Covid 19 cases

Introduction:

It can be very contagious and spreads quickly. Over one million people have died from COVID-19 in the United States. COVID-19 most often causes respiratory symptoms that can feel much like a cold, the flu, or pneumonia. COVID-19 may attack more than your lungs and respiratory system.

Project development steps:

This project contains many steps, methods, and processes. I will provide all the steps below:

Download the Dataset:

The first step of the process is to download the dataset in any format, such as CSV, TXT, Excel,

or a dataset. For this project.

link:https://www.kaggle.com/datasets/chakradharmattapalli/covid-19-cases

Import necessary libraries:

There are several libraries that are commonly used for natural language processing (NLP) and

building the covid_19 interface. Here are some necessary libraries that you may need to

import: NLTK (Natural Language Toolkit)

Numpy

Scikit-l

earn

Matplotlib pandas

1)Import and read dataset:

loads a CSV file named "covid.csv" from the directory "F:\anu" and stores it in a variable called "data". Then, it prints the first 10 rows of the dataset using the head() function. This helps to quickly get a visual representation of the data

1)**Error:**

This error occurs when there are issues with the data in the CSV file, such as invalid or missing values.

2)**ParserError:**

This error occurs when there are issues with the format of the CSV file, such as incorrect delimiters, missing values, or incorrect data types.

3**)ImportError:**

This error occurs when the pandas library is not installed or not properly

imported. ***2)Data Cleaning:***

a)**Missing value**

The first code block checks for missing values in the dataset using the isnull() function and sums up the number of missing values in each column using the sum() function.

b). **Duplicate data**

The second code block checks for duplicate rows in the dataset using the duplicated() function and sums up the number of duplicate rows using the sum() function.

c). **Drop unecessary columns**

The third code block drops the "day" and "month" columns from the dataset using the drop() function with the axis parameter set to 1 to indicate columns, and the inplace parameter set to True to modify the dataset in place. The dateRep column is then converted to a datetime object using the to_datetime() function, and set as the index of the dataset using the set_index() function.

Errors:

1) **SyntaxError: I**f there are syntax errors in the code, such as missing parentheses or quotes, you'll get a SyntaxError when running the code.

2) **NameError**: If pandas is not imported or not aliased as pd, you'll get a NameError when calling pd.read_csv() or other pandas functions.

3)Data Analysis:

1.**Count the total number of cases and deaths in the dataset.**

calculates the total number of COVID-19 cases and deaths in the given dataset using the sum() function of Pandas. This information is important as it provides an overview of the magnitude of the COVID-19 outbreak covered in the dataset.

2.**Calculate the percentage of cases and deaths by country.**
groups COVID-19 data by country, calculates the total cases and deaths for each country, calculates the percentage of cases and deaths for each country, and creates a Pandas DataFrame with the calculated percentages. The output can help in understanding the impact of COVID-19 in different countries and identifying regions that require more attention and resources to combat the pandemic.

3. **Find the country with the highest number of cases and deaths** finds the country with

the highest number of COVID-19 cases and deaths in the given dataset.

It first groups the data by country and calculates the total number of cases and deaths for each country. Then, it finds the country with the highest number of cases and deaths using the idxmax() function on the respective columns of the grouped data.

Finally, it prints the name of the country with the highest number of cases and deaths, along with the actual numbers.

This information is useful for understanding which countries are most affected by the pandemic and may require more attention and resources to combat it.

Errors:

KeyError: 'countriesAndTerritories'

This error occurs when you use an incorrect column name while grouping or selecting data. Doublecheck the column names in your dataset and make sure that they match the ones you are using in your code.

TypeError: unsupported operand type(s) for /: 'str' and 'int'

This error occurs when you try to perform a mathematical operation between a string and an integer. Check the data type of your variables and make sure that they are all numeric.

4.data visualization

1. **Find top five countries in terms of cases, store them in a new dataframe and Visualize them**

performs data visualization by creating a bar plot of the top 5 countries in terms of COVID-19

cases.

First, a new Pandas DataFrame is created called 'df' containing only the 'cases', 'deaths', and 'countriesAndTerritories' columns from the original data.

Then, the data is grouped by country and the total cases and deaths are calculated by calling the sum() function on the respective columns of the grouped data.

Next, the data is sorted in descending order by the 'cases' column and the top 5 countries are selected using the head() function and stored in the 'top_5_countries' variable.
The index of the 'top_5_countries' DataFrame is then reset to create a new column with the index values.

Finally, a bar plot is created using the Seaborn library's barplot() function with the 'countriesAndTerritories' column on the x-axis and the 'cases' column on the y-axis, and the 'top_5_countries' DataFrame as the data source.

This visualization helps to quickly identify the countries that have been most affected by COVID-19 and allows for easy comparison of the number of cases between these countries.

2. **Find top five countries in terms of deaths, store them in a new dataframe and Visualize them**

The code first creates a new DataFrame called deaths_df that contains the columns 'deaths', 'cases', and 'countriesAndTerritories'. It then groups the data by country using the groupby() function and calculates the total deaths and cases for each country using the sum() function.

The next step is to sort the data by deaths in descending order using the sort_values() function and selecting the top 5 countries using the slice notation [:5]. The resulting DataFrame is then reset to have a new index using the reset_index() function.

Finally, a bar plot is created using the sns.barplot() function from the Seaborn library to visualize the top 5 countries with the highest number of COVID-19 deaths. The plot is customized with a title, x-label, ylabel, and figure size using various functions from the matplotlib library.

Errors:

1)ValueError:

This error could occur if the values being operated on are not valid. For example, if there are missing values or NaN values in the dataset. To fix this, check the dataset for any missing values or NaN values and remove or replace them accordingly.

2)ModuleNotFoundError:

This error could occur if the required modules such as pandas or numpy are not installed. To fix this, make sure to install the necessary modules using pip install command.

Conclusion:

The coronavirus disease continues to spread across the world following a trajectory that is difficult to predict. The health, humanitarian and socio-economic policies adopted by countries will determine the speed and strength of the recovery.