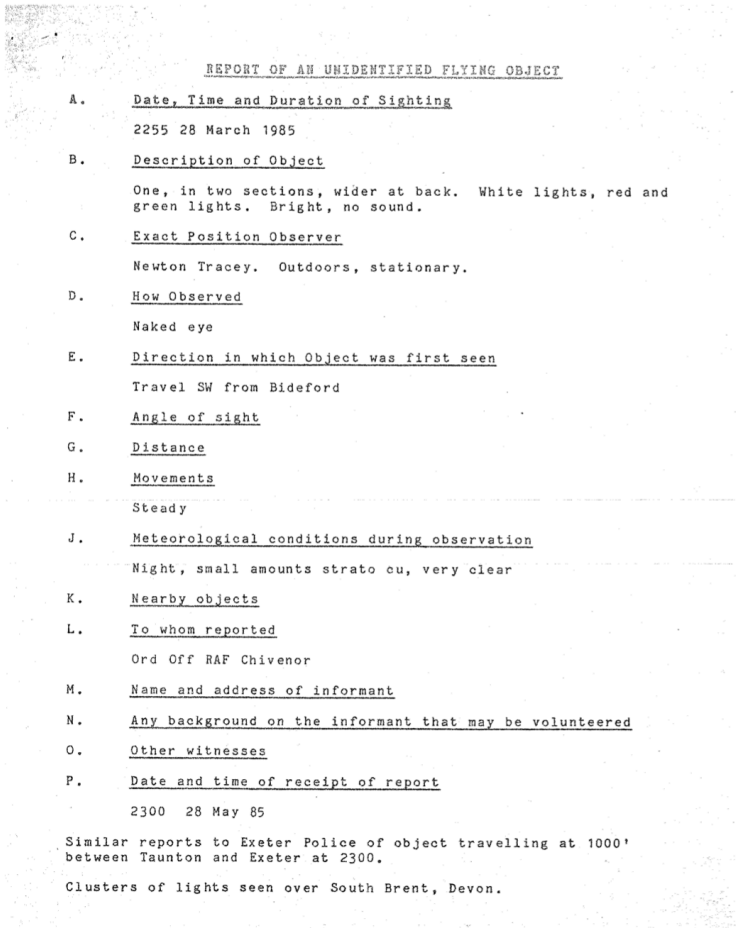
Homework: Large Scale OCR Extraction, Image Captioning and Object Recognition and Enrichment of UFO Sightings Data
Due: Friday, April 13, 2018 12pm PT

---

1. Overview



REPORT OF AN UNIDENTIFIED FLYING OBJECT

A. Date, Time and Duration of Sighting

2255 28 March 1985

B. Description of Object

One, in two sections, wider at back. White lights, red and green lights. Bright, no sound.

C. Exact Position Observer

Newton Tracey. Outdoors, stationary.

D. How Observed

Naked eye

E. Direction in which Object was first seen

Travel SW from Bideford

F. Angle of sight

G. Distance

H. Movements

Steady

J. Meteorological conditions during observation

Night, small amounts strato cu, very clear

K. Nearby objects

L. To whom reported

Ord Off RAF Chivenor

M. Name and address of informant

N. Any background on the informant that may be volunteered

O. Other witnesses

P. Date and time of receipt of report

2300 28 May 85

Similar reports to Exeter Police of object travelling at 1000' between Taunton and Exeter at 2300.

Clusters of lights seen over South Brent, Devon.

Figure 1: example of sighting from British UFO files dump.

In this second assignment, we will build off of the 60,000+ UFO ("unidentified flying objects") sightings data set that you mined and added features to in Assignment 1 in a handful of ways. First, we will perform some advanced extractions providing real world experience using Optical Character Recognition (OCR) with Apache Tika, and Tesseract, which are two state of the art technologies in the area. We will leverage OCR to integrate another extremely important dataset of sightings into your TSV sightings database. You are going to integrate the British Ministry of Defence's UFO Sightings data release into your TSV database. The only major problem is that the data release, as described in (http://www.theblackvault.com/documentarchive/united-kingdom-ufo-documents/) is 100s of pages of scanned PDF files. It's your job to take these sightings, some handwritten, some written in cursive, some typed, on image scans of various quality, color, black and white, etc., and to integrate these sightings into the data model from your

