Applied Data Engineering for Generative AI

Duration: 3-Days Hands-On Workshop

Course Modules

Day 1: Preprocessing & Data Cleaning

Morning Session (9:00 AM - 12:00 PM)

Module 1. Introduction to Data Cleaning

a. Importance of data cleaning

b. Types of data issues: missing values, duplicates, inconsistencies

Module 2. Hands-On: Data Cleaning Basics

a. Identifying and handling missing values

b. Removing duplicates

c. Correcting data inconsistencies

Module 3. Advanced Data Cleaning Techniques

a. Outlier detection and treatment

b. Data normalization and standardization

Module 4. Hands-On: Applying Advanced Techniques

a. Practice with real-world datasets

b. Group activities and discussions

Afternoon Session (1:00 PM - 5:00 PM)

Module 6. Introduction to Regular expression (Regex)

a. Examples of use cases

Module 7. Basic Regex Patterns

a. Character Classes and Sets

b. Quantifiers

c. Anchors and Boundaries

d. Special Characters and Escaping

Corporate Office
NQC Technology Sdn Bhd (604535-M) Unit 27-6 Block F2, Jalan PJU 1/42A,
Dataran Prima, 47301 Petaling Jaya, Selangor Darul Ehsan
Tel: +603-7805 2088 | Fax: +603.7803 6948
Email: ahi@nqc.com.my | www.nqc.com.my

Module 8. Intermediate Regex Patterns

a. Groups and Capturing

Module 9. Hands-On: Applying Regex in python

a. Practice with real-world datasets

b. Group activities and discussions

Day 2: Data Engineering

Morning Session (9:00 AM - 12:00 PM)

Module 10. Introduction to Feature Engineering

a. Definition and importance

b. Types of features: numerical, categorical, datetime

Module 11. Creating New Features

a. Feature extraction

b. Feature transformation

c. Interaction features

Module 12. Advanced Feature Engineering Techniques

a. Encoding categorical variables

b. Binning and discretization

c. Polynomial features

Module 13. Hands-On: Advanced Techniques

a. Real-world dataset exercises

b. Group discussions and feedback

Afternoon Session (1:00 PM - 5:00 PM)

Module 14. Overview on NLP

Module 15. Text Preprocessing

a. Tokenization

b. Stop words removal

Corporate Office
NQC Technology Sdn Bhd (604535-M) Unit 27-6 Block F2, Jalan PJU 1/42A,
Dataran Prima, 47301 Petaling Jaya, Selangor Darul Ehsan
Tel: +603-7805 2088 | Fax: +603.7803 6948
Email: ahi@nqc.com.my | www.nqc.com.my

c. Stemming and Lemmatization

Module 16. Text Feature engineering

a. Bag of Words (BoW)

b. Term Frequency-Inverse Document Frequency (TF-IDF)

Module 17. Advanced NLP

a. Sparse word embeddings (Word2Vec etc.)

b. Dense word embeddings (Bert etc.)

c. Semantic similarity

Day 3: Generative AI, Vector DB & Retrieval Augmented Generation

Morning Session (9:00 AM - 12:00 PM)

Module 18. Generative AI for text data

a. Large Language Models(LLM)

b. Key concepts: Tokens, embeddings, and attention mechanisms

c. Applications

i. Text generation and completion

ii. Machine translation and summarization

iii. Conversational agents and chatbots

iv. Sentiment analysis

Module 19. Vector Databases

a. Implementing ChromaDB in Generative AI

b. Data Ingestion, chunking etc.

c. Querying and Retrieval

d. Integration with AI Models

Afternoon Session (1:00 PM - 5:00 PM)

Module 20. Retrieval Augmented Generation

a. Retrieval Mechanisms: Techniques and algorithms

b. Generation Models: Overview of language models used

c. Integration of Retrieval and Generation: How they work together

Module 21. Hands on building a simple RAG Chatbot