

Apache Tika is a framework for content type detection and content extraction which was designed by Apache software foundation. It detects and extracts metadata and structured text content from different types of documents such as spreadsheets, text documents, images or PDFs including audio or video input formats to certain extent.