Problem Statement: #1 - File Extension Categorization

"Developing a Robust Classification Model for Document Categorization Based on File Extensions"

The objective of this academic project is to develop a robust classification model that accurately identifies the category of a document based on its file extension. For example, given a file with the extension ".docx", the model should classify it as belonging to the "Word Documents" category and ".jpg" indicates an image file and can be classified as the "Image Files" category. The project will involve collecting extension data from public sites and utilizing it to train a machine learning model. The model will be designed to accept a file extension as input and predict the corresponding category as output. The project aims to explore different machine learning models, train them with internet data, and incorporate feedback to continuously improve the model's accuracy. The final deliverable will include implementation details, approach documentation, and assumptions made during the project.

Objectives

1. Identify Suitable Machine Learning Models:

Evaluate different machine learning models, such as decision trees, random forests, support vector machines, or neural networks, to determine the most effective approach for classifying document categories based on file extensions.

2. Train the Model with Internet Data:

Collect extension data from public sites and use it as the training dataset for the machine learning model. Label the data with the corresponding document categories to enable the model to learn the patterns and relationships between extensions and categories.

3. Input-Output Mapping:

Develop a model that accepts a file extension as input and accurately predicts the document category as output. This mapping will provide a quick and efficient classification solution based solely on the file extension.

4. Feedback-Driven Model Improvement:

Incorporate a feedback loop mechanism to continuously improve the classification model. Gather user feedback on the accuracy of the model's predictions and use it to retrain the model periodically, incorporating new data and refining the classification capabilities.

5. Model Maturity through Correctness:

As the model receives feedback and is retrained with accurate data, it will mature and improve its classification accuracy. The model's performance will increase as more correct classifications are obtained.

6. Implementation and Approach Documentation:

Provide a comprehensive document outlining the implementation details and approach taken to develop the classification model. This documentation will include information on data collection, model training techniques, evaluation metrics, and assumptions made during the project.

7. Assumptions:

Make certain assumptions, such as assuming the collected extension data from public sites is reliable and representative of the document categories. Also, assume that the file extensions themselves provide sufficient information to accurately classify the documents into predetermined categories.

Additional Applicable Considerations

8. Data Preprocessing: Consider preprocessing the extension data to handle any inconsistencies, missing values, or outliers. This may involve data cleaning, normalization, or feature engineering techniques to enhance the model's performance.

9. Feature Selection: Explore different feature selection methods to identify the most relevant extension attributes that contribute to accurate document category classification. This can help reduce dimensionality and improve the efficiency and effectiveness of the model.

10. Model Evaluation: Utilize appropriate evaluation metrics, such as accuracy, precision, recall, or F1-score, to assess the performance of the classification model. Consider using techniques like cross-validation or holdout validation to ensure reliable evaluations.

11. Deployment and Scalability: Consider the deployment and scalability aspects of the developed model. Explore techniques to optimize the model's performance and ensure it can handle large-scale classification tasks efficiently.

By addressing these objectives and considerations, this project aims to create an effective and efficient system for accurately classifying document categories based on their file extensions.

Problem Statement: #2 - Behavior Tracking

"**Developing an Anomaly Detection System for detecting anomalous behavior based on document categories."**

Problem Statement:

The objective of this academic project is to develop an anomaly detection system for file type categorization, focusing on tracking the behavior of different file types across various categories. The system will monitor the addition, update, and deletion of files in a file management system, utilizing a dataset that captures the frequencies and patterns of these actions for each file type category. By analyzing the behavior of file type categorization, the system aims to identify anomalies and improve the accuracy of file classification. The system will utilize a suitable machine learning model trained on historical file data to learn normal behavior and distinguish anomalies. Real-time file data will be analyzed to flag potential anomalies that deviate significantly from the expected patterns. Feedback from users or administrators will be incorporated to continuously improve the accuracy of the model. Overall, the project aims to develop an effective anomaly detection system that enhances file type categorization processes and ensures proper file management.

Details:

In file management systems, accurately categorizing files based on their types is essential for efficient organization and retrieval of information. By tracking the behavior of different file types across various categories, companies can gain insights into the changes happening within their file repositories. Here's how this process works:

File Addition Analysis: Tracking the addition of new files allows companies to understand the growth and evolution of their file collections. By analyzing the frequency and patterns of file additions, they can identify trends and anticipate potential changes in file volume. For example, they may observe a sudden increase in file additions in specific categories during certain projects or periods of high activity. This analysis helps companies allocate resources, plan for storage capacity, and ensure proper categorization of newly added files.

File Update Analysis: Monitoring the updates made to existing files provides insights into the maintenance and modification of information. By analyzing the frequency and nature of file updates, companies can identify patterns and potential issues. For instance, if there is a significant increase in updates to files in a particular category, it could indicate ongoing revisions or corrections. This analysis helps companies prioritize file review, ensure accuracy and relevance of information, and identify the need for version control.

File Deletion Analysis: Analyzing the deletion of files helps companies understand the removal and retention of information. By tracking the frequency and patterns of file deletions, they can identify potential anomalies or unauthorized removals. For example, if there is an unexpected surge in file deletions in a specific category, it could indicate data loss or security breaches. This analysis helps companies ensure data integrity, enforce file retention policies, and detect any abnormal deletion behavior.

Anomaly Detection: By comparing the current behavior of file type categorization to historical data, companies can identify anomalies. For instance, if there is a sudden increase in the number of file additions or deletions compared to the average, it could indicate abnormal file management behavior. Anomalies can also be detected by analyzing the patterns of file updates, such as unusual update frequencies or unexpected modifications. This analysis helps companies detect and address any irregularities, ensure data accuracy, and maintain proper file type categorization.

The project will address the following key aspects:

1. Data Collection and Preprocessing: The project will involve collecting and preprocessing data that captures the behavior of file type categorization, including the frequency and nature of file additions, updates, and deletions for each file type category. The dataset will be structured to facilitate anomaly detection and analysis.

2. Machine Learning Model Selection: The project will explore different machine learning models suitable for anomaly detection in file type categorization behavior. This may include time series analysis, clustering, supervised learning, or hybrid approaches. The models will be evaluated based on their ability to capture the frequencies and patterns of file behavior across different categories and select the most appropriate model for the anomaly detection task.

3. Model Training: The selected machine learning model will be trained on historical file data, learning the normal behavior of file type categorization based on the frequencies and patterns of additions, updates, and deletions for each file type category.

4. Anomaly Detection and Flagging: The trained model will be applied to real-time file data to detect anomalies. The system will compare the behavior of file type categorization to the learned normal patterns and flag any deviations that significantly deviate from the expected behavior. This real-time anomaly detection will enable companies to take immediate action and address potential issues in file management.

5. Feedback and Model Improvement: The system will be designed to incorporate feedback from users or administrators. This feedback will be used to continuously improve the accuracy and effectiveness of the machine learning model. By leveraging the feedback, the system will refine its understanding of anomalies and enhance its detection capabilities over time.