DHANALAKSHMI SRINIVASAN ENGINEERING COLLEGE
(AUTONOMOUS)
(Approved by AICTE & Affiliated to Anna University, Chennai)
Re-Accredited with 'A' Grade by NAAC, Accredited by TCS
Accredited by NBA – BME, ECE & EEE
PERAMBALUR - 621 212. Tamil Nadu.
website : www.dsengg.ac.in

COURSE PLAN

Course Code/Name	U20AI703/NATURAL LANGUAGE PROCESSING			
Year/Section/Department	IV/-/AI&DS			
Credits Details	L: 3	T: 1	P: 0	C: 4
Total Contact Hours Required	60			

Syllabus:

UNIT I/ INTRODUCTION	**No. Of Periods: 12**
Origins and challenges of NLP – Language Modeling: Grammar-based LM, Statistical LM - Regular Expressions, Finite-State Automata – English Morphology, Transducers for lexicon and rules, Tokenization, Detecting and Correcting Spelling Errors, Minimum Edit Distance.	
UNIT II/ WORD LEVEL ANALYSIS	**No. Of Periods: 12**
Unsmoothed N-grams, Evaluating N-grams, Smoothing, Interpolation and Backoff – Word Classes, Part of-Speech Tagging, Rule-based, Stochastic and Transformation-based tagging, Issues in PoS tagging – Hidden Markov and Maximum Entropy models.	
UNIT III/ SYNTACTIC ANALYSIS	**No. Of Periods: 12**
Context-Free Grammars, Grammar rules for English, Treebanks, Normal Forms for grammar – Dependency Grammar – Syntactic Parsing, Ambiguity, Dynamic Programming parsing – Shallow parsing – Probabilistic CFG, Probabilistic CYK, Probabilistic Lexicalized CFGs - Feature structures, Unification of feature structures.	
UNIT IV/ SEMANTICS AND PRAGMATICS	**No. Of Periods: 12**
Requirements for representation, First-Order Logic, Description Logics – Syntax-Driven Semantic analysis, Semantic attachments – Word Senses, Relations between Senses, Thematic Roles, selectional restrictions – Word Sense Disambiguation, WSD using Supervised, Dictionary & Thesaurus, Bootstrapping methods – Word Similarity using Thesaurus and Distributional methods.	
UNIT V/ DISCOURSE ANALYSIS AND LEXICAL RESOURCES	**No. Of Periods: 12**

Discourse segmentation, Coherence – Reference Phenomena, Anaphora Resolution using Hobbs and Centering Algorithm – Co reference Resolution – Resources: Porter Stemmer, Lemmatizer, Penn Treebank, Brill's Tagger, WordNet, PropBank, FrameNet, Brown Corpus, British National Corpus BNC.

Objective:

- To learn the fundamentals of natural language processing
- To understand the use of CFG and PCFG in NLP
- To understand the role of semantics of sentences and pragmatics
- To apply the NLP techniques to IR applications

Text Book:

1. Daniel Jurafsky and James H. Martin "Speech and Language Processing", 3rd edition Prentice Hall, 2009
2. Steven Bird, Ewan Klein, Edward Loper, Natural Language Processing with Python– Analyzing Text with the Natural Language Toolkit(O'Reilly2009, website 2018)

Website:

W1: https://www.tutorialspoint.com//natural Language Processing/index.htm
W2: https://www.goseeko/Natural language processing/index.htm

Online Mode of Study:

NPTEL https://www.youtube.com/watch?v=BAUjhHmjgy4
- https://www.youtube.com/watch?v=J-Ny5qlW4F4
- https://www.youtube.com/watch?v=xq3Lqp4QHUA

Course Plan:

Topic Number	Topic	Reference Detail	Page Number	Mode of teaching	Number of Periods Required	Cumulative Period
UNIT 1-INTRODUCTION						
1	Origins and challenges of NLP	T1,w1	3-7	BB	1	1
2	Language Modeling: Grammar-based LM	T1,w2	9-16	BB	1	2
3	Statistical LM	T1, T2	18-23	BB	1	3
4	Regular Expressions	T1, T2	41-50	BB	2	5
5	Finite-State Automata	T1	84-90	BB	2	7
6	English Morphology	T1	52-59	BB	1	8
7	Transducers for lexicon and rules	T1	61-70	BB	1	9
8	Tokenization	T1, T2	61-70	BB	1	10
9	Detecting and Correcting Spelling Errors	T1, T2	70-76	BB	1	11
10	Minimum Edit Distance	W2	-	BB	1	12
Outcome of Unit I: **CO1:** Design an innovative application using NLP components.						
UNIT II – WORD LEVEL ANALYSIS						
11.	Unsmoothed N-grams	T1	97-106	BB	1	13
12.	Evaluating N-grams	T2	108-113	BB	1	14
13.	Smoothing, Interpolation and Backoff	T2	115-117	BB	1	15
14.	Word Classes	T1	117-118	BB	1	16
15.	Part of-Speech Tagging	T1	119-120	BB	2	18
16.	Rule-based Tagging	T1	150-151 169-174	BB	2	20
17.	Stochastic and Transformation-based tagging	T2,W1	176-181 226-232	BB	1	21

18.	Issues in PoS tagging	T1	186-189	BB	2	23
19.	Hidden Markov and Maximum Entropy models	T1,W1	192-197	BB	1	24
Outcome of Unit II: **CO3:** Explain the Brute Force method and Divide and Conquer method to solve computing problems. **CO4:** Analysis the sorting and search methods.						
UNIT III – SYNTACTIC ANALYSIS						
19.	Context-Free Grammars	T3	283-290	BB	1	25
20.	Grammar rules for English, Treebanks	T3	304-311	BB	1	26
21.	Normal Forms for grammar	T3	258-264	BB	2	28
22.	Dependency Grammar	T2, R1	297-302	BB	1	29
23.	Syntactic Parsing, Ambiguity	T2, R2	292-296	BB	2	31
24.	Dynamic Programming parsing, Shallow parsing	R2	197-198	BB	2	33
25.	Probabilistic CFG, Probabilistic CYK	T3	318-327	BB	1	34
26.	Probabilistic Lexicalized CFGs	T3	318-327	BB	1	35
27.	Feature structures, Unification of feature structures.	T3	318-327	BB	1	36
Outcome of Unit III: **CO3:** Compare the multistage graphs and optimal binary search trees.						
UNIT IV - SEMANTICS AND PRAGMATICS						
28.	Requirements for representation	T1,R1	345-358	BB	1	37
29.	First-Order Logic, Description Logics	T2,w2	345-358	BB	2	39
30.	Syntax-Driven Semantic analysis, Semantic attachments	T2,w2	362-367	BB	2	41
31.	Word Senses, Relations between Senses, Thematic Roles, selectional restrictions	T2,w2	368-374	BB	1	42
32.	Word Sense Disambiguation, WSD	W2	-	BB	2	44

	using Supervised					
33.	Dictionary & Thesaurus	T1,R1	361-370	BB	1	45
34.	Bootstrapping methods	T1,R3	372-378	BB	1	46
35.	Word Similarity using Thesaurus and Distributional methods.	T1,R3	372-378	BB	1	47

Outcome of Unit IV:

CO4: Describe how scientific problems can be solved using iterative method and how to cope with Limitations of algorithm power.

UNIT V - DISCOURSE ANALYSIS AND LEXICAL RESOURCES

36.	Discourse segmentation, Coherence	T2,R4	388-409	BB	1	48
37.	Reference Phenomena, Anaphora Resolution using Hobbs and Centering Algorithm	T2,R4	424-427	BB	1	49
38.	Co reference Resolution	R1	427-430	BB	3	52
39.	Resources: Porter Stemmer, Lemmatizer	T1	424-427	BB	3	55
40.	Penn Treebank, Brill's Tagger	T1	432-436	BB	2	57
41.	WordNet, PropBank	R2	436-438	BB	1	58
42.	FrameNet, Brown Corpus	T2,W1	438-440	BB	1	59
43.	British National Corpus BNC.	T2,T2	441-443	BB	1	60

Outcome of Unit V:

CO5: Compare & analyze the different algorithm design techniques for a given problem based on its Time and space complexity.

CO6: Build existing algorithms to improve efficiency

Course Outcome:

At the end of course: Students should be able to do:

CO1: Discuss the fundamental concepts problem solving algorithm, its types and the parameters to analyze those algorithms (K2)
CO2: Explain the Brute Force method and Divide and Conquer method to solve computing problems. (K2)
CO3: Explain the dynamic programming and greedy techniques to solve computing problems. (K2)

CO4: Describe how scientific problems can be solved using iterative method and how to cope with limitations of algorithm power. (K2)

CO5: Compare and analyze the different algorithm design techniques for a given problem based on its time and space complexity. (K3)
CO6: Build existing algorithms to improve efficiency (K3)

Course Outcome Vs Program Outcome Mapping:

COs	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2
CO1	3	1	-	-	-	-	-	-	-	-	-	-	2	2
CO2	3	1	-	-	-	-	-	-	-	-	-	-	2	2
CO3	3	1	-	-	-	-	-	-	-	-	-	-	2	2
CO4	3	1	-	-	-	-	-	-	-	-	-	-	2	2
CO5	2	2	1	1	-	-	-	-	-	-	-	-	2	2
CO6	2	2	1	1	-	-	-	-	-	-	-	-	2	2
AVG	2.33	1.33	1.00	1.00	-	-	-	-	-	-	-	-	2.00	2.00

Internal Evaluation Components:

Webportal	Assignment	Components	Topic Number with Topic / Unit Details	Relevance to CO
Webportal 1	--	Assessment – I (60)	Unit I and II	CO 1 & CO2
	1	Assignment – Handwritten (20)	1. Language Modeling: Grammar-based LM, 2. Statistical LM 3.Rule based Parts of Speech tagging	

	2	**Assignment – Poster Presentation / PPT (20)**	1.Finite state automata 2. N-grams 3. Stochastic and Transformation-based tagging	
Webportal 2	**--**	**Assessment – II (60)**		**CO3 & CO4**
	3	**Seminar (20)**	1.Normal forms for Grammar 2.Syntactic parsing, dynamic programming parsing 3.First order logic , Description logics	
	4	**Case Study Report (20)**	1.probabilistic CFG, probabilistic CYK 2. Word Senses, Relations between Senses 3. Word Similarity using Thesaurus and Distributional methods.	
Webportal 3	**--**	**Model Exam (75)**		**CO1 to CO6**
	5	**MCQ (15)**	1. Detecting and Correcting Spelling Errors 2. Hidden Markov and Maximum Entropy models. 3. Context-Free Grammars, Ambiguity 4. Bootstrapping methods 5. Porter Stemmer, Brill's Tagger	**CO1 to CO6**
	-	**Course Attendance (10)**	**Attendance Percentage** >=75% = 2 Mark >=80% = 4 Mark >=85% = 6 Mark >=90% = 8 Mark >=95% = 10 Mark	**--**

Submission Details:

Phase 1(Before AT 1)		**Phase 2 (Before AT 2)**		**Phase 3 (Model)**
Assignment 1	**Assignment 2**	**Assignment 3**	**Assignment 4**	**Assignment 5**

Google Class Code Details: qgcw42j

Class Name: IV year

PLAN OF ASSESSMENT TEST –DISTRIBUTION OF MARKS:

TEST	CO- MARK WISE DISTRIBUTION						BLOOM'S LEVEL MARK WISE DISTRIBUTION					
AT-1	CO1	CO2	CO3	CO4	CO5	CO6	BTL1	BTL2	BTL3	BTL4	BTL5	BTL6
AT-2	CO1	CO2	CO3	CO4	CO5	CO6	BTL1	BTL2	BTL3	BTL4	BTL5	BTL6
MODEL	CO1	CO2	CO3	CO4	CO5	CO6	BTL1	BTL2	BTL3	BTL4	BTL5	BTL6

Prepared By

Verified By
HOD / AI & DS

Approved By
PRINCIPAL