Python Text Analysis with NLP

Duration: 5 days / 40 hours
Prerequisites: Knowledge of Python Programming

Day 1: Introduction to NLP and Text Preprocessing

Module 1: Introduction to Natural Language Processing

- What is NLP?
- Applications of NLP
- Overview of the NLP pipeline

Module 2: Working with Text Data

- Text encoding (ASCII, UTF-8)
- Reading and writing text data in Python
- Tokenization concepts

Module 3: Text Cleaning and Preprocessing

- Lowercasing, punctuation removal, stopword removal
- Stemming vs Lemmatization
- Removing special characters and numbers

Lab 1: Text Cleaning Pipeline

- Build a preprocessing function using NLTK
- Tokenize text and remove noise from sample datasets (e.g., movie reviews)

Day 2: Linguistic Features and NLP Tools

Module 4: Part-of-Speech Tagging and Named Entity Recognition

- POS tagging with NLTK and spaCy
- Named Entity Recognition (NER) with spaCy

Module 5: Dependency Parsing and Sentence Structure

- Sentence segmentation
- Syntactic parsing with spaCy

Lab 2: Linguistic Analysis with spaCy

- Load and process text with spaCy
- Identify POS, Named Entities, and dependencies

Day 3: Text Representation Techniques

Module 6: Bag-of-Words and TF-IDF

- CountVectorizer
- TF-IDF Vectorizer
- Term frequency analysis

Module 7: Word Embeddings

- Introduction to word2vec, GloVe
- Using pre-trained word embeddings in spaCy and Gensim

Lab 3: Text Vectorization

- Convert a text dataset to numerical features using BoW and TF-IDF
- Visualize top keywords in documents

Day 4: Text Classification and Sentiment Analysis

Module 8: Text Classification with Machine Learning

- Naive Bayes, Logistic Regression
- Splitting data and model evaluation

Module 9: Sentiment Analysis

- Polarity and subjectivity (TextBlob)
- Sentiment classification with labeled datasets (e.g., IMDb, Twitter)

Lab 4: Building a Sentiment Classifier

- Train a model to classify sentiment from text
- Evaluate using accuracy, precision, recall, F1-score

Day 5: Advanced NLP & Project Work

Module 10: Topic Modeling

- Introduction to topic modeling
- Latent Dirichlet Allocation (LDA) with Gensim

Module 11: Text Summarization and Question Answering

- Extractive summarization with spaCy and transformers
- QA with pre-trained transformer models (HuggingFace)

Lab 5: Mini Project