Project Milestone 1 (the promytheus phase): Project Proposals, Team formation

Key dates:

- project proposals due: 09/25
- staff feedback: 09/29

Objectives:

For the first milestone, your team will **propose a project** that aligns with your personal, professional, and academic interests and passions.

Allowing you to propose your own projects, will enhance your engagement and lead to better learning outcomes. This approach will also foster your independence, critical thinking skills, and creativity, preparing you for real-world scenarios where you may be required to initiate and lead your own projects. Call on your inner data scientist and take charge of your project experience.

STEP 1: CREATE TEAMS (GROUPS OF 3-4 STUDENTS)

Platform for Team Formation:

You may use the Ed platform to find your teammates. Alternatively, you may form teams independently.

Team Registration:

Once you have finalized your team, please enter your team name and the names of team members in this shared spreadsheet.

STEP 2: SUBMIT STATEMENT OF WORK (PROJECT PROPOSAL)

Your Statement of Work should act as a blueprint for your project. It doesn't have to be extensive, but it should be clear and focused.

Components of the Statement of Work:

Title and Authors:

- **Title:** An engaging, relevant, and informative title that captures the essence of your project.
- **Authors:** Names of all team members and their respective email addresses.

Background and Motivation:

Provide a brief background on the topic you have chosen. Explain why you find it interesting or important, and mention any previous background, research interests, or readings that have influenced your choice.

Scope and Objectives:

Clearly outline the problem or question your project aims to solve. Make sure the scope is well-defined so that there is no ambiguity regarding your project's objectives.

Submission Guidelines:

- Length: 1-2 pages
- Format: PDF
- Submit via Canvas

STEP 3: DISCUSS DATA SOURCES

Data is the backbone of any data science project and therefore for any MLOps project, making it crucial to identify appropriate datasets for your endeavor. In your Statement of Work, you must address the

following aspects regarding data:

Source of Data:

- Identify where the data comes from (e.g., public repository, generated by the team, etc.).

Description of Dataset:

- Offer a brief overview of what the dataset contains. Is it time-series data, images, textual data, etc.?

Key Attributes:

- Describe the variables or features that are most relevant to your problem.

Relevance to the Project:

- Explain how the data is suited to solving the problem or question you've posed. Why is this data set useful or relevant?

Data Quality Concerns:

- If applicable, indicate any potential challenges related to data quality that you foresee (e.g., missing data, inconsistencies, or the need to merge multiple datasets). Mention your preliminary plan to tackle these issues.

Important Note:
Statements of Work that do not include information on available and relevant data will not be accepted.

STEP 4: DEFINE SCOPE AND PRELIMINARY DESIGN

The scope of your project is largely up to you and your team. Whether it's simple or complex, the aim should be to align with the course's learning objectives. However, for a project to be considered comprehensive, it should ideally include a few of the following minimum components:

Minimum Components for a Good Project:

- **Large or Heterogeneous Data:** Your project should involve a sizable or diverse dataset that requires careful handling and processing.
- **Scalability:** Consider how your solution will scale for many users, particularly in the application you intend to build.
- **Complex Models:** The project should explore models that are challenging to train, which will showcase your understanding of MLOps challenges.
- **Computationally Expensive Inference:** If your project involves inference models, they should be computationally intensive to align with real-world challenges.

Problem Statement:

- Clearly outline the problem or question your project addresses.

Objectives:

- List the primary goals or outcomes, which should align with your problem statement and the minimum components outlined above.

Learning Emphasis:

- Opt for models and methods that your team understands. The project should reflect your grasp of course concepts.

Application Mock Design:

- Include a preliminary design or sketch for the application you intend to develop. This could range from simple wireframes to a more detailed, clickable prototype.

Research and Development:

- Reference papers, blog posts, or other scholarly materials that aid your project and align with your objectives.

Fun Factor:

- The project should also be a space for you to enjoy both the subject matter and the developmental process.

Limitations and Risks:

- Discuss any anticipated challenges or limitations, such as data quality issues or technical constraints.

Milestones:

- List key milestones for both your project and application development. Include tentative deadlines, if possible.

Important Note:
Statements of Work that do not include both a well-defined scope and a preliminary design for the application will not be accepted.

Deliverables: Submit a PDF of your proposal on canvas

Below are two samples SOW for such apps:

Sample Proposal

ButterFlyer

Statement of Work for Project “”

AC215

Team Members

- [Pavlov Protovief]
- [Paolo Primopadre]
- [Pablo El Padron]

Contact Information

- [pavlos@pleasedonotemailme.com]

Problem Statement

To develop an application that can identify various species of butterflies in the wild using computer vision and offer educational content through a chatbot interface.

Minimum Components for a Good Project

- **Large Data:** Collection and utilization of a varied and substantial dataset of butterfly images.
- **Scalability:** Ability to handle multiple users querying the computer vision model and chatbot simultaneously.
- **Complex Models:** Use of deep learning-based computer vision models for accurate species identification.
- **Computationally Expensive Inference:** Implementation of efficient algorithms to minimize latency during species identification.

Objectives

1. Collect and preprocess a diverse dataset of butterfly images.
2. Develop a computer vision model to identify butterfly species.
3. Implement a scalable backend to handle multiple queries.
4. Design an intuitive and user-friendly frontend.
5. Integrate a chatbot for answering user questions about butterflies.

Learning Emphasis

The project will focus on employing a convolutional neural network for image recognition and natural language processing techniques for the chatbot, both areas covered in the course.

Application Mock Design

The application will feature two main interfaces:

1. A camera interface for capturing butterfly images.
2. A chatbot interface to interact with users. (Additional wireframes or mock-ups can be attached).

Objectives

1. **Collect and preprocess a diverse dataset of butterfly images:**
   - **Data Source:** The dataset will primarily come from public repositories like [X Dataset Source] and [Y Dataset Source]. We will also supplement this with images captured during field trips and those sourced from citizen science platforms.
   - **Data Attributes:** Images should ideally be labeled with the species name, geographic location, and date of capture.

- **Data Relevance:** The comprehensiveness of the dataset is vital for training a robust computer vision model capable of identifying a wide range of butterfly species.

(Note: If acquiring new images, all team members should adhere to ethical guidelines concerning wildlife photography and data collection.)

- **Source:** Data will be collated from a combination of open-source databases, user-generated content from platforms like iNaturalist, and field data collection by team members.
- **Description:** The dataset will comprise images of various butterfly species, ideally annotated with species names and other metadata like geographic location and date.
- **Key Attributes:** The dataset should include high-resolution images suitable for computer vision algorithms, along with associated metadata to enrich the model's understanding.
- **Relevance:** The dataset is fundamental to train the computer vision model to identify different species of butterflies accurately.
- **Data Quality:** We anticipate that some images might be poorly labeled or low in quality. These will either be cleaned or supplemented with additional data.

Research and Development

We will review literature and open-source projects related to computer vision in biological classification and natural language processing for educational chatbots.

Fun Factor

Exploring the intersection of biology and technology, while learning about butterflies, makes this project particularly engaging.

Limitations and Risks

- Possible challenges in obtaining a large and diverse enough dataset.
- Computational limitations when deploying complex models.

Milestones

1. Data collection and preprocessing: [Tentative Deadline]
2. Computer vision model development: [Tentative Deadline]
3. Backend implementation: [Tentative Deadline]
4. Frontend development: [Tentative Deadline]
5. Chatbot integration: [Tentative Deadline]
6. Final testing and deployment: [Tentative Deadline]