Name : Vedant Maruti Kadam

PRN : 202401050003 Roll No : CC-11

Theory Activity No : 1

Topic : Kaggle Text Classification Dataset

Problem Statements and Solutions

```
import numpy as np
import pandas as pd

Load the dataset
df = pd.read_csv('text_classification_dataset.csv')
```

Problem 1: What is the shape of the dataset?

```
print("Problem 1: Shape of the dataset")
print(df.shape)
```

Problem 2: What are the column names in the dataset?

```
print("\nProblem 2: Column names in the dataset")
print(df.columns)

Problem 3: What is the distribution of labels in the dataset?
print("\nProblem 3: Distribution of labels in the dataset")
```

```
print(df['label'].value_counts())
```

Problem 4: How many unique labels are there in the dataset?

```
print("\nProblem 4: Number of unique labels in the dataset")
print(len(df['label'].unique()))
```

Problem 5: What is the length of the longest text in the dataset?

```
print("\nProblem 5: Length of the longest text in the dataset")
print(df['text'].apply(len).max())
```

Problem 6: What is the average length of texts in the dataset?

```
print("\nProblem 6: Average length of texts in the dataset")
print(df['text'].apply(len).mean())
```

Problem 7: How many missing values are there in the dataset?

```
print("\nProblem 7: Number of missing values in the dataset")
print(df.isnull().sum())
```

Problem 8: What is the most common word in the dataset?

```
print("\nProblem 8: Most common word in the dataset")
```

```
from collections import Counter
words = ' '.join(df['text']).split()
print(Counter(words).most_common(1))
```

Problem 9: What is the frequency of the top 10 most common words in the dataset?

```
print("\nProblem 9: Frequency of top 10 most common words in the dataset")
print(Counter(words).most_common(10))
```

Problem 10: How many texts contain the word "machine"?

```
print("\nProblem 10: Number of texts containing the word 'machine'")
print(df['text'].apply(lambda x: 'machine' in x.lower()).sum())
```

Problem 11: What is the correlation between the length of texts and their labels?

Not directly applicable for text data

Problem 12: What is the distribution of text lengths in the dataset?

```
print("\nProblem 12: Distribution of text lengths in the dataset")
print(df['text'].apply(len).describe())
```

Problem 13: How many texts have a length greater than 100 words?

```
print("\nProblem 13: Number of texts with length greater than 100 words")
print((df['text'].apply(len) > 100).sum())
```

Problem 14: What is the average length of texts for each label?

```
print("\nProblem 14: Average length of texts for each label")
print(df.groupby('label')['text'].apply(lambda x: x.apply(len).mean()))
```

Problem 15: What are the top 5 most common labels in the dataset?

```
print("\nProblem 15: Top 5 most common labels in the dataset")
print(df['label'].value_counts().head(5))
```

Problem 16: How many texts are labeled as "spam"?

```
print("\nProblem 16: Number of texts labeled as 'spam'")
print((df['label'] == 'spam').sum())
```

Problem 17: What is the frequency of each label in the dataset?

```
print("\nProblem 17: Frequency of each label in the dataset")
print(df['label'].value_counts())
```

Problem 18: What is the standard deviation of text lengths in the dataset?

print("\nProblem 18: Standard deviation of text lengths in the dataset")

print(df['text'].apply(len).std())

Problem 19: How many unique words are there in the dataset?

print("\nProblem 19: Number of unique words in the dataset")

print(len(set(words)))

Problem 20: What is the distribution of word frequencies in the dataset?

print("\nProblem 20: Distribution of word frequencies in the dataset")

print(Counter(words).most_common())

These problem statements cover a range of topics, including:

- Dataset shape and structure
- Label distribution and frequency
- Text length and word frequency analysis
- Data quality and missing values

- Label-specific analysis

The solutions use various Numpy and Pandas methods, including:

- df.shape and df.columns for dataset structure
- value_counts() for label frequency
- apply() and lambda functions for text length and word frequency analysis
- isnull().sum() for missing values
- groupby() and mean() for label-specific analysis