NEW HORIZON
COLLEGE OF ENGINEERING

Data Viz: Natural Language Processing (NLP) and Sentiment Analysis using Machine Learning

Date: 30-07-2024 **Time:** 9:30 AM to 11:00 AM

On July 30, 2024, the Alumni Association and the Department of Computer Science and Engineering (Data Science) at New Horizon College of Engineering organized a workshop titled " Data Viz: Natural Language Processing (NLP) and Sentiment Analysis using Machine Learning" This report provides a detailed overview of the workshop on Natural Language Processing (NLP) and Exploratory Data Analysis (EDA), including the types and techniques of EDA, and how sentiment analysis is performed using Machine Learning (ML). The workshop aimed to impart foundational knowledge and practical insights into NLP, the importance of EDA, and the methodologies for conducting sentiment analysis.

Session Activities

The workshop session comprised both theoretical explanations and practical coding exercises, as seen in the provided image. The instructor used a whiteboard for conceptual explanations and a coding environment (likely VS Code) displayed on a screen for hands-on demonstrations.

Key Activities:

1. **Introduction to Regular Expressions:**
   - The instructor explained the use of regular expressions in text processing. Regular expressions are essential for pattern matching in strings, which is a fundamental step in NLP tasks.
2. **Importing Libraries:**
   - Participants learned how to import essential Python libraries for NLP and sentiment analysis. These included:
     - re for regular expressions.
     - string for string operations.
     - stopwords from nltk.corpus for removing common words that do not contribute to the sentiment.
     - CountVectorizer from sklearn.feature_extraction.text for converting text to a matrix of token counts.
3. **Data Preprocessing:**
   - The session covered loading and preprocessing text data. This involved cleaning the data by removing unnecessary characters, lowercasing text, and tokenizing sentences.

4. **Feature Extraction:**
   - The instructor demonstrated how to use CountVectorizer to convert text data into a format suitable for ML models, emphasizing the importance of transforming text into numerical features.
5. **Sentiment Analysis Model:**
   - Building and evaluating a sentiment analysis model using machine learning techniques. The practical exercise involved:
     - Splitting the data into training and testing sets.
     - Training a model (e.g., Naive Bayes) on the training data.
     - Evaluating the model's performance on the test data using accuracy, precision, recall, and F1 score metrics.

Natural Language Processing (NLP)

Overview: Natural Language Processing (NLP) is a branch of Artificial Intelligence that focuses on the interaction between computers and human languages. It aims to enable computers to understand, interpret, and generate human language in a valuable and meaningful way.

Key Components:

1. **Tokenization:** Splitting text into individual words or phrases.
2. **Stop Word Removal:** Eliminating common words that do not contribute much meaning (e.g., "and", "the").
3. **Stemming and Lemmatization:** Reducing words to their base or root form.
4. **Part of Speech Tagging (POS):** Identifying the grammatical category of words.
5. **Named Entity Recognition (NER):** Detecting and classifying proper names in text (e.g., names of people, organizations, locations).

Applications:

- **Machine Translation:** Automatically translating text from one language to another.
- **Chatbots and Virtual Assistants:** Automating customer service and providing real-time assistance.
- **Text Summarization:** Condensing long articles into shorter summaries.
- **Sentiment Analysis:** Determining the emotional tone behind a body of text.

Exploratory Data Analysis (EDA)

Definition: Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. It is a critical step in the data preprocessing phase.

Types of EDA:

1. **Univariate Analysis:** Examining a single variable to summarize and find patterns.
2. **Bivariate Analysis:** Analyzing the relationship between two variables.
3. **Multivariate Analysis:** Examining more than two variables to understand complex relationships.

Techniques:

- **Descriptive Statistics:** Using mean, median, mode, variance, and standard deviation to describe data.
- **Data Visualization:** Employing plots like histograms, box plots, scatter plots, and bar charts.
- **Data Transformation:** Normalizing or standardizing data to bring different scales into a common range.
- **Outlier Detection:** Identifying and handling anomalies that deviate significantly from other observations.

Sentiment Analysis using Machine Learning

Overview: Sentiment analysis, also known as opinion mining, involves determining the sentiment expressed in a piece of text. This can be positive, negative, or neutral. Machine Learning (ML) models can be trained to perform sentiment analysis by learning from annotated data.

Steps in Sentiment Analysis:

1. **Data Collection:** Gathering text data from sources like social media, reviews, and surveys.
2. **Data Preprocessing:** Cleaning text data by removing noise, stop words, and performing tokenization and lemmatization.
3. **Feature Extraction:** Converting text into numerical features using techniques like Bag of Words (BoW), TF-IDF, or word embeddings.
4. **Model Training:** Using ML algorithms such as Naive Bayes, Logistic Regression, or more advanced models like LSTM (Long Short-Term Memory) and BERT (Bidirectional Encoder Representations from Transformers) to train on the labeled data.
5. **Model Evaluation:** Assessing the performance of the model using metrics such as accuracy, precision, recall, and F1 score.
6. **Prediction and Analysis:** Applying the trained model to new text data to predict sentiment.

Challenges:

- **Ambiguity and Context:** Understanding the context and disambiguating words with multiple meanings.
- **Sarcasm Detection:** Identifying sarcastic comments which can alter the sentiment.
- **Language and Domain Variability:** Handling variations in language and specific jargon used in different domains.

The workshop provided valuable insights into the field of NLP, emphasizing its importance in modern AI applications. Understanding and performing EDA is crucial for preparing data and uncovering insights that inform model development. Sentiment analysis is a powerful tool for interpreting the emotional tone of text data, with broad applications in business and research. By leveraging NLP techniques and ML models, we can automate and enhance the process of sentiment analysis, leading to more informed decision-making and improved user experiences.

Faculty Coordinator **HOD-CSE(DS)**

Glimpses of the event: