Youtube Trending video analysis

Problem Description :

Background: YouTube is one of the world's most popular video-sharing platforms, with millions of videos uploaded and watched every day. The YouTube trending section features videos that have gained significant popularity and are trending among viewers. Analyzing the characteristics of trending videos can provide valuable insights into viewer preferences, content trends, and channel performance.

Dataset Information: The dataset used for this analysis is a CSV file containing information about trending videos in the United States. It includes various attributes for each video, such as video ID, title, description, publishing time, view count, like count, comment count, video category, channel title, and more. The dataset allows us to explore the factors that contribute to a video's popularity and understand patterns in trending videos.

Problem Statement: The objective of this project is to perform a comprehensive analysis of trending videos on YouTube in the United States. By analyzing the dataset, we aim to gain insights into the following aspects:

1. **Video Popularity:** Analyze the distribution of views, likes, and comments for trending videos and identify factors influencing video popularity.
2. **Video Categories:** Investigate the distribution of trending videos across different categories and identify the most popular video categories.
3. **Channel Performance:** Analyze the performance of channels with the most trending videos and identify influential content creators.
4. **Publishing Time Analysis:** Examine the relationship between the publishing time of videos and their chances of trending.

5. **Sentiment Analysis:** Perform sentiment analysis on video titles and descriptions to understand the overall sentiment of trending videos.
6. **Time Series Analysis:** Explore time series patterns in video views, likes, and comments to identify trends and seasonality.
7. **Geospatial Analysis:** Analyze the geographic distribution of trending videos and their popularity in different regions.
8. **User Engagement:** Analyze user engagement metrics, such as the ratio of likes to dislikes and comments to views, to understand viewer preferences.

Expected Outcomes: Through this analysis, we expect to gain valuable insights into the factors that contribute to a video's popularity on YouTube. By identifying trends and patterns, content creators and marketers can optimize their video content and engagement strategies to increase the chances of their videos trending. The findings can also be used to understand viewer preferences, improve content recommendations, and enhance user experiences on the platform.

Possible Framework :

1. **Import Libraries and Load Data:**
- Import necessary libraries like Pandas, NumPy, Matplotlib, Seaborn, and WordCloud.
- Load the dataset "USvideos.csv" containing information about YouTube trending videos in the United States using Pandas.
2. **Data Preprocessing:**
- Explore the dataset to understand its structure and check for missing values.
- Handle missing values in the "description" column by filling them with empty strings.
- Extract the year from the "trending_date" column to create a new "year" column for further analysis.
- Drop irrelevant columns like "Lattitude" and "Longitude."
- Convert the "trending_date" column to a proper date format and extract the publishing day and hour from the "publish_time" column.
3. **Data Visualization and Analysis:**
- Analyze the distribution of videos over the years and visualize it using a bar plot.
- Explore the distribution of video views and likes using histograms.
- Calculate the percentage of videos with views less than 1 million.
- Visualize the distribution of video likes and calculate the percentage of videos with likes less than 40,000.
- Visualize the distribution of comment counts and calculate the percentage of videos with comment counts less than 4,000.
- Analyze the correlation between numerical features using a heatmap.
4. **Text Analysis:**
- Check if video titles contain capitalized words and visualize the result using a pie chart.
- Calculate the percentage of video titles containing capitalized words.
5. **WordCloud Analysis:**
- Generate a WordCloud for video titles to visualize the most common words in trending video titles.
6. **Channel and Category Analysis:**
- Analyze the top 20 channels with the most trending videos and visualize the result using a bar plot.

- Extract the video category name from the "category_id" and visualize the distribution of trending videos across different categories.

7. **Publishing Time Analysis:**

- Visualize the distribution of trending videos based on the day of the week they were published.
- Visualize the distribution of trending videos based on the hour of the day they were published.

8. **Video Status Analysis:**

- Visualize the distribution of videos based on whether they are flagged as "video error or removed," "comments disabled," or "ratings disabled."

9. **Conclusion and Recommendations:**

- Summarize the key findings from the analysis of YouTube trending videos.
- Provide recommendations to content creators and marketers on how to optimize video content and engagement strategies to increase the chances of their videos trending.

