### **Step 3: Load and Prepare the Dataset**

#### **1. Locate the Dataset**

First, you need a dataset. We'll assume you have a dataset file named spam.csv. This file should be in your working directory or you should provide the correct path to it.

For this example, the dataset is expected to have columns like v1 for labels (spam/ham) and v2 for the email messages.

#### **2. Load the Dataset**

We will use the pandas library to load the CSV file into a DataFrame. This makes it easier to manipulate and analyze the data.

python

Copy code

import pandas as pd

# Load the dataset

data = pd.read\_csv('spam.csv', encoding='latin-1')

#### **3. Select and Rename Columns**

Since we are only interested in the v1 (label) and v2 (message) columns, we select these columns and rename them to label and message for clarity.

python

Copy code

# Select and rename columns

data = data[['v1', 'v2']]

data.columns = ['label', 'message']

#### **4. Encode Labels**

We convert the categorical labels (spam/ham) to numerical labels (1 for spam, 0 for ham). This is necessary because machine learning algorithms require numerical input.

python

Copy code

# Encode labels (spam=1, ham=0)

data['label'] = data['label'].map({'ham': 0, 'spam': 1})

### **Where to Run This Code**

You can run this code in a Python environment such as:

* **Jupyter Notebook**: An interactive environment for running Python code, useful for data analysis and visualization.
* **VS Code**: A popular code editor that supports Python development.
* **Python Script**: A standalone .py file that you can run from the command line or terminal.

#### **Example in Jupyter Notebook**

1. Open Jupyter Notebook.
2. Create a new notebook.
3. Copy and paste the following code into a cell and run it.

python

Copy code

import pandas as pd

# Load the dataset

data = pd.read\_csv('spam.csv', encoding='latin-1')

# Select and rename columns

data = data[['v1', 'v2']]

data.columns = ['label', 'message']

# Encode labels (spam=1, ham=0)

data['label'] = data['label'].map({'ham': 0, 'spam': 1})

# Display the first few rows of the dataframe to ensure everything is loaded correctly

data.head()

#### **Example in VS Code**

1. Open VS Code.
2. Create a new file named spam\_detection.py.
3. Copy and paste the following code into spam\_detection.py.

python

Copy code

import pandas as pd

# Load the dataset

data = pd.read\_csv('spam.csv', encoding='latin-1')

# Select and rename columns

data = data[['v1', 'v2']]

data.columns = ['label', 'message']

# Encode labels (spam=1, ham=0)

data['label'] = data['label'].map({'ham': 0, 'spam': 1})

# Display the first few rows of the dataframe to ensure everything is loaded correctly

print(data.head())

1. Open the terminal in VS Code.
2. Run the script using the command:

bash

Copy code

python spam\_detection.py

### **Detailed Explanation of the Code**

1. **Importing pandas**:
   * import pandas as pd
   * This imports the pandas library and aliases it as pd for ease of use.
2. **Loading the dataset**:
   * data = pd.read\_csv('spam.csv', encoding='latin-1')
   * This reads the spam.csv file into a DataFrame named data.
   * encoding='latin-1' is specified to handle any special characters in the dataset.
3. **Selecting and renaming columns**:
   * data = data[['v1', 'v2']]
   * This selects only the v1 and v2 columns from the DataFrame.
   * data.columns = ['label', 'message']
   * This renames the selected columns to label and message for better readability.
4. **Encoding labels**:
   * data['label'] = data['label'].map({'ham': 0, 'spam': 1})
   * This maps the ham label to 0 and the spam label to 1, converting the categorical labels to numerical labels.
5. **Displaying the first few rows**:
   * data.head()
   * This displays the first few rows of the DataFrame to verify that the data has been loaded and processed correctly.

By following these steps, you'll be able to load and prepare the dataset for your spam email detection project