

Apache Tika - a content analysis toolkit