My Observations while working on CS 572 Apache Tika Project

About Dataset

The dataset present in the FBI vault was scanned and then converted into PDF. The dataset was huge and consisted of over 2000 files from FBI. Most of the files were a digital version of declassified documents that have been scanned from paper and made available via a digital content management system. These documents consisted of scans of newspaper, confidential papers/reports from FBI.

While searching for the keywords in the dataset it was observed since the OCR quality of scan was not perfect there were many conversion from scan to text that were misspelled. There were few occurrences where an alphabet is rendered as a random special character. So, even if the word was ***ufo*** in the document it was rendered as ***ueo*** or ***ufo*** which may be undetected by a program while searching for ***ufo***. Also there was random suffix/prefix appended to words in the file. The file became seemingly impossible and difficult to read and interpret accurately by a human reader.

Did corpus really only contains 269 documents about UFOs ?

The number of 269 documents was present in the corpus of 3000 documents. However in the corpus which we were given contained 2067 files of which 239 files were found to be having a keyword related to UFO. This number could have been more if all files were read properly. Since the number of related files is 11.5 % while FBI mentioned this figure to be 9% we can deduce the corpus have more info about UFOs than FBI was telling.

Reasons for the discrepancy between numbers of documents found as compared to the reported figure of 269

The reasons must be:

1. The corpus given to us has only around 70% of the files while the figure of 269 is for the corpus with 100% of the files.
2. Other reasons might be improper OCR conversion from FBI files/reports.
3. There might be a reason where a word has an alphabet converted wrong and thus we could not find exact match.
4. Since we are searching only keyword which does not exist in between alphabet characters, it might be a reason two words are concatenated during OCR and thus our word becomes undetectable in the search.
5. We are ignoring different forms of a keyword like ***ufology*** for ***ufo*** is ignored in our search.
6. Around 800 files from 2067 were not read properly having content length less than 1000.

My thoughts About Apace TIKA

It's a really cool API and had made searching in a corpus very easy. Four lines of code and you could start parsing your document. As a developer this was one of the easiest API I have used ever. Probably a usage scenario like extracting metadata from a file using TIKA looks fun to me and a probable usage in a future web app I am planning to work over as well.

Vasu Jain vasujain@usc.edu 5019-3028-37

Extra Keywords Used

These are the extra keywords I used in keywords file for search in the PDF corpus:

- **UFOB**
- **UAP**
- **OVNI**
- **Project Blue Book**
- **Operation Majestic**
- **mystery airships**
- **extraterrestrial**
- **unidentified flying object**
- **ghost rockets**
- **Super Cosmos**
- **aerial phenomena**
- **extra sensory perception**
- **foo fighters**
- **alien spacecraft**