Alysa Xu, Catherine Lu, Justin Huh, Rohan Rane, Vishal Menon, Pratham Kambli
DSCI 550: Data Science at Scale
4th April 2025

DSCI 550 Group 8 Homework 2 Report

I. Observations from the Dataset

Q: Describe your observations (what you noticed about the dataset as you completed the tasks).
A:
Tika GeoTopicParser Tasks:

Using only the description and location for the parser resulted in only 18% extraction. By also combining the state of each haunted place for the parser, we were able to get 59% extraction. This means that most of the extracted geographic names ended up being state names. However, some of the extracted geographic names actually found specific places mentioned in the description, which was interesting. There were also rows where the parser misidentified a word as a geographic name, such as descriptions with the word "river" included resulting in the parser returning "River Nile" as the geographic name.

SPACY Tasks:

Even though there is structured metadata present with cities and witness counts, the descriptions have given a good sense of how people write and what is deemed necessary within the description. This made entity extraction both fascinating and challenging: while tools like SpaCy can pull out proper names and places, they often miss the emotional and ambiguous elements that give these stories life. We also encountered situations where cities like "Charlotte" or "Jackson" were misclassified as PERSON entities due to their dual use as names. In other cases, vague references such as "a shadowy figure" or "an unknown presence" held important meaning in context, but were not identified as entities at all. These gaps highlighted the limits of using out-of-the-box NLP tools on subjective, narrative-driven data. Despite these limitations, running SpaCy on the descriptions added a valuable layer of structured insight to our dataset. It allowed us to extract and compare frequently mentioned people, locations, and organizations across different haunted sites, giving us a new way to explore patterns that wouldn't be obvious from metadata alone.

Image Extraction Tasks:

The image extraction portion seemed to have some very interesting observations. We noticed that sometimes either captions were not found or NSFW images were tried to be generated but failed despite looking at the description to see why that might be. One thing we noticed about the dataset was that a lot of the images would continue to generate the same black silhouette that would show up in various ways across the images. We thought this was interesting, as when the description did not describe as much of what was seen, it gave us a quick generation of a figure or shadow this way. Apart from that, there were the objects in the image which were mentioned in the caption such as bridge, graveyard, church, etc. We also observed that extracting images using a GPU was significantly faster than using just CPU or remote machines. Furthermore, there were around only 20 rows out of the 10,000 where we couldn't identify the image paths for but this is most likely due to the fact that some datasets don't have images generated due to some NSFW images being generated (which were blocked from being added to the dataset automatically) or if the description was too long for the code to generate an image.

Tika Image Captioning Tasks:

Tika captioning gave interesting insights on the captions it extracted. Each image would have a few different captions generated with different levels of confidence, but we noticed that those captions wre very similar to each other and would only differ by one word or two. For example, the more detailed caption with higher confidence would include an adjective describing the main subject that the caption noted. We also noticed that the captions mostly described the main subject in the image and its relative position to other things in the image. The captions are all relatively simple and are not very long.

II. Questions of Interest

Q: Are there any haunted place correlations by location in the posts?
A: When looking at the locations of the haunted places in the posts, we saw a mix of what was considered a location. We had states show up (Washington, Ohio, Texas), we saw specific cities (e.g., *Chicago*), and we saw descriptors of locations (e.g., *Mansion, Main Street, Auditorium*). To our team, we were most intrigued by the specific location descriptors, such as Mansions and Auditoriums, giving us an understanding that there may be a correlation between specific types of locations and the presence of haunted locations.

Q: Are there correlations between the cities where the haunted places, and/or entities are identified with locations?
A: When looking at proportion, we noticed that only 531 rows, or 4.83% of the total rows that had the city listed in the dataset also appear as a named location in the corresponding description's extracted entities. Our team concluded that looking at the low proportion indicates many descriptions do not explicitly name the city, even if the metadata contains it, and that many haunted place descriptions have more of a reliance on indicating local references and assumptions, for example, referencing "the old mill" or referencing "Main Street". This may have been based on how data is collected and extracted. We can also therefore suggest that NER may not fully capture locality references in folklore-style text, and supplemental techniques might be necessary to use fuzzy matching or similar tools to help out with some references.

Q: Do the Entities provide any further context about the Haunted place? What about witness count? Does it allow you to further validate the witness count?
A: Over 25% of the haunted place descriptions showed some person entity however, those could either be eyewitnesses, victims, journalists, investigators, or just descriptive terms (nanny). We also saw that about 4% of the results showed more than 3 entities under person, which further analysis could give us a better understanding of the validity of claims since more people are involved. However, these entities could be investigators or police themselves – it is ultimately up to you to take this data with a grain of salt. At the end of the day, the person entities allowed us to extract and understand the context behind these haunted places. It would be interesting to understand whether the number of person entities plays a role in determining the validity of the data. The witness count we previously worked with may have inaccuracies based on the descriptions.

Q: Do the image captions accurately represent the image?
A: The image captions do not completely represent the images accurately. They are all very simplistic, which may also contribute to many details not getting captured. Some of the captions only vaguely identify the correct figures. For example, if there is a woman in the image, sometimes it is identified as a man. Other times, it identifies extra items in the image that do not exist but make sense in the context of what is represented in the image. The captions are mostly able to capture the type of object or figure present in the image, but not any of the specific details.

Q: Are the identified objects present in the image described in the original Haunted place and/or the generated caption?
A: When looking at this question we looked primarily to compare the original descriptions in addition to the generated image captions. We then used spaCY to ultimately help us extract entities and

evaluate if the entities in the generated captions matched the original descriptions. We saw that only 21.63% or 2,378 captions had at least one named entity that was showing up in the original description. In other words we saw that a majority of the captions do not include explicit mentions to entities from the original description, which we believe makes sense based on the random images we picked. We wanted to see from our view of the images whether they effectively matched and since many showcased a very simplistic version of the description, often leaving out complex multi-entity descriptions, we are not surprised to see this. Furthermore, we mentioned the shadow/black silhouette which overly simplifies the entities in the image. Going back to the entity analysis locations specifically versus location descriptors we noticed that location descriptors (such as old mill or church) were more likely to show up over Chicago or a specific location. Ultimately, the generated captions are based on visual object detection and due to the simplistic nature of the images, we are not surprised that there aren't many matches to the original descriptions (many of which had a lot of details and objects).

Q: Are there any specific trends you see in the text captions or identified objects in the image media?
A: Some common themes in the text captions include reccuring hauntings and ghost sightings. Frequent terms involved include the following: "figure", "apparition", "ghost", "haunting", "mysterious", "spotted", "witnesses". Captions often mention terms like "male", "female", "child" apparitions, or "unknown figure". This implies vagueness or shadowy entities. Terms like "described", "report", "believed" show it's based on anecdotal accounts and Time-Based Patterns: involved later parts of the day such as "evening", "morning", or "during the night". This suggests most sightings are claimed to occur in low-light or eerie times of day. Some of the associated events include common words such as "murder", "death", "disappearance", "school", "cemetery". Many captions are tied to tragic past events, especially involving children or schools. Cemeteries are especially prevalent and common haunted locations include roads, houses, universities, schools.

III. Thoughts About ML and Deep Learning Software

Q: Include your thoughts about the ML and Deep Learning software like GeoTopicParser, SpaCY, Tika Image Captioning, etc.–what was easy about using it? What wasn't?
A:

Setting up GeoTopicParser was the most difficult part. It required installing and learning to use Linux, then following modified README instructions from another student in the class to set up the path correctly. However, once all the lucene-geo-gazetteer library was set up properly, it was very easy to start the Tika server and make queries to the GeoTopicParser. It was also straightforward to write the Python code for running the haunted places TSV file and save the extracted geographic names, longitudes, and latitudes as new columns.

SpaCY was extremely easy to run; the biggest challenge was ultimately tackling it with our various members. Justin and Rohan both wrote code for it, and although their code was fairly similar the amount of empty entity responses varied. The one addition we felt was necessary was incorporating a progress bar, although that was due to the large dataset, and ensuring the code was running. We also noticed some discrepancies in city names potentially being persons and looked into troubleshooting by comparing outputs, adjusting how we passed the text into SpaCy, and analyzing entity labels more careully. As we reviewed the results, we began to see that SpaCy's performance was generally strong in pulling out proper nouns like people, cities, and landmarks, but seemed to struggle with ambiguity and context. Despite the occasional quirks, the tool was extremely helpful by giving us great insights and being extremely easy to use.

Tika Image Captioning was straightforward to use. The instructions on how to install it were easy to follow, and after that what we did was convert our zip file of generated images from PNG to JPEG. From there, we iterated through the entire new zip file through its pathway with all JPEGs to generate the captions for all images. The captioning was relatively quick to do. However, we noticed that the captions were all pretty simplistic and some were not very accurate to what the image was showing. It was interesting to compare the captions generated with the original descriptions we used to generate the images. Unfortunately, we were unsuccessful in running the object detection with Tika. After installing the dockers by following the instructions, they were built successfully. Although they were all built, when running the service, we could only access the API used for image captioning through this endpoint: http://localhost:8764/inception/v3/caption/image, and we were able to use it to iterate through our images to caption them. However, when trying to access the API for object detection, we encountered multiple errors and initially could not locate it. We tried cloning the repository for the dockers, which was able to be done. When we tried to run the API, there were problems with missing model files including being unable to extract the inception model. There were also some compatibility issues with TensorFlow where older APIs were deprecated, and missing dependencies which we added commands to install but there were still some issues with different versions. The Inception V4 Model was also needed to do the object detection, but we were unable to extract it and make it accessible. We tried fixing these errors, but ultimately could not find the correct models needed for the object detection. However, given that we were able to use SpaCY to identify entities and run the captioning on all images, we were still able to extract valuable insights from our dataset.

For image generation, it was simple utilizing Hugging Face's diffusers Library. Using the StableDiffusionPipeline from diffusers made it easy to generate high-quality images from text prompts with just a few lines of code. No model training was required. It simplifies the process of loading model weights, preprocessing text, and handling the diffusion sampling process. PyTorch also handles tensor and GPU memory management well, which is important when working with large models like Stable Diffusion. Some of the more difficult parts included getting the compatible versions of torch, transformers, and diffusers libraries, especially in Colab. The overall process is not optimized for speed, especially when running on CPUs or non-NVIDIA hardware.