10. **Additional Data Analysis (Optional):**

- Perform time series analysis to identify trends and seasonality in video views, likes, and comments.
- Conduct geospatial analysis to understand the geographic distribution of trending videos and their popularity in different regions.

11. **Future Work and Improvements (Optional):**

- Suggest potential areas of improvement for the analysis.
- Propose additional research and data sources that can enhance the understanding of YouTube trending video trends and patterns.

12. **Final Remarks:**

- Conclude the project with final remarks and acknowledgments.

Code Explanation :

*If this section is empty, the explanation is provided in the .ipynb file itself.

Importing Libraries and Loading Data: The code begins by importing necessary libraries like Pandas, NumPy, Matplotlib, Seaborn, and WordCloud. These libraries help us handle and visualize data effectively. Then, the code loads the dataset "USvideos.csv" into a Pandas DataFrame, which contains information about YouTube trending videos in the United States.

Data Preprocessing: Before diving into analysis and visualization, it's essential to clean and prepare the data. The code starts by exploring the dataset to understand its structure and check for missing values. It handles missing values in the "description" column by filling them with empty strings. Next, it extracts the year from the "trending_date" column and creates a new "year" column for further analysis. Some irrelevant columns like "Lattitude" and "Longitude" are dropped to focus on relevant information. The "trending_date" column is converted to a proper date format, and the publishing day and hour are extracted from the "publish_time" column for later analysis.

Data Visualization and Analysis: Data visualization is an essential aspect of data analysis. The code uses various plots, such as bar plots and histograms, to visually represent the data distribution. It analyzes the number of videos over the years, the distribution of views, likes, and comment counts. The code also calculates the percentage of videos with views, likes, and comment counts below specific thresholds. Furthermore, it examines the correlation between numerical features using a heatmap to understand their relationships.

Text Analysis and WordCloud: YouTube video titles play a crucial role in attracting viewers. The code checks if video titles contain capitalized words, and the results are visualized using a pie chart. Additionally, the code generates a WordCloud to display the most common words in trending video titles.

Channel and Category Analysis: The code conducts an analysis of the top 20 channels with the most trending videos. It also extracts the video category names from the "category_id" and visualizes the distribution of trending videos across different categories.

Publishing Time Analysis: The code visualizes the distribution of trending videos based on the day of the week and the hour of the day they were published. This helps understand the optimal timing for video publishing.

Video Status Analysis: The code visualizes the distribution of videos based on whether they are flagged as "video error or removed," "comments disabled," or "ratings disabled." This analysis provides insights into potential issues that may affect a video's performance.

Conclusion and Recommendations: At the end of the code, a summary of key findings and recommendations is provided based on the analysis. This helps content creators and marketers optimize their video content and engagement strategies.

Additional Data Analysis (Optional): The code suggests optional additional analyses that can be performed, such as time series analysis and geospatial analysis, to gain deeper insights into video trends and patterns.

Future Work and Improvements (Optional): The code may suggest potential areas for improvement or additional research that can enhance the analysis further.

Final Remarks: The code concludes the project with final remarks and acknowledgments.

Future Work :

Analyzing YouTube trending videos is an exciting project, and there are several ways to enhance the analysis and gain deeper insights. Here's a step-by-step guide on how to implement future work for this project:

Step 1: Sentiment Analysis One valuable addition to the project is sentiment analysis. By analyzing the comments on trending videos, we can determine the overall sentiment of the audience towards the video. This can be achieved by using Natural Language Processing (NLP) techniques and pre-trained sentiment analysis models. The steps for implementation are as follows:

- Use NLP libraries like NLTK or spaCy to process the comments and extract text features.
- Employ pre-trained sentiment analysis models to classify comments into positive, negative, or neutral sentiment.
- Visualize the sentiment distribution of the comments for each trending video to understand audience reactions better.

Step 2: Influencer Analysis Another interesting aspect to explore is influencer analysis. Identify top influencers or content creators whose videos trend frequently and analyze their common characteristics. The steps for implementation are as follows:

- Create a new column to identify influencers based on the number of trending videos they have.
- Analyze the common themes, titles, and categories among trending videos created by influencers.
- Visualize the distribution of influencers and their trending videos to understand their impact on the platform.

