Homework: Large Scale Data Extraction & Analysis for Haunted Places
Due: Friday, April 4, 2025 12pm PT

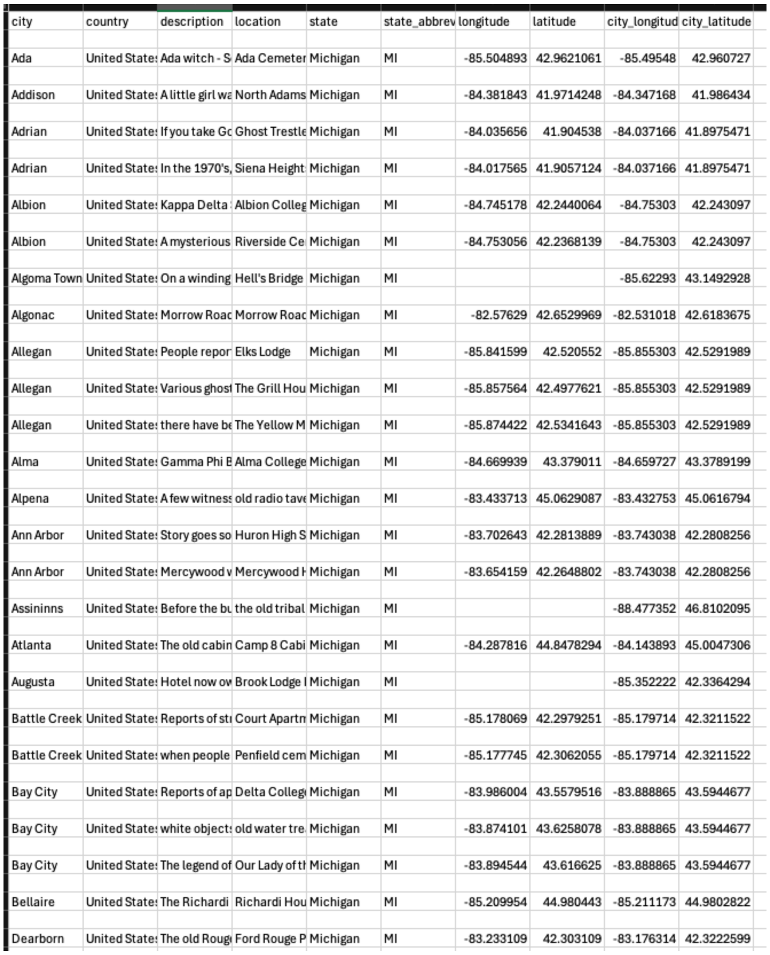
1. Overview

Figure 1. Images generated using Imagine from Meta, and ChatGPT/DALL-E based on descriptions from four of the haunted places records.

In the second assignment, you will build upon the work that you did in assignment 1 in identifying additional explanations and features that helped you understand the Haunted places data, and more information about Haunted places that could help you validate and verify them and their associated potential validity. You added features that correlated alcohol consumption, amount of hours in a day, you derived features that allowed you to identify visual, audio, and other evidence, and you investigated physical properties of the Haunted place. You cleaned and identified the appropriate count of witnesses. Further, beyond these new columns, you were asked to select 3 datasets of 3 different top level MIME types and for each dataset, add 3 new features (total 9 new features) to your data. At the end of the new you added a number of new features to the data exploring the unintended consequences of Big Data, and the five 5 V's.

In this assignment you will focus on large scale content extraction and data science by exploring parts of the Haunted Places data left unexplored in assignment 1. You will explore the unique properties of locations in the dataset by using a GeoParser tool developed in the USC Information Retrieval and Data Science (IRDS) group.

There are plenty of **locations** mentioned in the location field, the state field, and the description field. As we have discussed in class during the advanced extraction lectures, as well as the metadata lectures, and clustering lectures, and as we will talk about during the Named Entity Recognition (NER) lecture, it is possible to use machine learning and natural language processing (NLP) to scan text and extract locations. For example, as you can see from the City and State fields in the data, you could input the text based information and receive a coded Lat/Lng geolocation back.



city	country	description	location	state	state_abbrev	longitude	latitude	city_longitud	city_latitude
Ada	United State	Ada witch - S	Ada Cemeter	Michigan	MI	-85.504893	42.9621061	-85.49548	42.960727
Addison	United State	A little girl wa	North Adams	Michigan	MI	-84.381843	41.9714248	-84.347168	41.986434
Adrian	United State	If you take Gc	Ghost Trestle	Michigan	MI	-84.035656	41.904538	-84.037166	41.8975471
Adrian	United State	In the 1970's,	Siena Height	Michigan	MI	-84.017565	41.9057124	-84.037166	41.8975471
Albion	United State	Kappa Delta	Albion Colleg	Michigan	MI	-84.745178	42.2440064	-84.75303	42.243097
Albion	United State	A mysterious	Riverside Ce	Michigan	MI	-84.753056	42.2368139	-84.75303	42.243097
Algoma Town	United State	On a winding	Hell's Bridge	Michigan	MI			-85.62293	43.1492928
Algonac	United State	Morrow Roac	Morrow Roac	Michigan	MI	-82.57629	42.6529969	-82.531018	42.6183675
Allegan	United State	People repor	Elks Lodge	Michigan	MI	-85.841599	42.520552	-85.855303	42.5291989
Allegan	United State	Various ghost	The Grill Hou	Michigan	MI	-85.857564	42.4977621	-85.855303	42.5291989
Allegan	United State	there have be	The Yellow M	Michigan	MI	-85.874422	42.5341643	-85.855303	42.5291989
Alma	United State	Gamma Phi B	Alma College	Michigan	MI	-84.669939	43.379011	-84.659727	43.3789199
Alpena	United State	A few witness	old radio tave	Michigan	MI	-83.433713	45.0629087	-83.432753	45.0616794
Ann Arbor	United State	Story goes so	Huron High S	Michigan	MI	-83.702643	42.2813889	-83.743038	42.2808256
Ann Arbor	United State	Mercywood v	Mercywood H	Michigan	MI	-83.654159	42.2648802	-83.743038	42.2808256
Assininns	United State	Before the bu	the old tribal	Michigan	MI			-88.477352	46.8102095
Atlanta	United State	The old cabin	Camp 8 Cabi	Michigan	MI	-84.287816	44.8478294	-84.143893	45.0047306
Augusta	United State	Hotel now ov	Brook Lodge I	Michigan	MI			-85.352222	42.3364294
Battle Creek	United State	Reports of st	Court Apartn	Michigan	MI	-85.178069	42.2979251	-85.179714	42.3211522
Battle Creek	United State	when people	Penfield cem	Michigan	MI	-85.177745	42.3062055	-85.179714	42.3211522
Bay City	United State	Reports of ap	Delta Colleg	Michigan	MI	-83.986004	43.5579516	-83.888865	43.5944677
Bay City	United State	white object	old water tre	Michigan	MI	-83.874101	43.6258078	-83.888865	43.5944677
Bay City	United State	The legend of	Our Lady of th	Michigan	MI	-83.894544	43.616625	-83.888865	43.5944677
Bellaire	United State	The Richardi	Richardi Hou	Michigan	MI	-85.209954	44.980443	-85.211173	44.9802822
Dearborn	United State	The old Roug	Ford Rouge P	Michigan	MI	-83.233109	42.303109	-83.176314	42.3222599

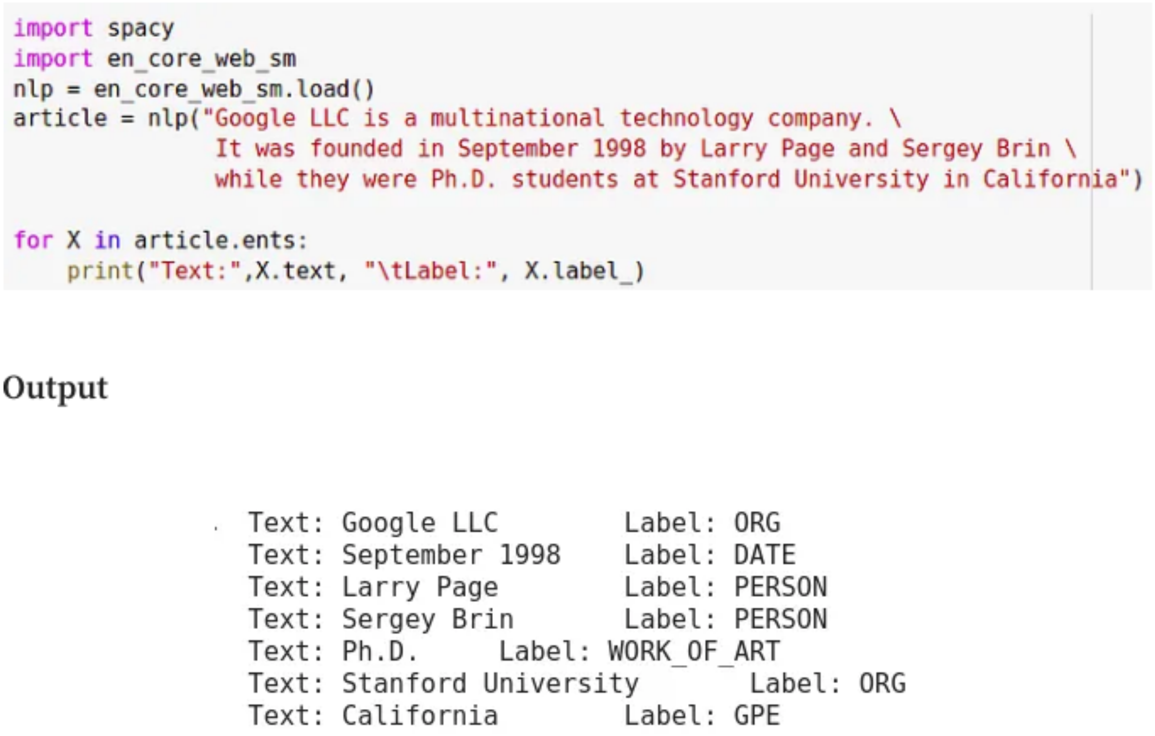
Figure 2. City, and State fields that you could input into the GeoParser tool.

Location names like "Sacred Heart Academy", "Alma College", and so on, can be geocoded, and their corresponding latitude and longitude can be extracted using a tool developed by the USC

Data Science Group called GeoTopicParser. The tool can take some text, and then perform an analysis using the Geonames.org database, and custom NLP and NER parsing code, using Tika, generating output that looks like the following if the text contained in the post mentioned, e.g., “China”:

```
[
   {
      "Content-Type":"application/geotopic",
      "Geographic_LATITUDE":"39.76",
      "Geographic_LONGITUDE":"-98.5",
      "Geographic_NAME":"United States",
      "Optional_LATITUDE1":"27.33931",
      "Optional_LONGITUDE1":"-108.60288",
      "Optional_NAME1":"China",
      "X-Parsed-By":[
         "org.apache.tika.parser.DefaultParser",
         "org.apache.tika.parser.geo.topic.GeoParser"
      ],
      "X-TIKA:parse_time_millis":"1634",
      "resourceName":"polar.geot"
   }
]
```

One last thing that we can do with the text, illustrating large scale content analysis is to run machine learning approaches on Haunted Places sightings to determine other named entities in the sightings text. We can do this in assignment 2 by leveraging machine learning, and the SpaCY tool is a named entity recognition library that will examine text and identify all of the associated entities present in content in the text.



```
import spacy
import en_core_web_sm
nlp = en_core_web_sm.load()
article = nlp("Google LLC is a multinational technology company. \
               It was founded in September 1998 by Larry Page and Sergey Brin \
               while they were Ph.D. students at Stanford University in California")

for X in article.ents:
    print("Text:",X.text, "\tLabel:", X.label_)
```

Output

```
Text: Google LLC        Label: ORG
Text: September 1998    Label: DATE
Text: Larry Page        Label: PERSON
Text: Sergey Brin       Label: PERSON
Text: Ph.D.     Label: WORK_OF_ART
Text: Stanford University       Label: ORG
Text: California        Label: GPE
```

Figure 3. The SPaCY tool and named entity recognition. Notice in the top portion an article is input into the tool in Python and the bottom portion is the output from SpaCY.