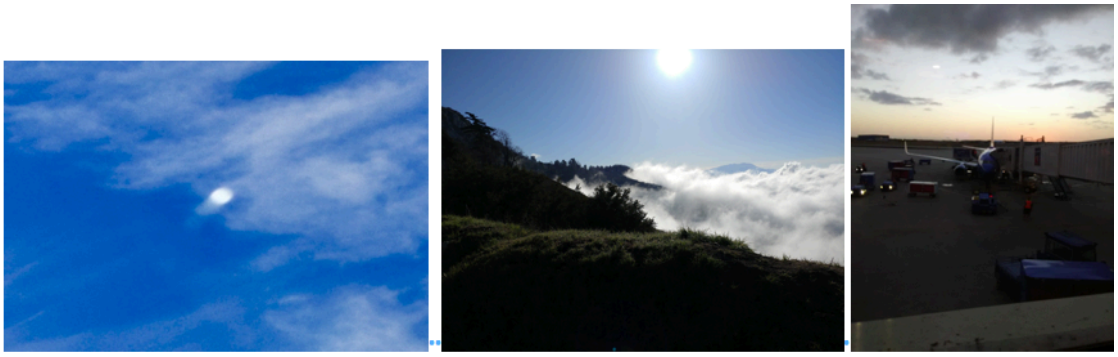
first assignment, in particular, to add, for each sighting in the British UFO files, the following fields into your TSV dataset:

1. Date of Sighting
2. Date of Report
3. Location
4. Shape (orb, light, etc.)
5. Duration
6. Description (Text)

The final tranche of UFO files released by The National Archives contain a wide range of UFO-related documents, drawings, letters, and photos and parliamentary questions covering the final two years of the Ministry of Defence's UFO Desk (from late 2007 until November 2009). It is certainly possible that a handful of these fields will not be directly available from the OCR. For example, location may not be directly discernable, date of report may only be available (but not sighting date, etc.) We expect that the scan data will be noisy, so you will do your best and we may supplement some of the extractions with manually curated "labels" and "fixes" to the data. Welcome to data cleaning in the real world! An example of such a UFO sighting from the British UFO files dataset is shown in Figure 1.

In addition to the sighting data you will also apply another advanced extraction technique in order to combine multimedia (image/*) data to your dataset. As opposed to the first assignment where this was difficult to do automatically since we had not covered the particular lecture topics, in this assignment we will leverage two easy to use Tika Docker files to identify objects present in an image and to generate a textual (human readable) caption for the image. Both of these Docker Files are available in Apache Tika and they leverage Machine Learning and Deep Learning extraction techniques that we have discussed in class and in particular Google's Tensorflow technology and custom Deep Learning models built in the USC IRDS group. You can see some examples of the Image Captioning and Image Object identification in action below in Figure 2a-c showing 3 automatically generated labels (with only generic training) run on UFO sightings images from the UFO stalker dataset http://www.ufostalker.com/tag/photo – which includes a collection of 4,047 sightings that have one or more images associated with them. We will integrate this dataset of sightings into your TSV file as well and add two new features related to identified objects (a list of them); and a generated image text caption.

Figure 2: a) a light/orb shown in the daylight; b) an orb present against a mountain background; and c) an orb in a cloudy sky.



Machine Generated Labels for a). b). c)

a plane flying in the sky over a field	*a view of a mountain range with a mountain in the background*	*an airplane is parked on the tarmac at an airport*

The integration of these two datasets will allow you to apply knowledge gained from the Parsing/Extraction Lecture, the lectures on advanced extraction (including Deep Learning and OCR), and also topics discussed including Large Scale Content Extraction. In particular, please consider techniques discussed in class to embark on this assignment.

2. Objective

The objective of the assignment is to apply both Optical Character Recognition (OCR) techniques to automatically extract and parse information from noisy image scans and to integrate the British Ministry of Defence Data; a treasure trove of additional UFO sightings and in addition to actually integrate image/* data into your sightings dataset as well by featurizing images and identifying objects in them and generating human readable text captions automatically using Deep Learning.

