ASSIGNMENT

Task 1: Exploratory Data Analysis and Data Visualization

Maximum Score: 05

Download the dataset from https://www.kaggle.com/rohitsahoo/sales-forecasting and perform exploratory data analysis.

1. Exploratory Data Analysis:
    a. Explore the dataset to understand its structure, variables, and distributions
    b. Identify missing values and outliers, and decide how to handle them.
2. Data Visualization:
    a. Create visualizations to show the trends and patterns in various features.
    b. Use appropriate visualization techniques (e.g., line chart, scatter plot, bar chart) to present your findings.
    c. Analyze the visualizations to identify trends and insights in the data.

Task 2: Data Analysis :Natural Language Processing

Maximum Score: 05

Download the dataset from: https://www.kaggle.com/andrewmvd/trip-advisor-hotel-reviews and perform preliminary analysis using common Natural Language Processing techniques.

Your analysis should include:

1. Pre-processing the text data to remove noise and irrelevant information.
2. Performing basic statistical analysis such as word frequency counts, n-grams, and sentiment analysis.
3. Visualizing the text data using techniques such as word clouds, scatter plots, and heatmaps.
4. Applying advanced NLP techniques such as topic modeling, text classification, or named entity recognition to gain deeper insights into the data.

You should use Python and relevant NLP libraries such as NLTK, spaCy, or gensim to perform your analysis. Finally, you should write a report that summarizes your findings and discusses any interesting patterns or insights you discovered during your analysis.

Task 3: Flask Application Development

Maximum Score: 10

Your task is to design a Flask application that can retrieve named entities from a news article.

Description:

1. The application should allow users to submit a URL of a news article, extract the text from the article, and display a list of named entities (e.g., people, organizations, locations) that appear in the text.
2. You should use Natural Language Processing (NLP) techniques to perform named entity recognition, and the extracted entities should be displayed in a user-friendly format.
3. You can train your own model or compare state of the art named entity recognition models(at least 3) and justify your choice of model.
4. Training code/ Comparision code should be shared in a separate notebook.

Notes:

1. All the codes/models/notebooks should be uploaded on google drive and shared with pallavi@pinacalabs.com and humanresource@pinacalabs.com
2. A separate folder should be created for each task and named accordingly .
3. All codes should be well documented and clear.