REPORT

Homework: Content extraction and search using Apache Tika.

Dataset: *Vault of Size 13GB approx 2067 pdf files*

Observation:

When we did a Search on keyword "UFO" on the site FBI vault site http://vault.fbi.gov we got 23 files found with matching keyword.

In our Apache Tika project: Using Tika parser and list of keywords which are synonyms and closely related to our key search "UFO" we get 212 files in the vault.

The reason for discrepancy between the number of documents you found as compared to the reported figure of 23 could be that since the dataset is collected over long duration of time and some dated back to 1930s-1940s, so those days UFO term was not coined, hence where referred to as "Ghost Objects", Flying "Objects" etc, hence the FBI site fails to parse the content of file and look for other similar keywords, and another reason for such large difference in count could be the parsed content (OCR quality) from the vault data set and probably the search was based on filename.

Apache Tika

A content analysis toolkit which was every easy to detects and extracts metadata and structured text content from various documents using existing parser libraries SAX parser. In our assignment we used PDFParser which was very handy to parse pdf contents and one of striking features we noticed was structured content and streamed parsing which doesn't allow to keep the full document content in memory or spooled to disk, thus allowing application to deal with large data set that has to be parsed with minimum resource requirement.