CSCI 599: Assignment 2 -

OCR, Data cleaning, Object recognition and Image Captioning

1. **What did you notice about the dataset as you completed the tasks ?**

- British_UFO data set though only 8 files, had a lot of pages for us to perform OCR on.
- OCR pipeline was helpful, but a little nudging around of the script to optimise the output was required.
- Though we didn't write any of the logic for OCR and pdf to image conversion, we were actually doing CV in this assignment.
- Using convert/magick with all the abundant options made us realise each page of every pdf unless strictly maintained to be of a particular format, OCR was difficult to automate. Every image required a different subset of options and different intensity of them to be used.
- Anything less, or more would give us garbled text after OCR.
- Surprisingly, we didn't have to set a lot of options for tesseract.
- Approximately 60-70% of the scraped images' output after OCR didn't relate to the images.
- Out of more than 4000 pages in the PDF files, only about 700 entries were added to the CSV file.
- We realised that preprocessing is not easy and at the worst case, each image/PDF needs a different set of options set for proper OCR.
- Tesseract worked better on images with black background and white text, so we tried converting all PDFs to that format before OCR.
- After OCR, the output had a lot of garbled values. A lot of cleaning went into the final output, yet, a lot of the output is not understandable.
- Few situations where the location of the sighting was drawn on a map and other situations where OCR is impossible, we resorted to using manual OCR for entries into V2.
- The dataset had different pictures of varying alpha(transparency coefficient), orientation, handwritten, typed, picturized, etc. images. So processing them was found to be hard.

2. What questions did your new joined datasets allow you to answer about the UFO sightings previously unanswered?

- The new dataset V2, has more entries(rows) than the existing V1 dataset.
- It includes a vaster geographical area, now that areas in the United Kingdom is considered for OCR. So questions like “Did any UFO sighting happen in such and such area of UK.
- Some entries have delineated stories and description about aliens (both graphic and textual) which brings in more entities for Named Entity Recognition and thus more questions to answer. The questions may range from, “Do the aliens look human?”, “Do aliens whose details are recorded, have bifocal vision?”, etc.
- Questions like “What organisations are listed in the dataset?”, “What is the most common name listed in the records of the UFO sightings?”
- What dialect people have interest in UFO?
- Were there any clouds in the sky during the following sighting?
- Objects/entities in the picture of UFO sighting.
- The questions can be extrapolated to the videos that were scraped as well.

Again, many other questions related to UFO, the geographical area, the cultural difference, language, etc. can be answered from V2 which we couldn’t do using V1.

Many CV related questions like what kinds of entities are most common in the images, are there more objects in the image than the number of UFOs? Etc.

3. How well did the image captions accurately describe the UFO object types?

Approximately, about 40% of all the image captions seemed to relate to the image when we considered a sample of the extracted images. It was even lesser, when we considered only the videos/output.

4. What about the identified objects in the image?

Objects identified from the image weren’t very precise, sometimes the context of the object present in the image would be wrong. Sometimes, the caption and the objects identification was right on the point. If I were to bet on the output of the object recognition, I would give it a 6:19 probability.

5. How well did OCR work? What did you have to do to clean up the noise in the data?

To clean up the noise, we had to change the several options while converting/preprocessing the data. Most images needed options subsets which wouldn't work with the rest of the dataset. So, the cleaning and the preprocessing took the majority of the work done. To clean up the noise, we had to look for options few of which would:

- change the contrast of the image
- Change the orientation of the image
- Flatten them
- Change transparency
- -despeckle, -compose darken to make page more dense to detect the low quality text
- Change the number of bits used to represent data tried 8 to 32 bits. 8 was the most suitable.

6. Of the incorporated British UFO sightings, how many of them could also similarly be explained akin to the sightings from the first assignment?

Very few sightings from the British UFO sightings "exactly" matched the entries in V1. The dates so not have a specific format in the OCR output, so exact matching to the V1 entries is not possible. Same logic goes with other features of the data.

7**: Were there any new object types introduced by the British UFO sightings?**

~75% of the object type/shape are those that we had in V1. The new entries are such that they describe the existing shapes but are worded differently.

8: How well were the British UFO sightings described? Was there a lot of missing data?

The British UFO sightings was fairly well documented and described. About 30% of the data was not in proper format and in the remaining 70% of it, there were 50% approximately, had tables in them and the other half had paragraph formatting.
There weren't many fields missing in the data which were well formatted. But, after OCR, due to attenuated print in the image, there was a lot of missing fields and text.
We managed to get only about 25% of all the formatted data into the final tabulation.

9. 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?

We scrapped around 1700 images and considering a sample of 100 images' output, we saw that there were no UFO related images. Then we checked for other entries randomly (without considering UFO related objects) we saw that there was high veracity in the captions generated according to the image fed for CV. The model provided wasn't trained on the UFO sightings pictures and the scraped images also didn't have any UFO related pictures. So, finally, the CV was done good but the UFO content in it was nil.

10. Our thoughts about OCR pipelining, and Image Captioning/Object identification – what was easy about using it? What wasn't?

The OCR pipeline: provided helped us start the assignment. It was easy to understand and edit. The inputs that were to be given to the script was individual pdf files. The options for the two commands was the main part we had to edit which was very clear from the given OCR pipeline script. No real difficult part in the pipeline to figure out. Just the right subset of options for the right kinds of images was needed for optimal output.

Image Captioning/Object identification: Using tika wasn't easy, we couldn't figure out it's errors. Documentation lacks details and you can understand the source code only if you have written it. Using the docker service directly was easier.