**Text Analytics**

(updated Jan. 2014)

Course Type: Elective

Credits: 03

***Objectives:***

The course aims to introduce the doctoral students to the problems of contemporary interest in Text Analytics and introduce algorithms and techniques to solve them. It describes different algorithmic paradigms and their application to recent problems in Text Analytics.

***Pre Requisites:***

The course is designed as an introductory course in the Text Analytics. It is positioned at the intersection of traditional Natural Language Processing, Machine Learning and Lexicon-based methods for Language Processing. The course as such does not require any specific know-how except a basic understanding of programming concepts. A prior course in Information Retrieval, Machine Learning and hands-on JAVA or R programming will, however, help improve the pace of learning.

***Contents:***

**Basics of Text Processing [20%]**: Statistical and Graphical NLP; Representation; Boolean and Vector Space Models; Feature Selection; Stop Words; Stemming; Parts of Speech Tagging; Graph Based Representations; IR view of Text Processing; Similarity Measures; Notion of Information Need, Precision and Recall.

**Classification and Clustering [30%]:** Supervised and Unsupervised Methods for Text Processing; Classification Methods such as Naïve Bayes, Nearest Neighbour, Rochio’s and Support Vector Machines; Clustering Methods such as Partitional and Hierarchical, Soft and Hard, K-Means, EM, Agglomerative Clustering; Datasets and Performance Measures.

**Applications [50%]:** Open and Targeted Information Extraction; Named Entity Recognition; Question Answering; Sentiment Analysis; Semantic Annotation; Document Summarization, Text Analytics in Commercial Settings.

**Laboratory Exercises.**

***TEXT AND REFERENCE BOOKS:***

* Christopher D. Manning, P. Raghvan and H. Schutze, Introduction to Information Retrieval, Cambridge University Press, 2008 [[**Companion Website**](http://nlp.stanford.edu/IR-book/)- contains downloadable book, slides and exercises]
* Christopher D. Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press. 1999. [[**Companion Website**](http://nlp.stanford.edu/fsnlp/)]
* R. Mihalcea and D. Radev, Graph based Natural Language Processing and Information Retrieval, Cambridge University Press, 2011.
* [Daniel Jurafsky](http://www.stanford.edu/~jurafsky/) and [James H. Martin](http://www.cs.colorado.edu/~martin/), Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition. Prentice-Hall, 2008. [[**Companion Website**](http://www.cs.colorado.edu/~martin/slp.html)]
* Ricardo Baeza-Yaets and Berthier Ribeiro-Neto, Modern Information Retrieval: The Concept and Technology behind Search, 2nd Edition, Addison-Wesley, 2011. [[**Companion Website**](http://www.mir2ed.org/) - contains certain downloadable chapters, slides and resources]
* G.S. Ingersol, T.S. Morton and A.L. Farris, Taming Text: How to Find, Organize and Manipulate It, Manning Publications, 2013.
* U.S. Tiwary and Tanveer Siddiqui, Natural Language Processing and Information Retrieval, Oxford University Press, 2008.

***WEB RESOURCES:***

Students are encouraged to visit following links about representative portals/ journals/ SIGs reporting research work in IR:

* [ACM Special Interest Group on IR (SIGIR)](http://www.acm.org/tocs/)
* [ACM Transactions on Information Systems](http://www.tois.acm.org)
* [ACM Transactions on Knowledge Discovery from Data](http://www.tkdd.acm.org)
* [ACM Transactions on Web](http://www.tweb.acm.org)
* [Data and Knowledge Engineering, Elsevier](http://www.sciencedirect.com/science/journal/0169023X)
* [Information Processing and Management, Elsevier](http://www.sciencedirect.com/science/journal/03064573)
* [IEEE Transactions on Knowledge and Data Engineering](http://www.computer.org/portal/web/tkde)
* [Journal of Information Science, Sage](http://www.jis.sagepub.com)
* [Data Mining and Knowledge Discovery, Springer](http://www.link.springer.com/journal/10618)
* [Information Retrieval, Springer](http://www.link.springer.com/journal/10791)
* [Knowledge and Information Systems, Springer](http://www.link.springer.com/journal/10115)
* [World Wide Web, Springer](http://www.link.springer.com/journal/11280)
* [Data Mining and Knowledge Discovery, Wiley](http://www.onlinelibrary.wiley.com/journal/10.1002/(ISSN)1942-4795)
* [Information Systems Journal, Wiley](http://www.onlinelibrary.wiley.com/journal/10.1111/(ISSN)1365-2575)
* [ACM SIGKDD](http://www.kdd.org)
* [Journal of Computational Linguistics](http://www.mitpressjournals.org/loi/coli)
* [Transactions of the Association for Computational Linguistics](http://www.transacl.org/)

***SOME leading CONFERENES:***

* [ACL: Association for Computational Linguistics](http://www.aclweb.org/)
* [EMNLP: Empirical Methods in Natural Language Processing](http://hum.csse.unimelb.edu.au/emnlp2013/)
* [NAACL: North American Chapter of the Association for Computational Linguistics](http://naacl.org/)
* [EACL: European Chapter of the Association for Computational Linguistics](http://www.eacl.org/)
* [COLING: International Conference on Computational Linguistics](http://www.coling-2014.org/)
* [CoNLL: Conference on Natural Language Learning](http://ifarm.nl/signll/conll/)
* [Conference on Information and Knowledge Management (CIKM)](http://www.cikmconference.org)
* [Text REtrieval Conference(TREC)](http://www.trec.nist.gov)
* [European Conference on Information Retrieval (ECIR)](http://www.ecir2014.org)

 Slides and PDF copies of some reading material will be shared as the class progresses. More printed and online material, particularly related to the assignments, will also be suggested.

***ASSESSMENT CRITERIA:***

Students will undergo a continuous evaluation process involving various elements at different stages of the course progress. A total of 20% and 40% marks of evaluation will be drawn from the mid-semester and end-semester examinations, respectively. The remaining 40% marks will be awarded based on successful completion of problem solving and programming assignments. More precisely the distribution of marks will be as follows:

Seminar - 20%

Home and Lab Assignment - 20%

Mid Semester Examination - 20%

End Semester Examination - 40%

***LAB and PRESENTATION ASSIGNMENTS***

* Assignment A
* Assignment B

***Queries and Feedback may be routed to***[***vivek@cs.sau.ac.in***](mailto:vivek@cs.sau.ac.in)