To perform OCR you will use Tika's automatic integration with Tesseract (https://wiki.apache.org/tika/TikaOCR). Tesseract is a powerful open source statistical OCR toolkit powered up with contributions from IBM and Google over many years. It works on image scans of varying languages and types. You will first leverage the OCR capabilities of Tika and Tesseract as described on the aforementioned Tika wiki page, and then in turn you will develop a Sightings Parser capability to take the OCR'ed text and to extract out sightings from it. We will provide you a default OCR pipeline that gives you a start on cleaning and preparing the data, and then you will use Tika and develop some of your own scripting to turn the extracted text into rows of sightings to add to your TSV file.

Secondly, you will leverage two Tika Docker images, the first (https://wiki.apache.org/tika/TikaAndVision) allowing you to leverage Tensorflow & Inception v4 – built from the Google ImageNet corpus – to automatically identify objects in UFO sightings images associated with the UFO stalker dataset. You will install the Docker and obtain a Tika REST service that you can use to iterate over and generate lists of objects identified in the image. Then you will perform a similar task and run the Tika Image Captioning Docker (https://wiki.apache.org/tika/ImageCaption) – also that leverages Inception-v4, and the USC IRDS Neural Image Caption Generator (see: https://github.com/USCDataScience/img2text) based on the Google Show & Tell paper: https://arxiv.org/abs/1411.4555. The extracted image captions you will see are trained against ImageNet, a rich corpus of 10M images and associated labels from WordNet. However, one thing that you quickly observe is that ImageNet is very broad and not particularly specific for UFO sightings and as such some of the captions and/or labels generated may be augmented with some refinement which we will specify later in the assignment.

You will generate in this assignment a new TSV dataset which includes joined British UFO files that you extract through your OCR pipeline, and in addition, will include

joined Image UFO sightings data via your Image Object identification and Captioning pipeline. Please generate a version 2 TSV file that includes this joined data.

The assignment specific tasks will be specified in the following section.

3. Tasks

1. Generate a copy of your TSV v1 dataset. Call it "v2" or something similar. You will add your new British UFO file sightings and UFO stalker Image based sightings to this dataset.
2. Download the British UFO files from the Dropbox. We have subset the British UFO sightings so you don't have to look at all of them.
    a. https://www.dropbox.com/sh/bwzhuhigz222rwr/AADNTCqrTdtD78sXdWrEHUxsa?dl=0
    b. Once downloaded, have a look at this Github GIST to get started on preparing the data to be extracted into sightings
        i. https://gist.github.com/chrismattmann/a5031c317bad35ca30cec7b9decd51a5
        ii. The basic OCR pipeline script requires the following dependencies
        iii. Image Magick (https://www.imagemagick.org/script/download.php) you can install with brew also brew install imagemagick
        iv. Poppler (https://poppler.freedesktop.org/) you can install with Brew (brew install poppler)
        v. Read this article once you have installed the dependencies
            1. http://kiirani.com/2013/03/22/tesseract-pdf.html
    c. Once you have read the GIST, and looked at the output in the outtxt/* directories, you will see that some of the text is garbled, some didn't come out right, etc.
        i. Explore this – could changing things with ImageMagick, convert, etc. improve the OCR? Image Orientation? Handwriting?
        ii. Develop some scripts that take the text and attempt to clean it some more
    d. Develop and deliver scripts that take outtxt/* and convert the OCR into rows in your v2 TSV dataset from #1.
        i. You are strongly encouraged to consider writing a Tika Parser and integrating it into your pipeline, see: http://tika.apache.org/1.16/parser_guide.html
        ii. If you don't write a Tika parser, identify in your README why you didn't and make sure you deliver your scripts.
3. Install Tika Dockers package for Image Captioning and Object Recognition
    a. git clone https://github.com/USCDataScience/tika-dockers.git
    b. Read and test out: https://wiki.apache.org/tika/TikaAndVision
    c. Read and test out: https://wiki.apache.org/tika/ImageCaption
    d. Write a Python program that uses Selenium (see: http://selenium-python.readthedocs.io/) to scrape down the images from the UFO stalker dataset: http://www.ufostalker.com/tag/photo

e. Run the Object identification from 3b and Image Captioning (3c) web services and generate objects identified and a text image caption for each of the images in 3d.
f. Format the extracted objects (as a simple list e.g., [object1,object2,object3]) and text caption as CSV sightings rows and add them to your V2 TSV dataset.

