Apache Tika

Um toolkit para análise de conteúdo

O que é

O tika consegue extrair meta dados e texto de mais de 1000 tipos de arquivos diferentes

Todos os tipos de arquivos podem ser passados por uma única interface, o que a torna útil para motores de busca fazerem indexação, para análise de conteúdo e tradução em larga escala

Quem usa

- **Goldman Sachs -** Corretora de investimentos e seguradora
- **Drupal -** Sistema de gerenciamento de conteúdo
- **Alfresco -** Softwares de gerenciamento de informação
- **Pesquisadores Acadêmicos**

Exemplos de uso

- A Goldman Sanches precisava pesquisar por cláusulas específicas em milhares de documentos em papel. Para resolver isso, eles digitalizaram e usaram o Tika para procurar os termos nos textos extraídos.

- PhD em Geografia com foco em sensoriamento remoto por satélite de gelo e neve usou Tika para extrair informações de inúmeros portais e repositórios sobre o assunto online

Principais softwares concorrentes

KFileMetaData - biblioteca em C++ que extrai texto e meta dados de diversos formatos de arquivos

DocSplit - Extrai texto de documentos e PDF’s