SI 618 Team Project Proposal

## AIGC Prompts Category Classifier

## 1. Team members

* **Yan Lu** (yanlunar) *section 001*
* **Yifan Li** (ivanlyf) *section 001*

## 2. Project summary

| **Background**  When using the AIGC (Artificial Intelligence Generated Content) platform, crafting effective prompts is a critical step in generating high-quality content. Prompts serve as the input or guidance for the AI model to generate meaningful responses or content. Good prompts are clear and structured instructions, questions, or cues provided to AI systems and can lead to better production of the generated content.  However, even though there are hundreds of AIGC models and application tools for generating text, images, and even videos (e.g. GPT, Llama, Stable Diffusion, Midjourney…), online guidance for producing high-quality prompts is few. People may find it hard to find appropriate and high-quality prompts based on their needs in a certain **category**.  Here, we propose a classifier to categorize AIGC prompts, which can correctly and precisely classify prompts into appropriate categories so as to make it easier for people to refer to. |
| --- |

## 3. Datasets

## 3.1 Primary dataset description

| We have already scrapped the primary dataset from the website. In our database, we organize and store data preparations by themes. Currently, there are 7,432 prompts with different themes, with Nature occupying the largest share. More themes will be added in the future.   * Name: Promptlibrary prompt datasets * Estimated Size: 20MB * Location: <https://promptlibrary.org/> * Format: CSV * Access Method: Web crawling |
| --- |

## 3.2 Secondary dataset description

| The dataset consists of 192, 280 high-quality diffusion prompts. The key features include id, prompts, urls of the source sites, .etc.   * Name: 900k Diffusion Prompts Dataset * Estimated Size: 247.34MB * Location: <https://www.kaggle.com/datasets/tanreinama/900k-diffusion-prompts-dataset> * Format: CSV * Access Method: Download |
| --- |

## 

## 3.3 [ Yes ] Affirm: datasets are public.

## 4. Cleaning and manipulation

| **Data Inspection:**  Start by inspecting the data to identify any anomalies, missing values, or inconsistencies. This may involve checking for duplicate entries, outliers, and errors in the data. We will not simply join the two datasets, but create some groups and do experiments to compare different kinds of categorying methods. For the sake of performance, we will also separate our data into several training sets and test sets. Several methods for division, including Cross-Validation and hosting another Validation Set, may be considered.  **Data Preprocessing:**   * Handling Missing Data: Decide on a strategy for handling missing data. * Removing Duplicates: Identify and remove duplicate entries to maintain data integrity. * Outlier Detection: Detect and handle outliers that may affect the categorization process.   **Text Data Processing:**   * Text Cleaning: Clean and preprocess textual data by removing special characters, lowercasing, or removing stopwords. * Feature Extraction: Convert text into numerical features using techniques like TF-IDF or word2vec word embeddings.   **Categorization Strategy:**   * Define the categories or labels to assign to the data. * Choose a categorization method, such as supervised learning (classification), unsupervised learning (clustering), or rule-based categorization.   **Model Training:**   * Annotate a subset of the data for training and evaluation if using a machine learning approach. * Train a categorization model using the preprocessed data and appropriate machine learning algorithms. * Fine-tune the model for better performance and generalization. |
| --- |

## 5. Analysis

| **Evaluation Metrics:** We will use evaluation metrics like F-1 Score, MAP, and NDCG to evaluate our classifier. Since all of them require some ground truth about how, we have to prepare a set of prompt-category relevance scores.  **Expected Outcome:** A classifier to classify AIGC prompts into appropriate categories, with high precision and recall performance. |
| --- |

## 6. Visualizations

| First, create a bar chart illustrating the distribution of number of prompts by category, with categories on the x-axis and numbers on the y-axis. This visualization will help identify the most and least profitable product categories.  Also, a bar chart showing within a certain category, the ranking of the number of keywords.  Additionally, I intend to create a heatmap that shows the correlation between all the features in a certain category. |
| --- |

## 7. Ethical considerations

Does your choice of data raise any ethical issues? If so, briefly describe the concern and how you plan to mitigate it.

| The selection of data for the prompt categorizer raises ethical concerns related to data privacy, especially if the dataset contains sensitive information. To address these issues, we intend to prioritize data pre-processing, ensuring the removal or encryption of any sensitive data. |
| --- |

## 8. Contributions

Indicate the contribution that each team member will make to the project.

| Yan Lu: Data preprocessing, data analysis, data visualization and report  Yifan Li: Data analysis, web crawling, model improvement, report |
| --- |