ESSENTIAL OF DATA SCIENCE

Theory Activity No. 1

Name – Kedar Sanjay jasud

Div – CS8

Roll No. – CS8-15

PRN – 202401120001

- 20 problem statements for Kaggel Text Classification Dataset using Numpy and Pandas.
- Kaggle Link - https://www.kaggle.com/datasets/pratspy/opinrank-dataset-processed

10 Problem Statements Using NumPy:

1.Find the maximum review length
2. Find the minimum review length
3. Find the average review length
4. Find the standard deviation of review lengths
5. Count reviews longer than 100 characters
6. Count reviews shorter than 50 characters
7. Find the median review length
8. Sort all review lengths

9. Count reviews with more than 150 characters
10. Print the length of every 1000th review

- **Solution:-**

```
# 1
np.max(df['Review'].str.len().to_numpy())

# 2
np.min(df['Review'].str.len().to_numpy())

# 3
np.mean(df['Review'].str.len().to_numpy())

# 4
np.std(df['Review'].str.len().to_numpy())

# 5
np.sum(df['Review'].str.len().to_numpy() > 100)

# 6
np.sum(df['Review'].str.len().to_numpy() < 50)

# 7
np.median(df['Review'].str.len().to_numpy())

# 8
np.sort(df['Review'].str.len().to_numpy())

# 9
np.sum(df['Review'].str.len().to_numpy() > 150)

# 10
df['Review'].str.len().to_numpy()[::1000]
```

```
Max length of review: 10512
Min length of review: 1
Mean length of review: 542.0892666083066
Standard deviation of review lengths: 466.74590372422733
Number of reviews longer than 100 characters: 110379
Number of reviews shorter than 50 characters: 10745
Median length of review: 438.0
Sorted lengths of reviews: [    1     1     1 ...  8821 10432 10512]
Number of reviews longer than 150 characters: 106334
Every 1000th review length: [1145  789  647  746  179  507  493  998 1410 1494  626  615  764  351
 1073  320  506  309 1073  779  541  505 1031  808  989  973  414  316
  238  479 1076  396  265  488    8  328  362  629  615    8  507    6
 1176  507  319  537  448  538  141  378  399  406  263  342  383  413
  804  910  252 1564  616  708  265  398 1555  927  255 1101  643   79
  883  425  785   60  993   19  521  673  424  470  454  535  482   95
  872  383  352 2054  367  594  900    6  385  308  603    8  239  458
  625  592  437  163  923  335  843  716  371  204  426  287  550   20
  387   22  132  364 1184  335  690  293   26  125  285  473]
PS C:\Users\Admin\.vscode\extensions\ms-vscode.cpptools-1.22.11-win32-x64\ui\New folder (2)>
```

10 Problem Statements Using Pandas:

1.Count positive and negative reviews
2. Find the number of missing reviews
3. Add a column for review length
4. Add a column for word count
5. Show top 5 longest reviews
6. Show top 5 shortest reviews
7. Find reviews that contain the word "clean"
8. Count how many reviews contain the word "good"
9. Count how many duplicate reviews exist
10. Find average word count grouped by Sentiment

- **Solution:-**

```
# 1
df['Sentiment'].value_counts()

# 2
df['Review'].isnull().sum()

# 3
df['Length'] = df['Review'].apply(len)

# 4
df['Words'] = df['Review'].apply(lambda x: len(x.split()))

# 5
df.sort_values(by='Length', ascending=False).head()

# 6
df.sort_values(by='Length').head()

# 7
df[df['Review'].str.contains('clean', case=False)]

# 8
df['Review'].str.contains('good', case=False).sum()

# 9
df['Review'].duplicated().sum()

# 10
df.groupby('Sentiment')['Words'].mean()
```

```
Sentiment
POSITIVE    62269
NEGATIVE    61103
Name: count, dtype: int64
Number of missing reviews: 0
Top 5 longest reviews:
                                                   Review Sentiment  Length
46441  grander pricewe holiday virgin include flight ...  NEGATIVE   10512
60620  epic fail many levels amazing ok go struggle d...  NEGATIVE   10432
18967  heinous recently misfortune spend royal carbig...  NEGATIVE    8821
60856  great lot patience tolerate incompetence fact ...  NEGATIVE    7756
99062  puli five year making finally urban resort nes...  POSITIVE    7615
Top 5 shortest reviews:
       Review Sentiment  Length
89412       j  POSITIVE       1
91582       u  POSITIVE       1
77229       e  POSITIVE       1
99654       z  POSITIVE       1
107892      u  POSITIVE       1
Reviews containing 'clean':
                                                   Review Sentiment  Length
0       stylish clean reasonable value poor glad first...  NEGATIVE    1145
1       clean good poor service check friend arrive di...  NEGATIVE     823
12      not great part tour bed hard could sleep floor...  NEGATIVE     183
13      ok ok use bit updating somewhat clean breakfas...  NEGATIVE     479
17      didnt like several good thing central not expe...  NEGATIVE     340
...                                                   ...       ...     ...
123366  excellent love time could not believe quality ...  POSITIVE     450
123367  still work kink stay dream business still work...  POSITIVE     769
123368  need need stay somewhere central one still not...  POSITIVE     673
123369  conveniently locate nice stay spend westbury m...  POSITIVE    1142
```