Question17:

Scenario: You are a data analyst working for a marketing research company. Your team hascollected a large dataset containing customer feedback from various social media platforms. Thedataset consists of thousands of text entries, and your task is to develop a Python program to analyze the frequency distribution of words in this dataset. Your program should be able to perform

the following tasks:

Load the dataset from a CSV file (data.csv) containing a single column named "feedback"

with each row representing a customer comment.

 Preprocess the text data by removing punctuation, converting all text to lowercase, and

eliminating any stop words (common words like "the," "and," "is," etc. that don't carry

significant meaning).

 Calculate the frequency distribution of words in the preprocessed dataset.

 Display the top N most frequent words and their corresponding frequencies, where N is provided as user input.

 Plot a bar graph to visualize the top N most frequent words and their frequencies.

Question: Create a Python program that fulfills these requirements and helps your team gain insights from the customer feedback data.

Answer:

import pandas as pd

import string

import matplotlib.pyplot as plt

from collections import Counter

import re

import os

stop\_words = {

    'the', 'is', 'and', 'in', 'to', 'with', 'a', 'for', 'of', 'on', 'it', 'this',

    'that', 'i', 'was', 'but', 'be', 'have', 'not', 'are', 'as', 'very', 'so', 'from'

}

file\_path = os.path.join(r"C:\Users\jampa\Downloads\data.csv")

def load\_data(filepath):

    try:

        df = pd.read\_csv(filepath)

        return df['feedback'].dropna().astype(str)

    except Exception as e:

        print("Error loading file:", e)

        return pd.Series()

def preprocess(texts):

    words = []

    for text in texts:

        text = text.lower()

        text = re.sub(f"[{string.punctuation}]", "", text)

        tokens = text.split()

        tokens = [word for word in tokens if word not in stop\_words]

        words.extend(tokens)

    return words

def plot\_words(word\_freq, n):

    common = word\_freq.most\_common(n)

    words, counts = zip(\*common)

    plt.figure(figsize=(8,3 ))

    plt.bar(words, counts, color='orange')

    plt.title(f"Top {n} Most Frequent Words")

    plt.xlabel("Words")

    plt.ylabel("Frequency")

    plt.xticks(rotation=45)

    plt.tight\_layout()

    plt.show()

def main():

    feedback\_data = load\_data(file\_path)

    if feedback\_data.empty:

        print("No feedback data found.")

        return

    words = preprocess(feedback\_data)

    word\_freq = Counter(words)

    try:

        n = int(input("Enter the number of top frequent words to display: "))

    except ValueError:

        print("Please enter a valid number.")

        return

    print(f"\nTop {n} Most Frequent Words:")

    for word, count in word\_freq.most\_common(n):

        print(f"{word}: {count}")

    plot\_words(word\_freq, n)

if \_\_name\_\_ == "\_\_main\_\_":

    main()

Output