However, we're not stopping at the text, and in this assignment, you will generate an image for each associated Haunted Place. You will accomplish this by leveraging a Generative AI tool called Stable Diffusion and providing it with the description text and experimenting to generate an image for each Haunted Place. Once you have the image associated with the haunted place, you can then use a machine learning based Image Captioning algorithm called Show & Tell originating from Google to automatically caption and generate text features about the Haunted place use these captions as additional features to describe it. 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 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 4a-c showing 3 automatically generated labels (with only generic training). We will integrate this Tika capability and generate labels and text captions for your haunted places images.

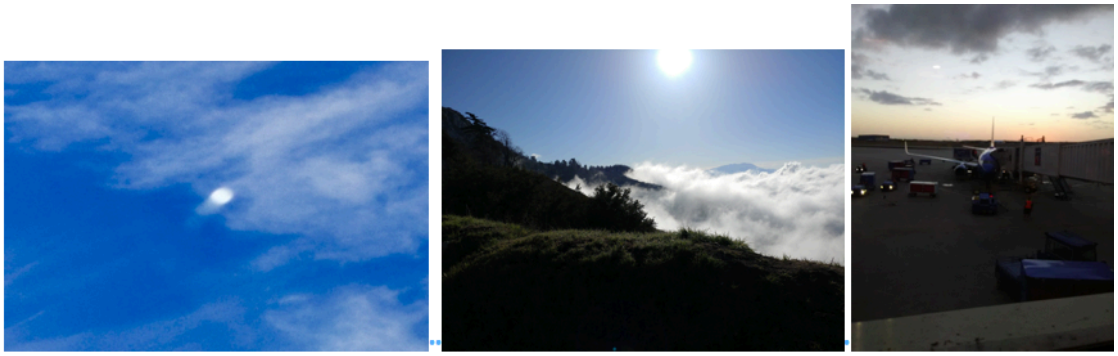


Figure 4: a) a light/orb shown in the daylight; b) an orb present against a mountain background; and c) an orb in a cloudy sky.

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

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

The combination of these techniques will allow you to apply knowledge gained from the Parsing/Extraction Lecture, the lectures on advanced extraction (including Deep Learning and Metadata), 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 this assignment is two-fold. First, you will expose the richness of the Haunted places and generate accurate, geolocated Lat/Lng that can be used to map and understand your data. You can identify location names using the GeoTopicParser and geocode the associated locations in the sightings and then finally you can run SpaCY to automatically classify and identify additional features in the Haunted places data. The location names will also aid you in comparing e.g., Haunted places, associated features, and information together from your prior assignment 1 work. In addition, using SpaCY will provide you with entities that you can use to further classify and study your Haunted places data.

In addition, the other objective of this portion of the assignment is to leverage the richness in the underlying images you generate for each Haunted place, along with large scale machine learning and data science, to generate text from the images (a caption), along with the list of objects present in the imagery. Both will provide additional features with which you can examine the Haunted place, and compare: are these images representative of the sighting and something that provides a visual aid to better understand the report? You will explore these questions in assignment 2.

