Approval: 7th Senate Meeting

Course Name: Text Retrieval and Mining

Course Number: CS 560

Credits: 3-0-0-3

Prerequisites: CS309 Information and Database Systems

Intended for: B Tech, MS and PhD

Distribution: Elective for CS

Semester:

Course Preamble: Availability of an enormous collection of text data and impact of information retrieval from the unstructured text data are the two key factors, which have triggered interest and research in text mining in recent years. In the beginning, the course briefly introduces the students to the field of Information Retrieval with the help of practical examples like web search engines. The course introduces the concept of text mining with a focus on exploratory data analysis in large collections of text along with techniques for text clustering, topic identification, and visualization of the results of those methods. Hands on assignments will focus on experimenting with clustering, topic identification, visualization etc with given datasets. Mini projects and assignments will carry weightage equivalent to one credit.

Course Modules:

- Introduction to IR: What IR means, what its goals are, what entities it attempts to retrieve, the criteria by which IR systems are evaluated. Web search engine as a case study. [3 Lectures]
- Introduction to TM: Structured vs Unstructured Data, Document Classification and Information Retrieval, Clustering and Organizing Documents, Textual Information to Numerical Vectors, Tokenization, Lemmatization, Stemming, Vector generation, Sentence boundary determination, Part of speech tagging, Parsing, Feature generation [8 Lectures]
- Text Mining Techniques: Similarity and nearest neighbor methods, Decision rules, Decision Trees, Linear Scoring methods, Key word search, Document matching, Inverted list, Clustering- K-means, Centroid clustering, Hierarchical clustering [8 Lectures]
- Looking for Information in Documents: Finding Patterns and entities, Coreference resolution, Relationship Extraction, Template filling and database construction [6 Lectures]
- Case Studies: Assigning topics to News Articles, Filtering email (Enron example) [3 Lectures]
- Hands on experience and mini projects

  Software R with text mining package will be used for text mining the following datasets:
  - Twitter dataset
  - Titanic survivor dataset
  - Blogging dataset
  - Spam training dataset

Projects will be based on the real world applications like (not an exhaustive list)

- Spam filtering
- Fraud detection by investigating notification of claims
- Automatic labeling of documents in business libraries
- Creating suggestion and recommendations (like amazon)
- Monitoring public opinions (for example in blogs or review sites)
- Measuring customer preferences by analyzing qualitative interviews
- Fighting cyberbullying or cybercrime in IM and IRC chat

Reference Books/Material:

1. Introduction to Information Retrieval, by C. Manning, P. Raghavan, and H. Schütze. Cambridge University Press, 2008.

2. Fundamentals of Predictive Text Mining by Sholom M. Weiss, Nitin Indurkhya, Tong Zhang, Editors David Gries Fred B. Schneider, Springer

3. W. Fan, L. Wallace, S. Rich, Z. Zhang, Tapping the power of text mining, Communications of ACM, 49(9), 76-82, 2006.

4. The Journal of Statistical Software article “Text Mining Infrastructure in R”

5. Introduction to the tm Package: Text Mining in R, By Ingo Feinerer, June 10, 2014 (http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf)