

**Kadi Sarva Vishwavidyalaya**  
Faculty of Engineering and Technology  
**First Year Master of Engineering (Computer Engineering)**  
In Effect from Academic Year 2017-18

<b>Subject Code: MECE204-N-B</b>	<b>Subject Title: NATURAL LANGUAGE PROCESSING &amp; INFORMATION RETRIVAL</b>
----------------------------------	--

Teaching scheme				Total Credit	Evaluation Scheme					Total
L	T	P	Total		Theory		Mid Sem Exam	CIA	Pract.	
Hrs	Hrs	Hrs	Hrs		Hrs	Marks	Marks	Marks	Marks	
04	00	02	06	05	03	70	30	20	30	150

**Learning Objectives:**

- Learn to write code for text indexing and retrieval.
- Learn to evaluate information retrieval systems
- Learn to analyze textual and semi-structured data sets
- Learn about text similarity measure and Understanding about search engine
- Text Classification & Clustering
- Machine Translation

**Outline of the Course:**

Sr. No	Title of the Unit	Minimum Hours
1	Overview of text retrieval systems	5
2	IR Models	12
3	Query expansion and feedback	5
4	Text classification	5
5	Text clustering	5
6	Introduction to NLP	5
7	N-Gram Language Model	6
8	Statistical Graphical Model	10
9	NLP Applications	7
10	NLP and Deep Learning	4

**Total hours (Theory): 64**

**Total hours (Lab): 32**

**Total hours: 96**

**DETAILED SYLLABUS:**

**Kadi Sarva Vishwavidyalaya**  
Faculty of Engineering and Technology  
**First Year Master of Engineering (Computer Engineering)**  
In Effect from Academic Year 2017-18

Sr. No	Topic	Lecture Hours	Weight age(%)
1	<b>Overview of text retrieval systems</b> <ul style="list-style-type: none"> <li>• Boolean retrieval</li> <li>• The term vocabulary and postings lists</li> <li>• Index compression</li> </ul>	5	8
2	<b>IR Models</b> <ul style="list-style-type: none"> <li>• Vector Space Model</li> <li>• TF-IDF Weight</li> <li>• Evaluation in information retrieval</li> <li>• Okapi/BM25;</li> <li>• Language models</li> <li>• KL-divergence;</li> <li>• Learning to Rank</li> </ul>	12	19
3	<b>Query expansion and feedback</b> <ul style="list-style-type: none"> <li>• Relevance feedback</li> <li>• pseudo relevance feedback</li> <li>• Query Reformulation</li> </ul>	5	8
4	<b>Text classification</b> <ul style="list-style-type: none"> <li>• The text classification problem</li> <li>• Naive Bayes text classification</li> <li>• k- nearest neighbors</li> <li>• Support vector Machine</li> <li>• Feature Selection</li> </ul>	5	8
5	<b>Text Clustering</b> <ul style="list-style-type: none"> <li>• Flat Clustering</li> <li>• K-means algorithm</li> <li>• PAM and PAMK</li> <li>• Hierarchical clustering</li> <li>• DBSCAN algorithm</li> </ul>	5	8
6	<b>Introduction to NLP</b> <ul style="list-style-type: none"> <li>• Introduction to NLP</li> <li>• Regular Expression, tokenization</li> <li>• Minimum Edit distance</li> </ul>	5	8
7	<b>N-Gram Language Model</b> <ul style="list-style-type: none"> <li>• Intro. N-Gram</li> <li>• N-Gram probability estimation and perplexity</li> <li>• Smoothing technique(Laplace/good Turing/Kneser-Ney/Interpolation/JM smoothing/Dirichlet)</li> </ul>	6	9
8	<b>Statistical Graphical Model</b> <ul style="list-style-type: none"> <li>• Hidden Markov Model(HMM)</li> </ul>	10	16

**Kadi Sarva Vishwavidyalaya**  
Faculty of Engineering and Technology  
**First Year Master of Engineering (Computer Engineering)**  
In Effect from Academic Year 2017-18

	<ul style="list-style-type: none"> <li>• Maximum Entropy Markov Model(MEMM)</li> <li>• Conditional Random field(CRF)</li> <li>• Latent Semantic Indexing</li> <li>• Topic Model</li> </ul>		
<b>9</b>	<b>NLP Applications</b> <ul style="list-style-type: none"> <li>• Sentiment Analysis</li> <li>• Information Extraction</li> <li>• Text Summarization</li> </ul> Machine Translation	<b>07</b>	<b>10</b>
<b>10</b>	<b>NLP and Deep Learning</b> Simple Word Vector representations: word2vec, GloVe	<b>04</b>	<b>06</b>
	<b>Total</b>	<b>64</b>	<b>100</b>

**Instructional Method and Pedagogy:**

- At the start of course, the course delivery pattern, prerequisite of the subject will be discussed.
- Lectures will be conducted with the aid of multi-media projector, black board, OHP etc.
- Attendance is compulsory in lecture and laboratory which carries 10 marks in overall evaluation.
- One internal exam will be conducted as a part of internal theory evaluation.
- Assignments based on the course content will be given to the students for each unit and will be evaluated at regular interval evaluation.
- Surprise tests/Quizzes/Seminar/tutorial will be conducted having a share of five marks in the overall internal evaluation.
- The course includes a laboratory, where students have an opportunity to build an appreciation for the concepts being taught in lectures.
- Experiments shall be performed in the laboratory related to course contents.

**Learning Outcome:**

- To Understand Document as Vector
- Performance evolution metric for IR
- To understand search Engine functionality
- Various Supervised and Unsupervised learning Method
- Basic technique for language processing
- Text analysis
- Machine translation

**Text Book:**

- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. <http://nlp.stanford.edu/IR-book/information-retrieval-book.html>
- ChengXiang Zhai, Statistical Language Models for Information Retrieval (Synthesis Lectures Series on Human Language Technologies), Morgan & Claypool Publishers, 2008. <http://www.morganclaypool.com/doi/abs/10.2200/S00158ED1V01Y200811HLT001>

**Kadi Sarva Vishwavidyalaya**  
Faculty of Engineering and Technology  
**First Year Master of Engineering (Computer Engineering)**  
In Effect from Academic Year 2017-18

- D. Jurafsky and J. Martin “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition”,
- C. Manning and H. Schutze, “Foundations of Statistical Natural Language Processing”,

**LIST OF PRACTICALS:**

Sr. No	Name of Experiment
1	Introduction to Lucerne/terrier/Indri and Sample index creation in Java/Python.
2	Basic Text Processing method on text document
2	Query Expansion and Ranking in Lucene
3	Implement Language Model with all smoothing technique
4	Implementation of various classification algorithm on text
5	Implementation of various Clustering algorithm on text
6	Develop Pos Tagger
7	Sentiment Analysis on Social Media data
8	Implement LSA and Topic model
9	Various track at TREC 2017 conference (students will be encouraged to participate in such track) <ul style="list-style-type: none"> <li>• Clinical Decision Support Track</li> <li>• Contextual Suggestion Track</li> <li>• Real Time Summarization Track</li> <li>• Temporal Summarization Track</li> <li>• Tasks Track</li> </ul>
10	Various track at CLEF 2015 Conference(students will be encouraged to participate in below track) <ul style="list-style-type: none"> <li>• Linked Data Track</li> <li>• Tweet Contextualization track</li> <li>• Relevance Feedback Track</li> </ul>