Step 3: Time Series Analysis Performing time series analysis can reveal the trending patterns and trends over time. This can help predict the best times to publish videos and understand seasonal trends. The steps for implementation are as follows:

- Use time series analysis techniques to identify trends and seasonal patterns in the number of trending videos over different months or years.
- Create time series plots to visualize the trends and identify any recurring patterns.

- Analyze the performance of videos based on their publishing time to optimize content release strategies.

Step 4: Video Length and Engagement Investigate the relationship between video length and audience engagement. Analyze how video length affects the number of views, likes, comments, and overall performance. The steps for implementation are as follows:

- Group trending videos into categories based on their length (e.g., short, medium, long).
- Calculate the average number of views, likes, and comments for each category.
- Visualize the engagement metrics across different video lengths to understand the audience's preference for video duration.

Step 5: Geospatial Analysis Geospatial analysis can provide insights into regional preferences and audience behavior. Analyze how trending videos vary across different regions or states. The steps for implementation are as follows:

- Use geospatial data or APIs to map the distribution of trending videos across different regions.
- Analyze the top categories and themes that trend in specific regions to identify regional preferences.
- Visualize the geospatial data using maps to understand regional video trends.

Step 6: Content Recommendation Build a content recommendation system based on user preferences and past viewing history. The recommendation system can suggest relevant trending videos to users, improving user experience and engagement. The steps for implementation are as follows:

- Collect user data and preferences through YouTube API or user interactions.
- Use collaborative filtering or content-based filtering techniques to recommend videos to users based on their preferences.
- Evaluate the recommendation system's performance using metrics like click-through rate and user engagement.

Step 7: YouTube Algorithm Analysis Explore the YouTube algorithm's impact on video trends and visibility. Analyze factors that contribute to videos trending on the platform

and understand how the algorithm selects and promotes videos. The steps for implementation are as follows:

- Gather data on video metadata and features that affect video rankings in YouTube's algorithm.
- Analyze the correlation between video attributes and trending status using statistical techniques.
- Visualize the insights to understand the key factors influencing video visibility and success.

Step 8: YouTube Trending Topics Identify trending topics or keywords that are popular among YouTube viewers. This analysis can help content creators align their video content with current trends and maximize audience reach. The steps for implementation are as follows:

- Use text analysis techniques to extract keywords and trending topics from video titles and descriptions.
- Identify common themes or topics among trending videos during specific periods.
- Visualize the trending topics to guide content creation strategies.

Step 9: Competitor Analysis Conduct a competitor analysis to understand how competing channels or content creators perform in terms of trending videos. Identify the strengths and weaknesses of competitors to develop a competitive strategy. The steps for implementation are as follows:

- Identify competitor channels or content creators in the same niche or category.
- Analyze the performance of competitor videos in terms of views, likes, and comments.
- Visualize the comparison to identify areas where your channel can improve.

Step 10: Machine Learning for Trend Prediction Develop a machine learning model to predict which videos are likely to trend in the future. This predictive model can be based on historical video data and various video attributes. The steps for implementation are as follows:

- Select relevant features and preprocess the data for machine learning.
- Split the dataset into training and testing sets.

- Use classification or regression algorithms to build the predictive model.
- Evaluate the model's performance using metrics like accuracy, precision, and recall.

Concept Explanation :

Random Forest: Unleashing the Power of Forest Wisdom!

Oh, hello there! Welcome to the enchanted forest of machine learning algorithms, where trees grow wild and free. Today, let me introduce you to the most mystical of them all - the Random Forest!

Imagine you're in a magical world filled with decision-making trees, each with its own unique wisdom. These trees love to work together, just like a group of friends having a grand adventure! That's what a Random Forest is - a magical coalition of decision trees!

Let's Unravel the Magic: Imagine you have a treasure trove of data about YouTube videos - their views, likes, comments, and more! Your quest is to find out which videos will trend and become legendary among viewers. But how can you predict the future, even in a mystical forest?

- Fear not, for the Random Forest is here to lend you its powers! It's like having a council of wise trees that will vote on the best decision for each video. These wise trees, called decision trees, have the power to ask questions and make choices.
- For example, a decision tree might ask, "Has the video received more than 1 million views?" If the answer is "yes," it might say, "Trend Alert!" But if the answer is "no," it might consult with its fellow trees and gather more insights.