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 columns for
   a. GeoTopic name, along with associated lat/lng
   b. Columns for all generated SPaCY named entities

    c. A column pointing on disk to your generated AI image for the report.
    d. Image Caption generated by Tika's Show & Tell caption generator
    e. Detected objects in the image using Tika's Inception Rest service
2. Download and install Tika Python using PIP and the instructions at http://github.com/chrismattmann/tika-python
3. Install GeoTopicParser using the instructions here https://cwiki.apache.org/confluence/display/tika/GeoTopicParser
    a. The result of this should be the Lucene GeoGazetteer REST server running as specified here: https://github.com/chrismattmann/lucene-geo-gazetteer
    b. You can connect the GeoGazetteer to Tika-Python using the instructions here: https://github.com/chrismattmann/tika-python#changing-the-tika-classpath
4. Install SpaCY using PIP and the instructions here: https://spacy.io/usage
5. Use a Generative AI image generator service such as any of the following: Stable Diffusion, Imagine from Meta, Midjourney, DALL-E 3 from OpenAI
    a. Leverage the text from your sighting columns, such as Description, etc.
    b. Combine that text together into a caption
    c. Input the caption into the Image Generation service to receive the generated image
    d. Note that you can use the API calls to these services to script/automate your process
6. Install Tika Image Dockers and generate captions for your Haunted places images from step 5
    a. To access the images, use a local file URL from your generated Haunted places images
    b. Install Tika Dockers package for Image Captioning and Object Recognition
        i. git clone https://github.com/USCDataScience/tika-dockers.git and https://hub.docker.com/r/uscdatascience/im2txt-rest-tika
        ii. Read and test out: https://cwiki.apache.org/confluence/display/TIKA/TikaAndVisionDL4J
        iii. Read and test out: https://github.com/apache/tika/pull/189
    c. Iterate through all the Haunted places images and add the generated image caption and the detect object(s) column to your dataset
7. Iterate through all the Haunted places and then run Tika GeoTopicParser and extract out Location name, including Lat/Lng based on your text fields for the sighting
    a. Write a Python program to do this
    b. Add the new column(s) to your dataset
8. Run SpaCY on all the Haunted places and extract the associated entities present in the description text
    a. Add these columns and scores to each of your Haunted places in your new dataset

4. Assignment Setup

4.1 Group Formation

You can work on this assignment in groups sized at **minimum 2, and maximum 6**. You may reuse your existing groups from discussion in class. If you have any questions, contact the TA via his email address with the subject:
DSCI 550: Team Details.

5. Report

Write a short 4-page report describing your observations, i.e. what you noticed about the dataset as you completed the tasks. For example, the following questions are of interest.

1. Are there any Haunted places correlations by location in the posts?
2. Are there correlations between the cities where the haunted places, and/or entities identified with locations?
3. 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?
4. Do the image captions accurately represent the image?
5. Are the identified objects present in the image described in the original Haunted place and/or the generated caption?
6. Are there any specific trends you see in the text captions or identified objects in the image media?

Also 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?

6. Submission Guidelines

This assignment is to be submitted ***electronically, by 12pm PT*** on the specified due date, via Gmail dsci550.2025a@gmail.com for the Thursday class, or dsci550.2025b@gmail.com for the Tuesday class. Use the subject line: DSCI 550: Mattmann: Spring 2025: EXTRACT Homework: Team XX. So, if your team was team 15, and you had the Thursday class, you would submit an email to dsci550.2025a@gmail.com with the subject "DSCI 550: Mattmann: Spring 2025: EXTRACT 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 join three other datasets, and what you used to extract additional features.
- Use relative paths {not absolute paths} when loading your data files so that we can execute your script/notebook files without changing everything.
- If using a notebook environment, use markdown cells to indicate which tasks/questions you are solving.
- Include your updated dataset TSV. We will provide a Dropbox or Google Drive location for you to upload to {you don't need to attach it inside the zip file}.
- 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_EXTRACT.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_DSCI550_HW_EXTRACT.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 dsci550.2025a@gmail.com (Thursday class) or dsci550.2025b@gmail.com (Tuesday class).

When submitting, please organize your code and data file as the directory structure shown:

```
Data
      dataset1 {leave it empty for now}
Source Code
      script1
      notebook
Readme.txt
Requirements.txt
```

Important Note:

- Make sure that you have attached the file 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.

6.1 Late Assignment Policy

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