Milestone 1: Project Initialization and Planning Phase

The "Project Initialization and Planning Phase" for the SMS Spam Detection project focuses on setting up the groundwork for identifying and filtering spam messages using NLP techniques. This phase defines the project's goals, scope, and stakeholders, allocates resources, and outlines timelines. It also involves assessing risks and planning mitigation strategies to ensure a smooth project execution. Proper initiation guarantees alignment among team members and establishes a roadmap for successful delivery.

Activity 1: Define Problem Statement

- **Problem Statement:** SMS spam messages pose a significant challenge to mobile users, causing inconvenience and potential security risks. Identifying and filtering these messages requires an effective classification model that can differentiate between spam and legitimate messages using textual analysis.
  - **Github link :** https://github.com/SubhashMishra700/SMS-Spam-Detection-Using-NLP/commit/5813900c3cf8dd72b03f419374f4a6e01f5ceae4

  - **Activity 2: Project Proposal (Proposed Solution)**

- **Proposed Solution:** The proposed project, "SpamFilter NLP Classifier," aims to utilize natural language processing techniques to build an efficient and scalable SMS spam detection model. The system will analyze text data from SMS messages, extract relevant features, and classify messages as spam or ham (legitimate). By leveraging pre-processing techniques, feature engineering, and machine learning algorithms, this project will provide a robust solution to minimize spam and enhance user experience.
- **Github link :** https://github.com/SubhashMishra700/SMS-Spam-Detection-Using-NLP/tree/main/Project%20Initialization%20and%20planning%20phase

Activity 3: Initial Project Planning

- **Description:**
  Initial project planning for the SMS spam detection system involves:

○ Setting objectives: Building a model capable of achieving high accuracy in spam detection.

○ Defining the scope: Ensuring the system works across diverse SMS datasets and language nuances.

○ Identifying stakeholders: Project team members, end-users, and relevant data providers.

○ Planning workflow: Establishing steps for data collection, preprocessing, feature extraction, and model building.

○ Setting timelines: Defining deliverables and milestones for each phase of the project.

○ Allocating resources: Identifying tools, libraries, and computational resources required.

- Clear initial planning ensures a structured approach for the successful implementation of the SMS Spam Detection system.
  - **Github link :** https://github.com/SubhashMishra700/SMS-Spam-Detection-Using-NLP/tree/main/Project%20Initialization%20and%20planning%20phase

  - **Milestone 2: Data Collection and Preprocessing Phase**

The **Data Collection and Preprocessing Phase** is crucial for building an SMS spam detection system. It involves gathering high-quality SMS datasets and preparing them for analysis through cleaning, encoding, and preprocessing steps. This phase establishes a reliable foundation for feature extraction and model development. **Activity 1: Raw Data Sources and data Quality**

- **Description:**

  The dataset for the "SpamFilter NLP Classifier" project will be sourced from platforms like Kaggle or UCI ML Repository, featuring SMS messages labeled as spam or ham. Steps include:
- Verifying dataset quality (e.g., format consistency and presence of labels).
- Addressing missing data and duplicates.
- Ensuring adherence to ethical guidelines for data usage, such as anonymization.
- **Outcome:** Reliable and diverse SMS data for effective spam classification.
- **Github link :** https://github.com/SubhashMishra700/SMS-Spam-Detection-Using-NLP/tree/main/Project%20Initialization%20and%20planning%20phase

Activity 2: Data Quality Report

- **Description:**

The Data Quality Report assesses the SMS dataset to ensure:

■ No missing or corrupt entries.

■ A balanced representation of spam and ham messages. ■ Correct and consistent text encoding (e.g., UTF-8).

Cleaning steps include token normalization, punctuation removal, and filtering out non-textual elements.

- **Github link :** https://github.com/SubhashMishra700/SMS-Spam-Detection-Using-NLP/tree/main/Project%20Initialization%20and%20planning%20phase

Activity 3: Data Exploration and Preprocessing

- **Description:**

■ **Exploration:** Analyze the dataset to identify word distributions, message lengths, and spam keywords.

- ■ **Preprocessing:** Steps include:
  - ○ Tokenizing text data.
  - ○ Lowercasing for consistency.
  - ○ Removing stop words and stemming/lemmatization to reduce noise.
  - ○ Converting text into numeric formats like TF-IDF or embeddings.
- **Outcome:** A well-processed dataset ready for feature extraction and model building. ● **Github link:** https://github.com/SubhashMishra700/SMS-Spam-Detection-Using-NLP/tree/main/Project%20Initialization%20and%20planning%20phase

Milestone 3: Model Development Phase

The **Model Development Phase** focuses on building and evaluating machine learning models for SMS spam detection using preprocessed data. This phase emphasizes feature selection, model selection, training, and performance evaluation to ensure optimal results.

