YouTube Data Analysis and Prediction Project README

Introduction
This project focuses on extracting, analyzing, and predicting trends based on YouTube data. It encompasses modules for data extraction, exploratory data analysis (EDA), and machine learning predictions related to YouTube trending videos.

Prerequisites to Run the Project
- Python 3.x: A Python 3 environment is required.
- Jupyter Notebook: For running and viewing .ipynb files.
- API Key: A valid YouTube Data API key is needed.
- Libraries:
    - googleapiclient.discovery for YouTube API interaction.
    - pandas, numpy for data manipulation.
    - matplotlib, seaborn for data visualization.
    - sklearn for machine learning tasks.
    - Other libraries for data handling and analysis as used in the notebooks.

Running the Application
- Install all the prerequisites as mentioned.
- Ensure that the input data is present in the 'input' directory. This data is crucial for the proper functioning of the application and should be extracted from YouTube as per the first two modules.
- Run the Jupyter notebooks sequentially to maintain the flow of data and dependencies:
    - youtube_category_data_extraction_script.ipynb
    - youtube_trending_videos_data_extraction_script.ipynb
    - data_analysis_of_trending_videos.ipynb
    - ML_predictions_from_yt_data.ipynb

- Execute all cells in each notebook for complete data processing and analysis

Modules of the Code:

1) YouTube Category Data Extraction
This module is responsible for extracting video category data from YouTube. Here's an overview of its functionality:

API Initialization: It initializes the YouTube Data API client using the provided API key.
Data Fetching: The script fetches all YouTube video categories for a specified region (default is 'US').
Data Formatting and Saving: The fetched data is formatted into a desired JSON structure and saved to a file named 'US_category_id.json'.

2) YouTube Trending Videos Data Extraction
This module is focused on fetching and saving YouTube trending video data for multiple countries. The functionality can be summarized as follows:

Initialization and Setup: It uses the YouTube Data API client initialized with a valid API key.
Country Selection: The script is set up to fetch trending data for a predefined list of countries.
Data Storage: It creates a directory ('input') for storing CSV files containing the trending video data.
Data Extraction and Writing: For each country, the script fetches data on trending videos, including details like video ID, title, publication date, view count, likes, etc. This data is then saved into a CSV file for each country.
Duplication Avoidance: The script checks if data for the current day has already been fetched to avoid duplication.

YouTube Data Analysis and Prediction Project README

Libraries and Modules:
googleapiclient.discovery for accessing YouTube API.
csv for handling CSV file operations.
os for directory and file path operations.
datetime for date and time manipulation.

3) Data Analysis of Trending Videos
This module focuses on performing exploratory data analysis (EDA) on the trending YouTube videos dataset. The functionalities include:

Data Loading and Cleaning: It reads the trending video data from CSV files, cleans the data (e.g., handling NaN values, date formatting), and prepares it for analysis.
Analysis Techniques: The script performs various analysis tasks, such as:
Trend analysis based on video categories, publish and trending dates.
Generating word clouds for different video categories.
Analyzing the frequency of videos by different channels.
Visualization: It uses data visualization techniques to present insights from the data, including bar graphs and word clouds.

Libraries and Modules:
pandas for data manipulation and analysis.
matplotlib and seaborn for data visualization.
datetime, re, nltk for handling dates, regular expressions, and natural language processing tasks.

4) Machine Learning Predictions from YouTube Data
This module is dedicated to applying machine learning techniques to predict various aspects of YouTube trending videos. Key functionalities include:

Data Preprocessing: The module handles data loading, cleaning, and preprocessing necessary for machine learning.
Feature Engineering: It involves creating new features from the existing data to improve the predictive models.
Model Training and Evaluation: Various machine learning models are trained and evaluated to find the best-performing model. These models include linear regression, decision tree, and gradient boosting among others.
Predictions: The module makes predictions on the view velocity of videos and the number of days a video might trend. These predictions are visualized using bar plots and other visualizations.

Libraries and Modules:
pandas for data manipulation.
matplotlib for data visualization.
sklearn for machine learning tasks, including model training, evaluation, and prediction.
Other data preprocessing and machine learning-related libraries as used in the notebook.