Unity is Strength: Here's where the magic happens! The Random Forest gathers many such decision trees, each with its own knowledge and expertise. Together, they form an unbreakable alliance to make the ultimate prediction.

- Each decision tree casts its vote, and the majority's decision wins! It's like a democracy, but for predictions! So, if most trees say, "Trend Alert!" for a particular video, you can be pretty confident that the video will indeed become a trendsetter.
- A Shield Against Evil: But wait, there's more! The Random Forest is not just great at predicting trends; it's also a shield against evil overfitting monsters. These monsters can trick a single decision tree into making biased decisions, leading to inaccurate predictions.

- But fear not! When the Random Forest comes together, it forms a protective barrier against overfitting monsters. The alliance of decision trees keeps each other in check, ensuring that the predictions are not swayed by random noise.

How to Summon the Random Forest: To summon the Random Forest's power, you need the magical language of programming! In Python, you can use libraries like scikit-learn to bring the Random Forest to life. Just feed your data to the forest, and it will work its magic!

- And voilà! The Random Forest will bestow upon you the gift of predictions, revealing which YouTube videos are destined for greatness. It's like having a crystal ball, but way cooler!
- In a Nutshell: So, dear friend, remember the Random Forest is not just a single decision tree, but a magnificent collaboration of many trees, working together to unlock the secrets of trending videos. With its wisdom and unity, the Random Forest will guide you on your journey to YouTube fame!
- Now go forth, harness the power of the forest, and unveil the wonders of YouTube trending videos! May the Random Forest be with you!

Exercise Questions :

Exercise 1: Data Exploration

Question: What is the purpose of the "USvideos.csv" dataset in this project, and what information does it contain?

Answer: The "USvideos.csv" dataset contains information about YouTube videos that trended in the United States. It includes details such as video titles, channel titles, trending dates, views, likes, dislikes, and comment counts.

Exercise 2: Missing Data Handling

Question: How did the code handle missing data in the "description" column?

Answer: The code filled in missing values in the "description" column with an empty string, represented as "".

Exercise 3: Visualizing Trends

Question: Using the "USvideos.csv" dataset, plot a bar chart showing the number of videos that trended each year.

Answer: To plot the bar chart, we can use the "trending_date" column to extract the year, count the occurrences of each year, and then visualize the results using a bar chart.

Exercise 4: Distribution Analysis

Question: Create a histogram to show the distribution of views for the YouTube videos in the dataset.

Answer: We can use a histogram to visualize the distribution of views. The x-axis will represent the views range, and the y-axis will show the number of videos falling within each range.

Exercise 5: Correlation Analysis

Question: Is there any correlation between the number of likes and the number of views for the trending videos? If yes, explain the relationship.

Answer: To determine the correlation between likes and views, we can calculate the correlation coefficient using the "views" and "likes" columns. A positive correlation would indicate that as the number of views increases, the number of likes also tends to increase.

Exercise 6: Title Analysis

Question: Can you create a word cloud to visualize the most frequently occurring words in video titles?

Answer: Yes, we can use a word cloud to visualize the most common words in the video titles. The size of each word in the cloud corresponds to its frequency in the dataset.

Exercise 7: Video Publishing Analysis

Question: How does the number of videos published vary across different days of the week and hours of the day?

Answer: To analyze the publishing pattern, we can extract the day of the week and hour of the day from the "publish_time" column and then plot bar charts to show the number of videos published on different days and hours.

Exercise 8: Categorical Analysis

Question: Which YouTube category has the highest number of trending videos in the dataset?

Answer: To find the YouTube category with the highest number of trending videos, we can group the data by the "category_name" column and count the occurrences of each category.

Exercise 9: Video Removal Analysis

Question: What percentage of videos in the dataset have been removed or had errors?

Answer: To find the percentage of videos that were removed or had errors, we can count the occurrences of "video_error_or_removed" being True and divide it by the total number of videos in the dataset.

Exercise 10: Trending Video Prediction

Question: Can you build a machine learning model using the "views," "likes," and "comment_count" columns to predict whether a video will trend or not?

Answer: Yes, we can use the "views," "likes," and "comment_count" columns as features and create a target variable to indicate whether a video trended or not. We can then use a classification algorithm like Logistic Regression or Random Forest to build the model and make predictions based on the features.