# TEAM\_02\_CSCI599\_HW\_BIGDATA\_Report

## Input File: British UFO Files

## Output file: TSV\_OCR\_v2.tsv

## Data Preparation and Cleaning

We started with running the provided ocr-pipeline.sh bash script (<https://gist.github.com/chrismattmann/a5031c317bad35ca30cec7b9decd51a5)> on the British UFO pdf files. We modified the script and used some options of image magick to fix errors like ‘spp not in set {1,3,4}’ using ‘-background white -alpha Off’ option and improve the OCR quality for files which were not giving any text output. We did border removal by using -border , -fuzz, -trim options.

The most important pages were “Reports for Unidentified Objects”, from where we extracted the information about UFO sightings, like Date, Time and Duration of Sightings, Description of object, How was it observed, What were the nearby objects or the weather conditions when the observation was made or Were there any other witnesses to the particular UFO sightings.

Observations made about the extracted dataset:

1. While implementing OCR on the pdf files, most of the data we initially received was improper and most of it was just garbled text.
2. Since the original documents are old and little bit tattered, some the scanned pages have low visibility due to faded text and ink blots. So no useful data is being extracted from them.
3. Some of the records are written in cursive, which Tesseract is unable to read.
4. We considered only those pages in the PDF which contains the keywords “REPORT OF UNIDENTIFIED OBJECT”

### Parsing OCR extracted data into TSV:

We used the following British UFO files dataset: <https://www.dropbox.com/sh/bwzhuhigz222rwr/AADNTCqrTdtD78sXdWrEHUxsa?dl=0>

There are 8 British UFO pdf files. The pages of interest in the pdf files contained the keywords “FLYING OBJECT OR AERIAL”. We ignored the pages which did not contain the mentioned keywords. Since the extracted text from OCR scanning contained few garbled characters, we decided to apply regex to look for the following keywords:

1. Flying or aerial (if page is of interest)
2. Date & Sighting (for Date of Sighting)
3. Mins, Secs, Hours, Still there (for Duration)
4. Description (for Shape)
5. Position (for Location)
6. How Observed + Direction (for Description)
7. Receipt (for Date of Report)

We followed the following steps to extract data from scanned PDF files and parse them into TSV:

1. We split each PDF file into individual pdf files with one page per file.
2. We converted the image into its respective tiff format using imagemagick library. Since the scanned PDFs have low visibility and have a lot of noise, we applied few image enhancement features of imagemagick like noise reduction, darkening the text and redefining the borders.
3. We then used Tesseract to extract the data into corresponding text files. Since the output text had different kind of spelling errors, garbled text –we tried to use autocorrect lib from python for correcting spelling mistakes and enchant lib from python to identify if a word belongs to en-US dictionary. we could not fix them as two or more words combined and we had no way to identify them as distinct words. We had to manually edit some of the text files to correct the keywords we are looking for.

#### Insights from the joined dataset:

1. What we noticed about the dataset?

The UFO Sightings dataset provided by British UFO files are properly formatted reports which have more detailed description than the original dataset we received for Assignment 1. The details like what nearby objects could have been misunderstood as a UFO or what Meteorological conditions could have triggered the observation, can actually provide us with hints whether the observation was true or was just a deception.

1. What questions did the newly joined answer about UFO sightings previously unanswered?
2. How well did the image captions accurately describe the UFO object types? What about the identified objects in the image?
3. How well did the OCR work? What did we do to clean up the noise in the data?

Due to a lot of noise in the original British UFO files, some of the pages in PDFs produced garbled texts when OCR was operated on them to extract text. It was because most of the characters, even though were typewritten, had low visibility, overlapped stamped ink or had some dirt on the pages because the scanned files are very old. To clean up the noise, in order to extract a better readable data, we played around with few features of magick (imagemagick). For example- we redefined the border color and border thickness, we defined the density of the generated image to maximum 300 dpi level, we introduced fuzz for color matching upto 20% and trim to trim the borders of the identical color as the corners of an image.