NEW HORIZON
COLLEGE OF ENGINEERING
AUTONOMOUS COLLEGE Permanently Affiliated to VTU, Approved by AICTE & UGC
Accredited by NAAC with 'A' Grade

Data Viz: NLP Foundations-Workshop

Date: 22-07-2024 **Time:** 9:30 AM to 11:00 AM

On July 22, 2024, the Alumni Association and the Department of Computer Science and Engineering (Data Science) at New Horizon College of Engineering had organized a workshop titled "Data Viz: NLP Foundations." This workshop was held from 09:30 AM to 11:00 AM at the Data Science Lab-1. The session was conducted by resource person from PreProd Corp. This report summarizes the activities and learnings from the workshop.

The workshop aimed to provide overview on natural language processing (NLP) foundations. Natural Language Processing (NLP) is a critical field of artificial intelligence that focuses on the interaction between computers and human languages. The NLP Foundations session provided an in-depth understanding of essential concepts and techniques in NLP. This report documents the key points discussed during the session.

Key Topics Covered

NLP formulation is categorized into two primary approaches:

1. Statistical NLP

Statistical NLP leverages statistical methods to analyse and interpret large datasets of text. The session covered several important components of Statistical NLP:

a. Bag of Words (BOW)

- **Definition:** A method for representing text data by treating each word independently and disregarding grammar and word order.
- **Usage:** It helps in creating word clouds and analysing the frequency of words within a text.

b. TF-IDF Vectorizer

- **Term Frequency (TF):** Measures the frequency of a word in a document.
- **Inverse Document Frequency (IDF):** Evaluates the importance of a word across multiple documents.
- **TF-IDF Vectorizer:** Combines TF and IDF to provide a weighted score, highlighting significant words in a document while filtering out common, less informative terms.

2. Sequential NLP

Sequential NLP focuses on the order and context of words to derive meaning from text. This approach is crucial for applications like language translation and text generation.

a. People-centric Approaches

- **Focus:** Personalized and context-based text processing that takes into account the sequence and overall context of words.

Document-Term Matrix (DTM)

The Document-Term Matrix (DTM) is a fundamental concept in NLP that represents the frequency of terms in a collection of documents. Each row corresponds to a document, and each column corresponds to a term. This matrix is essential for various text analysis techniques, including topic modelling, sentiment analysis, and information retrieval.

Exploratory Data Analysis (EDA) Methods

Exploratory Data Analysis (EDA) is a crucial step in understanding and preprocessing text data. The session highlighted several key EDA methods:

a. Stemming

- **Definition:** Reduces words to their root form by removing suffixes.
- **Example:** "Running," "runner," and "ran" are all reduced to "run."

b. Lemmatization

- **Definition:** Reduces words to their base or dictionary form, considering the context.
- **Example:** "Better" is reduced to "good," reflecting the word's meaning.

c. Case Conversion

- **Process:** Converting all characters to lowercase to standardize text data and improve algorithm efficiency.

Sparse Matrix Representation

In NLP, text data is often represented as a sparse matrix, where most elements are zero due to the infrequent occurrence of many words. Sparse matrices are efficient for storing and processing large text datasets, as they save memory by only storing non-zero elements.

The NLP Foundations session provided participants with a robust understanding of the core concepts and methods used in NLP. By covering both Statistical and Sequential NLP approaches, as well as foundational techniques like Bag of Words, TF-IDF, and DTM, attendees are well-equipped to apply these methods to real-world text analysis tasks. The session emphasized the importance of systematic text analysis, including data cleaning and preprocessing, to extract meaningful insights from text data.

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

Glimpses of the event: