ST. ALOYSIUS COLLEGE(AUTONOMOUS), JABALPUR

Reaccredited 'A+ 'Grade by NAAC(CGPA:3.68/4.00)
College with Potential for Excellence by UGC
DST-FIST Supported & STAR College Scheme by DBT

Faculty of Science

Bachelor of Science (B.Sc.) VI Semester

Paper-DSE-IIB

Text Mining using NLP

Course Outcomes

CO. No.	Course Outcomes	Cognitive Level
CO 1	Analyze and apply morphological analysis techniques such as lemmatization, finite automata, and finite state transducers.	U, A
CO 2	Perform Part-of-Speech (POS) tagging using rule-based and stochastic methods, and understand sequence labeling with HMM and Maximum Entropy models.	K
CO 3	Understand lexical semantics and perform word sense disambiguation using various approaches including dictionary-based methods and WordNet.	U
CO 4	Apply selectional restrictions and word similarity techniques using thesaurus and distributional methods for improved pragmatics and word sense disambiguation.	U, An
CO 5	Conduct discourse analysis, including anaphora and coreference resolution, and utilize lexical resources such as Penn Treebank, WordNet, and FrameNet.	U

Credit and Marking Scheme

	Credits	Marks		Total Marks
		Internal	External	
Theory	3	40	60	**100**
Practical	1	40	60	**100**
Total	**4**	**200**		

Evaluation Scheme

	Marks	
	Internal	**External**
Theory	3 Internal Exams of 20 Marks (During the Semester) (Best 2 will be taken)	1 External Exams (At the End of the Semester)
Practical	3 Internal Exams (During the Semester) (Best 2 will be taken)	1 External Exams (At the End of the Semester)

ST. ALOYSIUS COLLEGE(AUTONOMOUS), JABALPUR

Reaccredited 'A+ 'Grade by NAAC(CGPA:3.68/4.00)
College with Potential for Excellence by UGC
DST-FIST Supported & STAR College Scheme by DBT

Bachelor of Science (B.Sc.) **Subject: Computer Science**
VI Semester **Paper: DSE-IIB, Text Mining using NLP**

No. of Lectures (in hours per week): 2 Hrs. per week
Total No. of Lectures: 60 Hrs. **Maximum Marks:** 60

Units	Topics	No. of Lectures
I	History of NLP, Generic NLP system, levels of NLP, Knowledge in language processing, Ambiguity in Natural language, stages in NLP, challenges of NLP, Applications of NLP.	10
II	Morphology analysis –survey of English Morphology, Inflectional morphology & Derivational morphology, Lemmatization, Regular expression, finite automata, finite state transducers (FST), Morphological parsing with FST, Lexicon free FST Porter stemmer. N –Grams- N-gram language model, Self-learning topics: N-gram for spelling correction.	15
III	Part-Of-Speech tagging (POS)- Tag set for English (Penn Treebank), Rule-based POS tagging, Stochastic POS tagging, Issues –Multiple tags & words, Unknown words. Introduction to CFG, Sequence labeling: Hidden Markov Model (HMM), Maximum Entropy	10
IV	PRAGMATICS Selectional restrictions – Word Sense Disambiguation, WSD using Supervised, Dictionary & Thesaurus, Bootstrapping methods – Word Similarity using Thesaurus and Distributional methods.	10
V	Text summarization- LEXRANK, Optimization-based approaches for summarization, Summarization evaluation, Text classification. Sentiment Analysis introduction, Sentiment Analysis - Affective lexicons, Learning affective lexicons, Computing with affective lexicons, Aspect-based sentiment analysis.	15

TEXTBOOKS:

- Daniel Jurafsky, James H. Martin, "Speech and Language Processing: An Introduction to NaturalLanguageProcessing, Computational Linguistics and Speech", Pearson Publication, 2014.
- Steven Bird, Ewan Klein and Edward Loper, "Natural Language Processing with Python, First Edition,O'Reilly Media, 2009.

REFERENCE BOOK:

- Breck Baldwin, "Language Processing with Java and LingPipe Cookbook", Atlantic Publisher, 2015.
- Richard M Reese, "Natural Language Processing with Java", O'Reilly Media, 2015.

ST. ALOYSIUS COLLEGE(AUTONOMOUS), JABALPUR

Reaccredited 'A+ 'Grade by NAAC(CGPA:3.68/4.00)
College with Potential for Excellence by UGC
DST-FIST Supported & STAR College Scheme by DBT

List of Practical

1. Design and implement an NLP pipeline that performs tokenization, lemmatization, POS tagging, and named entity recognition on a given text corpus.
2. Develop a morphological parser using finite state transducers (FST) for English words, and demonstrate its ability to handle inflectional and derivational morphology.
3. Construct an N-gram language model for a given text corpus and use it to perform tasks such as next-word prediction and spelling correction.
4. Implement rule-based and stochastic POS tagging on a sample text, and evaluate the accuracy of each method using the Penn Treebank tag set.
5. Train a Hidden Markov Model (HMM) for POS tagging and use it to tag a new text. Compare its performance with a Maximum Entropy model.
6. Implement a word sense disambiguation system using dictionary-based and supervised learning approaches. Evaluate the system on a set of ambiguous sentences.
7. Use WordNet to explore relationships among lexemes (homonymy, polysemy, synonymy, hyponymy) and implement a robust word sense disambiguation algorithm.
8. Perform discourse segmentation and anaphora resolution using Hobbs and Centering algorithms on a given text. Analyze the coherence and reference phenomena in the discourse.
9. Implement a text summarization system using LEXRANK or an optimization-based approach. Evaluate the summarization quality using standard evaluation metrics.
10. Develop an aspect-based sentiment analysis system that uses affective lexicons to analyze customer reviews. Implement the system and evaluate its accuracy on a given dataset.