ESSENTIAL OF DATA SCIENCE

Theory Activity No. 1

Name – Himani Bhurkunde

Div – CS8

Roll No. – CS8-55

PRN – 202401120076

- 20 problem statements for Kaggel Text Classification Dataset using Numpy and Pandas.
- Kaggle Link - https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

10 Problem Statements Using NumPy:

1. Calculate the standard deviation of the review lengths.
2. Count how many reviews have lengths greater than the average review length.
3. Compute the 25th, 50th, and 75th percentiles of review lengths.
4. Normalize the review lengths using z-score normalization.
5. Determine the average word count of reviews longer than 1000 characters.
6. Count how many reviews have a word count between 100 and 200.

7. Find the minimum and maximum word count in the dataset.
8. Identify which reviews are shorter than 100 characters using a boolean mask.
9. Calculate the mean and variance of the word counts.
10. Count how many reviews fall within one standard deviation from the mean review length.

- **Solution:-**

```
std_review_length = np.std(review_lengths_np)
above_avg_count = np.sum(review_lengths_np >
np.mean(review_lengths_np))
percentiles = np.percentile(review_lengths_np, [25, 50, 75])
normalized_lengths = (review_lengths_np -
review_lengths_np.mean()) / review_lengths_np.std()
avg_word_long_reviews =
np.mean(word_counts_np[review_lengths_np > 1000])
mid_word_range = np.sum((word_counts_np >= 100) &
(word_counts_np <= 200))
min_word_count = np.min(word_counts_np)
max_word_count = np.max(word_counts_np)
short_mask = review_lengths_np < 100
mean_word_count = np.mean(word_counts_np)
variance_word_count = np.var(word_counts_np)
within_1_std = np.sum(
   (review_lengths_np > np.mean(review_lengths_np) -
np.std(review_lengths_np)) &
   (review_lengths_np < np.mean(review_lengths_np) +
np.std(review_lengths_np))
)
```

```
Standard Deviation of Review Lengths: 989.7181170827175
Reviews Longer Than Average: 16620
Review Length Percentiles (25/50/75): [ 699.    970.   1590.25]
Normalized Review Lengths (first 5): [ 0.45626019 -0.31466638 -0.38741437 -0.56726356  0.00764761]
Average Word Count in Long Reviews (>1000 chars): 348.92504478980044
Reviews with Word Count Between 100 and 200: 23553
Min and Max Word Count: 4 / 2470
Short Review Mask (first 5): [False False False False False]
Mean & Variance of Word Count: 231.15694 / 29358.1782298364
Reviews Within 1 Std Dev of Mean Length: 41618
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 the number of reviews for each sentiment.
 # 2. Determine the average character length of reviews.
 # 3. Find the longest review in the dataset.
 # 4. Identify the shortest review.
 # 5. Count how many reviews contain the word 'excellent'.
 # 6. Count how many reviews have more than 1000 characters.
 # 7. Compute the average review length grouped by sentiment.
 # 8. Determine how many reviews mention the word 'bad'.
 # 9. Find the review with the maximum word count.
 # 10. Display the first 5 reviews containing the word 'boring'.

- **Solution:-**

```
sentiment_counts = df['sentiment'].value_counts()
avg_review_length = df['review_length'].mean()
longest_review = df.loc[df['review_length'].idxmax(), 'review']
shortest_review = df.loc[df['review_length'].idxmin(), 'review']
excellent_reviews = df[df['review'].str.contains("excellent",
case=False, na=False)].shape[0]
long_reviews_count = df[df['review_length'] > 1000].shape[0]
avg_length_by_sentiment =
df.groupby('sentiment')['review_length'].mean()
bad_mentions = df['review'].str.contains("bad", case=False,
```

```
na=False).sum()
max_word_review = df.loc[df['word_count'].idxmax(), 'review']
boring_reviews = df[df['review'].str.contains("boring",
case=False, na=False)].head(5)
```

```
Sentiment Counts:
 sentiment
positive    25000
negative    25000
Name: count, dtype: int64

Average Review Length: 1309.43 characters

Longest Review (first 300 chars):
 Match 1: Tag Team Table Match Bubba Ray and Spike Dudley vs Eddie Guerrero and Chris Benoit Bubba Ray and Spike Dudley started things off with a Tag Te
am Table Match against Eddie Guerrero and Chris Benoit. According to the rules of the match, both opponents have to go through tables in order to get

Shortest Review:
 Read the book, forget the movie!

Number of Reviews with 'Excellent': 3625

Number of Reviews > 1000 characters: 24001

Average Review Length by Sentiment:
 sentiment
negative    1294.06436
positive    1324.79768
Name: review_length, dtype: float64
```