DSCI 550 HW2 Report: Haunted Places Dataset - Entity Extraction and Geolocation Analysis

Team 03 Members: Colin Leahey, Zili Yang, Chen Yi Weng, Aadarsh Sudhir Ghiya, Niromikha Jayakumar, Yung Yee Chia

1. Introduction

In this project, we expanded upon the haunted places dataset previously explored in Homework 1 by incorporating multiple layers of contextual enrichment using natural language processing and image processing tools. The main objective was to extract additional metadata from text descriptions using SpaCy and Apache Tika, perform geolocation using GeoTopicParser, generate visual imagery using Stable Diffusion, and integrate image captioning and object detection through Tika Docker.

By transforming each record into a more semantically rich and multimedia-aware format, we aimed to explore deeper patterns related to geography, story content, and visual representations.

2. Methodology Overview

2.1 Entity Recognition (NER) with SpaCy

We used SpaCy's en_core_web_sm model to extract named entities from the description column. The result was stored as a new field called Named_Entities. The entities extracted included PERSON, DATE, LOCATION, ORG, TIME, CARDINAL, among others.

One interesting challenge we noted was that while SpaCy captured relevant information, such as people's names and specific dates, its utility in confirming witness counts or verifying spatial mentions was limited due to lack of domain tuning. Still, the contextual layer added valuable narrative cues that aided understanding.

2.2 Geolocation with GeoTopicParser

We used the GeoTopicParser REST service to retrieve geographic coordinates. However, we encountered a key challenge: GeoTopicParser often returns **multiple or ambiguous locations** when presented with compound input. For example, when a description included both a specific town and a vague regional name, the parser's output frequently prioritized the wrong match.

To overcome this, we decided to **not append additional ambiguous geo-strings** if a clear city, state, and location were already present. Out of over 10,000 entries, only ~40 had coordinates that differed from the original dataset, indicating:

- Either the original lat/lon values came from the same gazetteer (likely GeoNames)
- Or GeoTopicParser returned no valid result, defaulting to the original values

This highlighted a broader lesson: **NER models like GeoTopicParser need domain-specific tuning**. While effective for general news, they struggle with historical or narrative datasets like ghost sightings, where geographic references vary in style and granularity.

2.3 Image Generation with Stable Diffusion

Using a reduced version of the Stable Diffusion v1.5 model, we generated one image per haunted record. The prompt was constructed from the description, city, state, and apparition_type columns. Due to limited compute resources, we adjusted the model to use smaller resolution (384x384) and fewer inference steps (30) for efficiency.

Generated images were saved with unique IDs and referenced in the dataset via an ai_image_path column.

2.4 Image Captioning and Object Detection with Tika

We ran a Tika Docker container to extract:

- Image captions (descriptive text of the image)
- Detected objects (visual entities in the image)

Though the Docker container initialized properly, we encountered frequent connection errors during batch processing (e.g., No connection adapters were found for "b'://None:80/meta/text'"), indicating an unstable internal API endpoint or timeout configuration. Despite this, a subset of images was successfully processed, and we included those results in caption and detected_objects fields.

3. Analysis and Observations

3.1 Is there a geographical correlation between haunted places and story content?

Yes. Cities with high population or historical significance tended to exhibit more detailed or narrative-rich descriptions (e.g., Chicago, New Orleans). Haunted locations near historical landmarks were especially common.

3.2 Are certain cities associated with specific types of apparitions or events?

From the apparition_type column, we observed that:

- **Schools and dorms** had a higher rate of ghost children or poltergeists
- **Rural towns** more frequently referenced witches or cursed grounds This spatial-narrative relationship can be leveraged for further classification or story type clustering.

3.3 Did SpaCy's Named Entities provide additional context?

While not always directly useful for spatial analysis, the extracted entities offered **richer narrative elements**—identifying people, dates, or institutions within the story. These can support future entity linking or historical verification efforts.

3.4 Do the image captions align with the stories?

In cases where captioning succeeded, the image descriptions were often vague or generic (e.g., “A haunted house at night”). There was **limited alignment** with the text-based descriptions unless the prompt included strong visual cues like “graveyard” or “classroom.”

3.5 Were the objects detected in images reflective of the story?

Similar to captioning, object detection returned common objects (e.g., "building", "tree", "room") that lacked thematic alignment. The abstract and eerie nature of haunted-themed prompts likely challenged the object detector's general-purpose training.

3.6 Did any trends emerge in captions or object detections?

Notable trends included:

- **Frequent mentions of buildings, rooms, and lights** in captions
- Sparse object variety, dominated by generic indoor/outdoor categories This suggests that while useful for basic classification, these tools require more task-specific fine-tuning to deeply understand horror narratives.

4. Tool Reflection and Limitations

Tool	**Strengths**	**Limitations**
SpaCy (NER)	Easy to use, flexible, identifies diverse entities	Needs domain tuning, not always relevant to story
GeoTopicParser	Good for geo-coordinates if input is clean	Sensitive to ambiguity, requires structured input
Stable Diffusion	Generates visually compelling haunted scenes	Limited control over output consistency
Tika (Docker)	Integrates OCR, captioning, and detection	Unstable API, connection issues, generic results

5. Conclusion

This project demonstrated the power and limitations of integrating NLP, geospatial, and vision-based tools into a unified dataset enrichment pipeline. While many of the generated insights were constrained by the limitations of generic models, the process revealed valuable strategies for improving future pipelines—such as better entity disambiguation, custom-tuned image prompts, and stable Docker-API interfacing.

We believe this enhanced dataset has the potential to serve as a foundation for more advanced modeling in the future—such as classifying ghost stories by type, visualizing haunted hotspots, or even generating interactive narrative maps.