Activity 1: Feature Selection Report

- **Description:**
  - ■ Features like word frequency, n-grams, and TF-IDF scores are selected based on their relevance to spam detection.
  - ■ Pretrained embeddings (e.g., Word2Vec or GloVe) may be used for deeper insights into message semantics.
  - ■ The report evaluates each feature's impact on classification accuracy.
- **Github Link:** https://github.com/SubhashMishra700/SMS-Spam-Detection-Using-NLP/tree/main/Model%20Development%20Phase

- **Activity 2: Model Selection Report**

- **Description:**
  - ■ Models like Naive Bayes, Logistic Regression, Random Forest, SVM, and LSTM are considered.
  - ■ Model selection criteria include:
    - ○ Handling of high-dimensional text data.
    - ○ Speed and scalability.
    - ○ Performance on the dataset (accuracy, precision, recall).
  - ■ The report documents the rationale behind choosing specific models for experimentation.
- **Github Link:** https://github.com/SubhashMishra700/SMS-Spam-Detection-Using-NLP/tree/main/Model%20Development%20Phase
- **Activity 3: Initial Model Training Code, Model Validation, and Evaluation Report**

- **Description:**
  - ■ Initial training involves applying selected algorithms to the processed dataset.
  - ■ Validation techniques (e.g., k-fold cross-validation) are used to evaluate models.
  - ■ Metrics like accuracy, precision, recall, F1-score, and ROC-AUC are calculated to assess performance.
  - ■ The final report documents the training process, hyperparameter tuning, and evaluation results.
- **Github link :** https://github.com/SubhashMishra700/SMS-Spam-Detection-Using-NLP/tree/main/Model%20Development%20Phase

- **Milestone 4: Model Optimization and Tuning Phase**

The **Model Optimization and Tuning Phase** focuses on enhancing the performance of the SMS spam detection model by refining algorithms and parameters. This phase includes fine-tuning hyperparameters, comparing key metrics, and selecting the best-performing model for robust spam detection. The goal is to achieve optimal predictive accuracy and efficiency.

Activity 1: Hyperparameter Tuning Documentation

- **Description:**
  - ■ The **Support Vector Machine (SVM)** and **Random Forest** models were selected for their strong initial performance in text classification tasks.
  - ■ **Hyperparameter Tuning** steps include:
    - ○ For SVM: Optimizing kernel type, regularization parameter (C), and gamma.
    - ○ For Random Forest: Tuning the number of trees, maximum depth, and minimum samples split.
  - ■ Grid search and Bayesian optimization techniques were used to identify the best parameter combinations, ensuring improved accuracy and reduced overfitting.
- **Outcome:** SVM achieved better spam classification accuracy post-tuning, aligning with the project's objective to minimize misclassification.

Activity 2: Performance Metrics Comparison Report

- **Description:**
  - ■ A detailed comparison was performed between baseline models and their optimized counterparts.
  - ■ Metrics evaluated include:
    - ○ **Accuracy:** Proportion of correctly classified messages.
    - ○ **Precision:** Spam identification accuracy.
    - ○ **Recall:** Ability to detect all spam messages.

○ **F1-Score:** Harmonic mean of precision and recall.

■ The report highlights that SVM outperformed other models after tuning, achieving an F1-score improvement of 5%.

- **Outcome:** A comprehensive analysis of model improvements, showcasing the enhanced predictive performance of the final model.

Activity 3: Final Model Selection Justification

- **Description:**

■ The **Final Model Selection Justification** outlines the rationale for choosing **SVM** as the ultimate model. ■ Key reasons include:

○ Exceptional performance across all metrics, especially F1-score and recall, critical for spam detection.

○ Robustness in handling textual data with balanced spam and ham classes.

○ Efficiency in real-time classification scenarios post-optimization.

■ The report also emphasizes SVM's scalability and minimal risk of overfitting after hyperparameter tuning.

- **Outcome:** SVM was selected as the final model due to its alignment with the project objectives of high accuracy, efficiency, and reliability.
- **Github link :** https://github.com/SubhashMishra700/SMS-Spam-Detection-Using-NLP/tree/main/Model%20Development%20Phase

Milestone 5: Project Files Submission and Documentation

For project file submission in Github, Kindly click the link and refer to the flow.
https://github.com/SubhashMishra700/SMS-Spam-Detection-Using-NLP.git
Click Here For the documentation, Kindly refer to the link. Click Here
https://github.com/SubhashMishra700/SMS-Spam-Detection-Using-NLP/tree/main/Documentation%20%26%20Demonstration

Milestone 6: Project Demonstration

For Demonstration refer to this link
https://github.com/SubhashMishra700/SMS-Spam-Detection-Using-NLP/tree/main/Documentation%20%26%20Demonstration