Semester: VI						
NATURAL LANGUAGE PROCESSING AND TRANSFORMERS **Category: Professional Core Course** **(Theory and Practice)**						
Course Code	**:**	**AI363IA**		**CIE**	**:**	**100 + 50 Marks**
Credits: L:T:P	**:**	**3:0:1**		**SEE**	**:**	**100 + 50 Marks**
Total Hours	**:**	**45T + 30L**		**SEE Duration**	**:**	**3.00 Hours**

Unit-I	**9 Hrs**

Introduction to NLP: NLP in the Real-world, NLP Tasks, what is Language: Building Blocks of Language, Why NLP is Challenging, Machine Learning, Deep Learning, and NLP: An Overview, Approaches to NLP: Heuristic-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

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

Unit II	**9 Hrs**

Tokenizing Text and WordNet Basics: Introduction, Tokenizing text into sentences, Tokenizing sentences into words, Tokenizing sentences using regular expressions, training a sentence tokenizer, Filtering stop words in a tokenized sentence Looking up Synsets for a word in WordNet, looking up lemmas and synonyms in WordNet, Calculating WordNet Synset similarity, Discovering word collocations. Word similarity, Minimum Edit Distance algorithm.

Replacing and Correcting Words: Introduction, stemming words, Lemmatizing words with WordNet, replacing words matching regular expressions, removing repeating characters, Spelling correction with Enchant, replacing synonyms, Replacing negations with antonyms, word sense disambiguation, Feature-Based WSD, The Lesk Algorithm as WSD Baseline

Unit –III	**9 Hrs**

Part-of-speech Tagging: Pos Tagging approaches, The General Framework Rule-Based approaches, Transformation-Based learning, Modifications to TBL and Other Rule-Based Approaches, Markov Model Approaches, HMM-Based Taggers, Maximum Entropy Approaches, Taggers Based on ME Models, Default tagging, training a unigram part-of-speech tagger, combining taggers with backoff tagging, Training and combining n-gram taggers, creating a model of likely word tags, tagging with regular expressions, Affix tagging, training a Brill tagger, Training the TnT tagger, Using WordNet for tagging, tagging proper names, Classifier-based tagging, Training a tagger with NLTK-Trainer

Unit IV	**9 Hrs**

Transformers Basics
The Encoder-Decoder Framework, Attention Mechanisms, Transfer Learning in NLP, Hugging Face Transformers: Bridging the Gap, A Tour of Transformer Applications: Text Classification, Named Entity Recognition, Question Answering, Summarization, Translation, Text Generation, The Hugging Face Ecosystem: The Hugging Face Hub, Hugging Face Tokenizers, Hugging Face Datasets, Hugging Face Accelerate, Main Challenges with Transformers.

Text Classification
The Dataset: A First Look at Hugging Face Datasets, From Datasets to Data Frames, looking at the Class Distribution, How Long Are Our Tweets? From Text to Tokens: Character Tokenization, Word Tokenization, Sub-word Tokenization, Tokenizing the Whole Dataset, Training a Text Classifier: Transformers as Feature Extractors, Fine-Tuning Transformers

Unit V	9 Hrs
Transformer Anatomy The Transformer Architecture, The Encoder: Self-Attention, The Feed-Forward Layer, Adding Layer, Normalization, Positional Embeddings, adding a Classification Head, The Decoder, Meet the Transformers: The Transformer Tree of Life, The Encoder Branch, The Decoder Branch, The Encoder-Decoder Branch **Text Generation** The Challenge with Generating Coherent Text, Greedy Search Decoding, Beam Search Decoding, Sampling Methods, Top-k and Nucleus Sampling **Summarization** Text Summarization Pipelines, Summarization Baseline: GPT-2, T5, BART, PEGASUS	

PART-A • Implement the following application of Natural Language Processing • Demonstrate the working of the programs by considering appropriate datasets	
1	**Text Summarization**: Text summarization refers to the technique of shortening long pieces of text. The intention is to create a coherent and fluent summary having only the main points outlined in the document.
2	**World Cloud**: A word cloud is a collection, or cluster, of words depicted in different sizes. The bigger and bolder the word appears, the more often it's mentioned within a given text and the more important it is.
3	**Sentiment Analysis**: Design a program(without using library functions) to perform Sentiment analysis that analyzes digital text to determine if the emotional tone of the message is positive, negative, or neutral using the following vectorization techniques: • TF-IDF • N-GRAMS • Bag of words • One-hot encoding
4	**Topic Modelling**: Topic modeling is an unsupervised machine learning approach that can scan a series of documents, find word and phrase patterns within them, and automatically cluster word groupings and related expressions that best represent the set.

