

EXPERIMENT-17

17. Scenario: You are a data analyst working for a marketing research company. Your team has collected a large dataset containing customer feedback from various social media platforms. The dataset 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.

Code:

```
import pandas as pd
import string
from collections import Counter
import matplotlib.pyplot as plt
stop_words = {
    "the", "is", "and", "a", "an", "in", "on", "of", "to", "for",
    "with", "was", "were", "it", "this", "that", "as", "be"
}
df = pd.read_csv("data.csv") # Make sure data.csv is in same folder
text = ".join(df["feedback"]).lower()
clean_text = text.translate(str.maketrans("", "", string.punctuation))
words = clean_text.split()
filtered_words = [w for w in words if w not in stop_words]
freq = Counter(filtered_words)
N = int(input("Enter N: "))
top_words = freq.most_common(N)
print("Top", N, "Most Frequent Words:")
print(top_words)
labels, values = zip(*top_words)
plt.figure(figsize=(8,5))
plt.bar(labels, values)
plt.title("Top N Most Frequent Words")
plt.xlabel("Words")
```

```
plt.ylabel("Frequency")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

Output:

