CTMTAIDS SII P3: Natural Language Processing

Teaching Scheme					Evaluation Scheme									
					Theory							Practical		
L	T	P	C	TCH	Internal Exams					University Exams		University Exams (LPW)		Total
					TA-1		MSE		TA-2 *					
					Marks	Hrs	Marks	Hrs	Marks	Marks	Hrs	Marks	Hrs	
03	00	00	03	03	25	00:45	50	01:30	25	100	03:00	-	-	200

* Note: TA-2 will be in form of assignments or workshops.

Objectives

1. To learn the concept of scripting for cyber security.
2. To learn Python basics for scripting.
3. To learn PowerShell scripting for information security concept.
4. To learn bash scripting for cyber security related task.
5. To learn the concept related secure development and threat hunting.

UNIT – I

Origins and challenges of NLP, Language Modelling, 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

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

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

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 and Thesaurus, Bootstrapping methods, Word Similarity using Thesaurus and Distributional methods.

UNIT – V

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

Reference Books:-

1. Daniel Jurafsky, James H. Martin Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech, Pearson Publication, 2014.
2. Steven Bird, Ewan Klein and Edward Loper, Natural Language Processing with Python, First Edition, O_Reilly Media, 2009.
3. Breck Baldwin, Language Processing with Java and LingPipe Cookbook, Atlantic Publisher, 2015.
4. Richard M Reese, Natural Language Processing with Java‖, O_Reilly Media, 2015.
5. Nitin Indurkhya and Fred J. Damerau, Handbook of Natural Language Processing, Second Edition, Chapman and Hall/CRC Press, 2010.
6. Tanveer Siddiqui, U.S. Tiwary, Natural Language Processing and Information Retrieval‖, Oxford University Press, 2008.