**Project**

**Project Task Breakdown**

1. **Project Topic Selection**

Define the scope by choosing a specific domain for web scraping and big data analysis, such as big data, computer network, intelligent traffic, autonomous vehicle, social media, e-commerce, or news websites. Establish clear project goals, such as collecting specific data through scraping, analyzing it, and applying machine learning models for predictions.

1. **Data Scraping**

Design the scraper by selecting a suitable web scraping framework, such as Scrapy, Beautiful Soup, or Requests. Identify target websites and their data structures and develop a scraping strategy. Implement the scraper by writing code to gather data, including product information, user reviews, or social media posts. Ensure to handle anti-scraping mechanisms by setting appropriate request intervals and using proxies if necessary. Choose a storage format for the scraped data, such as CSV, JSON, or a database, and ensure data integrity and accuracy throughout the process.

1. **Data Cleaning and Preprocessing**

Perform data cleaning by handling missing values and outliers, and formatting the data appropriately, such as adjusting dates and numerical types. This step is critical for ensuring the quality of the dataset before analysis.

1. **Feature Engineering**

Identify relevant features that impact the model and consider any derived features that may enhance the analysis. This could include techniques like keyword extraction or sentiment analysis.

1. **Machine Learning Modeling**

Split the dataset into training and testing sets to facilitate model training. Train the selected model and tune hyperparameters based on the data characteristics. Evaluate the model using suitable metrics, such as accuracy, F1-score, or mean squared error, to assess performance. Perform cross-validation to ensure the model's stability.

1. **Result Analysis and Visualization**

Present the results by visualizing model predictions in comparison to actual values. Analyze feature importance to identify key factors influencing the model's outcomes. Summarize the key findings and conclusions drawn from the data analysis and discuss any limitations of the model along with suggestions for improvement.

**Example Topics**

1. **Social Media Sentiment Analysis**: Scrape data from platforms like Twitter or Reddit to analyze user sentiment towards specific events.
2. **E-commerce Price Prediction**: Scrape product prices and reviews from e-commerce websites and utilize machine learning to predict price changes.
3. **News Trend Analysis**: Scrape articles from news websites to explore trending topics and their changes over time.
4. **Online Review Analysis**: Scrape product reviews to perform sentiment analysis and investigate factors influencing user ratings.

**Evaluation Criteria**

1. **Project Understanding (20%)**: Assessing the clarity of project goals and the relevance of the chosen topic.
2. **Scraping Implementation (10%)**: Evaluating the effectiveness and efficiency of the web scraper, including the handling of anti-scraping measures.
3. **Data Cleaning and Preprocessing (20%)**: Measuring the thoroughness in data cleaning steps and the overall quality and integrity of the cleaned dataset.
4. **Feature Engineering (20%)**: Evaluating the relevance and creativity in feature selection.
5. **Modeling and Evaluation (30%)**: Assessing the appropriateness of the chosen machine learning models and the robustness of the evaluation methods and metrics used. It is also important to provide the collected data and the corresponding code.