INTRODUCTION TO NATURAL LANGUAGE PROCESSING

COURSE CODE: 20ITH102

L	T	P	C
3	1	0	4

Pre-requisites: Data Structures & Algorithms, Python Programming Lab, Numerical Methods, Probability and Statistics, Calculus and Linear Algebra.

COURSE OUTCOMES:

At the end of the course the student shall be able to

CO1: Explain how NLP is applied in the real world. (L2)

CO2: Compare a traditional NLP pipeline and a DL-based NLP pipeline. (L2)

CO3: Select appropriate text representation from vectorization features, embeddings and handcrafted features. (L3)

CO4: Apply text classification to real world problems. (L3)

CO5: Build solutions for different Information Extraction tasks. (L3)

UNIT-I **(10 LECTURES)**

NLP: A PRIMER:

NLP in the Real World, NLP Tasks, What Is Language? Building Blocks of Language, Why Is NLP Challenging? Machine Learning, Deep Learning, and NLP: An Overview, Approaches to NLP, Heuristics-Based NLP, Machine Learning for NLP, Deep Learning for NLP, Why Deep Learning Is Not Yet the Silver Bullet for NLP, An NLP Walkthrough: Conversational Agents.

Learning Outcomes: At the end of the module student will be able to:

1. describe the building blocks of language. (L2)
2. explain various applications of NLP in the real world. (L2)
3. describe various NLP tasks. (L2)

UNIT-II **(10 LECTURES)**

NLP PIPELINE:

Data Acquisition, Text Extraction and Cleanup, HTML Parsing and Cleanup, Unicode Normalization, Spelling Correction, System-Specific Error Correction, Pre-Processing, Preliminaries, Frequent Steps, Other Pre-Processing Steps, Advanced Processing Feature Engineering, Classical NLP/ML Pipeline, DL Pipeline, Modeling, Start with Simple Heuristics, Building Your Model, Building THE Model, Evaluation, Intrinsic Evaluation Extrinsic Evaluation, Post-Modeling Phases, Deployment, Monitoring, Model Updating Working with Other Languages, Case Study.

Learning Outcomes: At the end of the module student will be able to:

1. describe main components of a generic NLP. (L2)
2. summarize various pre-processing techniques to clean the data. (L2)
3. illustrate feature engineering while building NLP models. (L2)

UNIT-III **(10 LECTURES)**

TEXT REPRESENTATION:

Vector Space Models, Basic Vectorization Approaches, One-Hot Encoding, Bag of Words, Bag of N-Grams, TF-IDF, Distributed Representations, Word Embeddings, Going Beyond Words, Distributed Representations Beyond Words and Characters, Universal Text Representations, Visualizing Embeddings, Handcrafted Feature Representations.

Learning Outcomes: At the end of the module student will be able to:

1. describe vector space models. (L2)
2. choose one of the various vectorization approaches. (L3)

3. prepare Handcrafted Feature Representations. (L3)

UNIT-IV **(12 LECTURES)**

TEXT CLASSIFICATION:

Applications, A Pipeline for Building Text Classification Systems, A Simple Classifier Without the Text Classification Pipeline, Using Existing Text Classification APIs, One Pipeline, Many Classifiers, Naive Bayes Classifier, Logistic Regression, Support Vector Machine, Using Neural Embeddings in Text Classification. Word Embeddings, Subword Embeddings and fastText, Document Embeddings, Deep Learning for Text Classification, CNNs for Text Classification, LSTMs for Text Classification, Text Classification with Large, Pre-Trained Language Models, Interpreting Text Classification Models, Explaining Classifier Predictions with Lime, Learning with No or Less Data and Adapting to New Domains, No Training Data, Less Training Data: Active Learning and Domain Adaptation, Case Study: Corporate Ticketing.

Learning Outcomes: At the end of the module student will be able to:

1. describe pipeline for building text classification systems. (L2)
2. Utilize existing text classification APIs. (L3)
3. apply pre-trained models to classify the data. (L3)

UNIT-V **(10 LECTURES)**

INFORMATION EXTRACTION:

IE Applications, IE Tasks, The General Pipeline for IE, Keyphrase Extraction, Implementing KPE, Practical Advice, Named Entity Recognition, Building an NER System, NER Using an Existing Library, NER Using Active Learning, Practical Advice, Named Entity Disambiguation and Linking, NEL Using Azure API, Relationship Extraction, Approaches to RE, RE with the Watson API, Other Advanced IE Tasks, Temporal Information Extraction, Event Extraction, Template Filling, Case Study.

Learning Outcomes: At the end of the module student will be able to:

1. describe temporal information extraction. (L2)
2. explain various relationship extraction approaches. (L2)
3. apply various APIs to relationship extraction. (L3)

TEXT BOOKS:

1. Sowmya Vajjala,Bodhisattwa Majumder, Anuj Gupta & Harshit Surana, *"Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems"*, 1st Edition, y O'Reilly Media, Inc., June 2020.

REFERENCES:

1. Daniel Jurafsky & James H Martin, *"Speech and Natural Language Processing"*, 2nd Edition, Pearson Publications, 2013.
2. Tanvier Siddiqui, U.S. Tiwary, *"Natural Language Processing and Information Retrieval"*, 1st Edition, Oxford Higher Education, 2008.

WEB-REFERENCE:

1. https://nptel.ac.in/courses/106106211