
7. "Text Extract for Drupal using Tika | Drupal.org". *www.drupal.org*. Retrieved 2016-04-15.
8. "Content Transformation and Metadata Extraction with Apache Tika - alfrescowiki". *wiki.alfresco.com*. Retrieved 2016-04-15.
9. Fox-Brewster, Thomas. "From Encrypted Drives To Amazon's Cloud -- The Amazing Flight Of The Panama Papers". *Forbes*. Retrieved 2016-04-15.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Apache_Tika&oldid=732794003"

Categories: Apache Software Foundation | Java platform | Free software programmed in Java (programming language) | Java (programming language) libraries | Software using the Apache license

---

- This page was last modified on 3 August 2016, at 08:34.
-