4. (**EXTRA CREDIT**) Try out TikaNER on text associated with your TSV v2 sightings
    a. See: https://wiki.apache.org/tika/FrontPage#Named_Entity_Recognition_.28NER.29_support
    b. Try out OpenNLP, CoreNLP, NLTK, MITIE and Grobid Quantities against textual descriptions of the sightings in your v2 dataset.
    c. Add the additional generated entities (People, Organizations, Locations, Measurements, etc.) to your dataset.
5. (**EXTRA CREDIT**) Improve the ImageCaptioning and Image Recognition Deep Learning model by re-training the last layer on images and UFO object types
    a. See: https://www.tensorflow.org/tutorials/image_retraining
    b. Generate two new models, one for UFO object types and another for UFO image captions based on those types.
    c. Submit a pull request to img2text (http://github.com/USCDataScience/img2txt.git) to add the new models.
    d. Comment on how well the new models perform.

4. Assignment Setup

4.1 Group Formation

You should keep the same group from your assignment one. There is no need to send any emails for this step.

5. Report

Write a short 4 page report describing your observations, i.e. what you noticed about the dataset as you completed the tasks. What questions did your new joined datasets allow you to answer about the UFO sightings previously unanswered? How well did the image captions accurately describe the UFO object types? What about the identified objects in the image? How well did OCR work? What did you have to do to clean up the noise in the data?

Thinking more broadly, do you have enough information to answer the following:

1. Of the incorporated British UFO sightings, how many of them could also similarly be explained akin to the sightings from the first assignment?
2. Were there any new object types introduced by the British UFO sightings?
3. How well were the British UFO sightings described? Was there a lot of missing data?

4. Of the UFO images, how many of the images actually generated image captions and/or objects that described the UFO and not just the background scenery?

Also include your thoughts about OCR pipelining, and Image Captioning/Object identification – what was easy about using it? What wasn't?

6. Submission Guidelines

This assignment is to be submitted ***electronically, by 12pm PT*** on the specified due date, via Gmail csci599spring2018@gmail.com. Use the subject line: CSCI 599: Mattmann: Spring 2018: ENRICHMENT Homework: Team XX. So if your team was team 15, you would submit an email to csci599spring2018@gmail.com with the subject "CSCI 599: Mattmann: Spring 2018: ENRICHMENT Homework: Team 15" (no quotes). **Please note only one submission per team**.

- All source code is expected to be commented, to compile, and to run. You should have at least a few Python scripts that you used to clean up the OCR extractions and turn them into sightings, along with scripts to scrape the images from UFO stalker, and likely to run through the image captioning and object identification.
- Include your updated dataset TSV. We will provide a Dropbox location for you to upload to.
- Also prepare a readme.txt containing any notes you'd like to submit.
- If you used external libraries other than Tika Python, you should include those jar files in your submission, and include in your readme.txt a detailed explanation of how to use these libraries when compiling and executing your program.
- Save your report as a PDF file (TEAM_XX_ENRICHMENT.pdf) and include it in your submission.
- Compress all of the above into a single zip archive and name it according to the following filename convention:

  TEAM_XX_CSCI599_HW_ENRICHMENT.zip

  Use only standard zip format. Do **not** use other formats such as zipx, rar, ace, etc.
- If your homework submission exceeds the Gmail's 25MB limit, upload the zip file to Google drive and share it with csci599spring2018@gmail.com.

Important Note:

- Make sure that you have attached the file the when submitting. Failure to do so will be treated as non-submission.
- Successful submission will be indicated in the assignment's submission history. We advise that you check to verify the timestamp, download and double check your zip file for good measure.
- Again, please note, only **one submission per team**. Designate someone to submit.
- Make sure you have your team members listed on your report
- Make sure your report is self-contained. Any plots, stats, etc. should be included in your report.
- Make sure you clearly answer and specify all the questions listed.

6.1 Late Assignment Policy

- -10% if submitted within the first 24 hours
- -15% for each additional 24 hours or part thereof