Prioritized ideas:

Library Photo Morgue

Description: The NU Library has received thousands of images from the Boston Globe. Their goal is to digitize and share this resource with the larger public. They're interested in a few capabilities that might be delivered by Data Science methodologies:

Automated metadata generation/enhancement: Many of these images already have metadata (e.g. year, subject) attached to them, but others do not. Even for those with metadata, sometimes the data is limited and would make searching for particular images difficult. By using the images and their metadata, there may be ways of generating metadata for new images or enhancing the metadata (e.g. with pre-trained image tagging models).

Some examples:

- Using a pre-trained image model to add tags to the image metadata
- Using caption generation models to generate captions and combining those with available metadata

Enhanced navigation and searchability: The ultimate goal of the project is to enable public use of the images. Currently, search is limited to string matching a user's query to the image metadata. There are likely to be better ways of navigation of these images.

- Improved search based on query-metadata similarity metrics
- Grouping and visualization of images based on similarity to one another and content they contain (see an example here: https://dhlab.yale.edu/projects/pixplot/)

Currently we need to confirm the dataset will be available and under what restrictions. One proxy dataset might be the Microsoft COCO image captioning dataset, which has images paired with captions (treated as metadata): https://cocodataset.org/#captions-2015

Synthetic data generation

Description: The AI Solutions Hub (AISH) often has to scope projects without having seen a client's datasets. This access is sometimes delayed even past the start of a project. AISH would like to be able to identify potential issues and begin work on a solution before access is obtained. In order to do that, we are aiming to develop a flexible data generation framework that would, given a set of parameters, generate a dataset similar to the expected dataset from the client:

Automated metadata generation/enhancement: Many of these images already have metadata (e.g. year, subject) attached to them, but others do not. Even for those with metadata, sometimes the data is limited and would make searching for particular images difficult. By using

the images and their metadata, there may be ways of generating metadata for new images or enhancing the metadata (e.g. with pre-trained image tagging models).

Some examples:

- Using a pre-trained image model to add tags to the image metadata
- Using caption generation models to generate captions and combining those with available metadata

Enhanced navigation and searchability: The ultimate goal of the project is to enable public use of the images. Currently, search is limited to string matching a user's query to the image metadata. There are likely to be better ways of navigation of these images.

- Improved search based on query-metadata similarity metrics
- Grouping and visualization of images based on similarity to one another and content they contain (see an example here: https://dhlab.yale.edu/projects/pixplot/)

Currently we need to confirm the dataset will be available and under what restrictions. One proxy dataset might be the Microsoft COCO image captioning dataset, which has images paired with captions (treated as metadata): https://cocodataset.org/#captions-2015

Older ideas:

Client: Mark Waggy IA Solutions Hub

Project Description:

- Notebook to report workflow: Build a Github Action workflow that generates a report from a series of Jupyter notebooks when annotated properly. Likely will make use of Scrapbook and Papermill; but open to other tooling.
- EAI Data Lake. Set up a Data Lake that is searchable and self-documenting. It must allow for unstructured, semistructured and structured data and be minimal cost that we can use to track public (and internal) data for use with projects.
- ML Deployment Pipeline. A pipeline for monitoring and deployment of machine learning models.
- Knowledge Graph for AI Solutions. Build a knowledge graph that can track, visualize and export entities and their relationships with expandable features. This one probably needs a lot more explanation but it is a cool project. Believe me 🙂

- ML Bias Reference. Build a tool that generates synthetic dataset that we can use as an "ML Bias" point of reference. Operationally it would take in an ML algorithm and provides an indication of the level of bias in an ML algorithm with respect to a synthetic a reference dataset.

EAI -- Solutions Hub

Project definition -- spring 2023

* Stakeholder -- Mark Wagy, PhD, Senior Data Scientist at The Roux Instute

* Story/Goal: We build machine learning models on a lot of datasets for various partners and industries. We would like to be able to have an automated process that we can run our algorithms against ground truth datasets to check whether they are biased or not. In this project we propose building such a tool.

* Data: [Law School Admission Dataset](http://www.seaphe.org/databases.php)

* **Other EIA project ideas that need to be fleshed out...**

  * Notebook to report workflow: Build a Github Action workflow that generates a report from a series of Jupyter notebooks when annotated properly. Likely will make use of Scrapbook and Papermill; but open to other tooling.

  * EAI Data Lake. Set up a Data Lake that is searchable and self-documenting. It must allow for unstructured, semistructured and structured data and be minimal cost that we can use to track public (and internal) data for use with projects.

  * ML Deployment Pipeline. A pipeline for monitoring and deployment of machine learning models.

  * Knowledge Graph for AI Solutions. Build a knowledge graph that can track, visualize and export entities and their relationships with expandable features. This one probably needs a lot more explanation but it is a cool project.

  * ML Bias Reference. Build a tool that generates synthetic dataset that we can use as an "ML Bias" point of reference. Operationally it would take in an ML algorithm and provides an indication of the level of bias in an ML algorithm with respect to a synthetic a reference dataset.