Apache Tika - a content analysis toolkit

The Apache Tika™ toolkit detects and extracts metadata and text ...