Course Outcomes: After completing the course, the students will be able to:-	
CO1	Apply various concepts, architectures, and frameworks of NLP to engineering problems
CO2	Demonstrate proficiency in utilizing the core and popular NLP libraries to provide solutions to real-world applications in Healthcare, Smart Cities, Agriculture, etc.
CO3	Design and Develop agents that use Transformers for natural language understanding and generation
CO4	Demonstrate the use of modern tools in solving day-to-day problems by exhibiting teamwork through oral presentations and reports
CO5	Collaborate in a group to build NLP solutions for the benefit of society

Reference Books	
1	Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems, Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta & Harshit Surana ,1st Edition, 2020, O'Reilly, ISBN: 978-1-492-05405-4
2	Python 3 Text Processing with NLTK 3 Cookbook, Jacob Perkins 2014, 1st Edition, Packt Publishing, ISBN 978-1-78216-785-3
3	Natural Language Processing with Transformers: Building Language Applications with Hugging Fac,Lewis Tunstall, Leandro von Werra, and Thomas Wolf, 2022, 1st Edition, O'Reilly Media, ISBN: 978-1-098-10324-8
4	Jurafsky, Dan., Martin, James H.Speech and Language Processing, 2nd Edition. United Kingdom: Pearson Prentice Hall, 2008.

5	Natural language processing, Eisenstein, Jacob, Online verfügbar unter https://princeton-nlp. github. io/cos484/readings/eisenstein-nlp-notes. pdf, zuletzt geprüft am 14 (2018): 2022.

RUBRICFOR THE CONTINUOUS INTERNAL EVALUATION

#	COMPONENTS	MARKS
1.	**QUIZZES:** Quizzes will be conducted in online/offline mode. **TWO QUIZZES** will be conducted & Each Quiz will be evaluated for 10 Marks. Each quiz is evaluated for 10 marks adding up to 20 MARKS	**20**
2.	**TESTS:** Students will be evaluated in test, descriptive questions with different complexity levels (Revised Bloom's Taxonomy Levels: Remembering, Understanding, Applying, Analyzing, Evaluating, and Creating). **TWO tests will be conducted**. Each test will be evaluated for **50Marks**, adding upto 100 Marks. **FINAL TEST MARKS WILL BE REDUCED TO 40 MARKS.**	**40**
3.	**EXPERIENTIAL LEARNING:** Students will be evaluated for their creativity and practical implementation of the problem. Case study based teaching learning (10), Program specific requirements (10), Video based seminar/presentation/demonstration (10) Designing & Modeling (10) **Phase 2 will be done in the exhibition mode (Demo/Prototype/any outcome). ADDING UPTO 40 MARKS**.	**40**
4.	**LAB:** Conduction of laboratory exercises, lab report, observation, and analysis (20 Marks), lab test (10 Marks) and Innovative Experiment/ Concept Design and Implementation (20 Marks) adding up to 50 Marks. THE FINAL MARKS WILL BE 50 MARKS	**50**
	MAXIMUM MARKS FOR THE CIE(THEORY+LAB)	**150**

RUBRIC FOR SEMESTER END EXAMINATION (THEORY)

Q.NO.	CONTENTS	MARKS
	PART A	
1	Objective type of questions covering entire syllabus	20
	PART B (Maximum of THREE Sub-divisions only)	
2	Unit 1 : (Compulsory)	16
3 & 4	Unit 2 : Question 3 or 4	16
5 & 6	Unit 3 : Question 5 or 6	16
7 & 8	Unit 4 : Question 7 or 8	16
9 & 10	Unit 5: Question 9 or 10	16
	TOTAL	**100**

RUBRIC FOR SEMESTER END EXAMINATION (LAB)

Q.NO.	CONTENTS	MARKS
1	Write Up	10
2	Conduction of the Experiments	20
3	Viva	20
	TOTAL	**50**