DA323: Multimodal Data Analysis and Learning 2.0

Assignment-cum-Project-01
Jan-May 2025, IIT Guwahati

***About*—This assignment-cum-project focuses on scalable data collection, multimodal analysis, and computational matching techniques across various data modalities. You will implement automated methods to collect, process, and analyze image, text, audio, and structured weather data, gaining hands-on experience with data scraping, preprocessing, and documentation. Additionally, a multimodal challenge requires you to match audio tracks with corresponding silent video clips based on extracted features, simulating real-world multimodal data problems. Another task explores national flags and anthems, analyzing color symbolism and linguistic themes to uncover potential cultural correlations. The final submission requires a structured GitHub repository, demonstrating well-documented code, insightful visualizations, and a clear presentation of findings.**

Task: Scalable Data Collection

In this task, you will explore and implement scalable methods to collect, curate, and document datasets across different modalities—image, text, audio, and structured weather data. The goal is to gain hands-on experience in automated data collection techniques, preprocessing, and dataset documentation, which are fundamental for building data-driven technologies.

A. Image Dataset Collection

(a) List 20 different categories of your choice. Search for images for each of these categories using Google Images or another image website.
(b) Implement an automated script to download 50 images for each topic.
(c) Store metadata, including image URL, filename, and resolution in a CSV file.
(d) Curate and organize the dataset into labeled folders based on the predefined categories.

Challenge: Give a name to the dataset. Demonstrate a use case for this data set.
Suggested Python tools: selenium, requests, BeautifulSoup, google-images-search

B. Text Dataset Collection

(a) List 20 different categories of your choice. Select three websites for each of the categories.
(b) Implement a web crawler to extract relevant text (e.g., article titles, content, publication date) from the above websites for each of the 20 categories.
(c) Store the collected text data for each category into a text file (that, is make 20 text files for 20 categories).
(d) Clean the text using NLP preprocessing techniques (e.g., removing HTML tags, punctuation, stop words).

Challenge: Give a name to the dataset. Demonstrate a use case for this data set.
Suggested Python tools: requests, beautifulSoup, scrapy

C. Audio Dataset Collection

(a) Identify online AM/FM radio stations that stream publicly available audio.
(b) Implement an automated script to record and store audio streams for a given duration ($30 - 90$ s).
(c) Save 30 audio files in WAV/MP3 format and store metadata such as station name, timestamp, and duration.

Challenge: Give a name to the dataset. Demonstrate a use case for this data set.
Suggested Python tools: ffmpeg, pydub, requests

D. Weather Dataset Collection

(a) Use an open-source weather API (e.g. OpenWeatherMap API) to collect historical and real-time weather data.
(b) Implement periodic data collection to observe weather trends over a month.
(c) Query data for 20 Indian cities and store parameters like temperature, humidity, and wind speed in a CSV file.

Challenge: Give a name to the dataset. Demonstrate a use case for this data set.
Suggested Python tools: requests, matplotlib, pandas

E. Analyzing India with Data

The **https://data.gov.in** data portal is meant to facilitate open access to shareable data owned by the Government of India (along with its usage information) in machine-readable form through the Internet across the country in a periodically updated manner. Data sets are published in an open format and are available for usage by the public in formats such as CSV, XLS, JSON, XML, etc.

(a) Select a dataset from https://data.gov.in that interests you.
(b) Clean and pre-process the data if needed.
(c) Perform exploratory data analysis (EDA) and identify trends, correlations, and build insights.

Suggested Python tools: matplotlib, seaborne, plotly

Task: Search for a match

We generated simulations of a ball following projectile motion inside a 2D rectangular frame. These simulations produced videos depicting the ball's movement within the frame, accompanied by an audio track correlated with the motion. An example can be seen at: https://www.youtube.com/watch?v=9acydOSvVj4.

After generating 45 such videos, we provide you with a dataset containing *audio and (muted) video tracks*. Your task is to *match*

Course Instructor: Neeraj Kumar Sharma (e-mail: neerajs@iitg.ac.in), Mehta Family School of Data Science & AI.

each audio track with its corresponding muted video. The dataset (.zip file) includes the following folders:

- *audio_only*: Contains 45 audio files (.wav) named as `audio_only_<ID>_.wav`.
- *video_only*: Contains 45 video files (.mp4) named as `video_only_<ID>_.mp4`.

Your task is to match the audio files in the *audio_only* folder with the corresponding video files in the *video_only* folder. The final matches should be submitted as a *CSV file* with two columns. The required format is provided inside the dataset folder. The dataset is available at: click here.

To accomplish this, design and implement a computational approach that *analyzes visual features in the videos and matches them with acoustic features extracted from the audio files*. This challenge reflects real-world problems in *multimodal data analysis*, offering an engaging and intellectually stimulating experience.

Task: Analyzing flags and anthems

Analyze the visual characteristics of national flags and the linguistic features of national anthems to identify correlations between color symbolism and national identity, history, or cultural themes.

A. Data Collection

(a) Obtain images of national flags of atleast 100 nations. *Hint:* Click here for one source.
(b) Obtain English translations of the national anthems of the above nations. *Hint:* Check https://nationalanthems.info/.
(c) Obtain music compositions (mp3 files) of the national anthems of a subset of the above nations. *Hint:* Check https://nationalanthems.info/.

B. Visual Analysis

(a) Read the blog post available at: click here.
(b) Now, do your own analysis of the flag images using data analysis techniques. In addition to validating some of the observations made in the above blog post, come up with newer observations based on your analysis.

C. Textual Analysis

(a) Pre-process the anthem translations for stop word removal.
(b) Pursue your own analysis of the obtained text, similar in spirit to what you explored for the images.

D. Audio Analysis

(a) Pursue your own analysis of the obtained music audio files, similar in spirit to what you explored for the images.

E. Multimodal correlation

(a) Think and explore if their exist some multimodal correlations between the above three modalities, in the spirit of the above context.

Submission Protocol

Prepare a Github repository. For each of the tasks, create a folder inside the repository. This repository should contain the codes, plots, jupyter notebook(s), created datasets, and anything else of relevance.

Keep the notebook neat by following best practices for code writing, plotting, and using bulleted text to state observations. You are welcome to be creative in solving the tasks as long as you demonstrate clarity in presentation.

Assignment due by 11:55 PM, 09-Mar-2025.