
Introduction to Text Mining: Theory and Applications

Mark J. Embrechts

Rensselaer Polytechnic Institute, Troy, NY, U.S.A.

http://www.rpi.edu/locker/82/001182/public_html/files/people/embrechts/embrechts.htm



Text mining is the process of automatically deriving novel, interesting and potential useful information from a single document or from vast amounts of documents. A key issue in text mining is how to express text in numbers. Typical text mining applications are text categorization, fingerprinting text for authorship, clustering, document summarization, deriving ontologies from collections of related texts, and the development of text-driven expert systems.

This short course is aimed for a general audience with no text mining experience, and will give an overview of text mining based on several case studies. Software for text mining developed by the author will be demonstrated at the short course and made available to the participants. Participants are encouraged to bring their laptop for hands-on experience.

This short course will also introduce visualization methods for text mining based on neural networks and statistics. More specifically visualization software and methods based on self-organizing maps, principal component analysis, partial-least squares, and independent component analysis will be demonstrated.

Demonstrations include:

- (i) Text categorization
- (ii) Text fingerprinting
- (iii) Application of textmining to bioinformatics (base-pair space and amino-acid space)
- (iv) Customer service analysis
- (v) Derivation of ontologies

A general text mining software package developed by the instructor will be distributed at the course. This software supports visualization with Partial-Least Squares, Principal Component Analysis, Independent Component Analysis and Self-Organizing Maps for text mining.

.....
Mark J. Embrechts is an Associate Professor of Decision Sciences and Engineering Systems at Rensselaer Polytechnic Institute, Troy NY. He has been with Rensselaer since 1983 and was previously employed as a postdoctoral staff member at Los Alamos National Laboratory (1981-1983). He is a pioneer in introducing neural networks, data mining and computational intelligence to the Graduate Engineering Curriculum at Rensselaer and has been teaching courses related to these topics since 1988. He has worked as a consultant or staff member at Los Alamos National Laboratory, Oak Ridge National Laboratory, the Max Planck Institute for Plasma Physics (Germany), the Paul Scherrer Institute (Switzerland), the World Bank (D.C. Headquarters), Nikkei Research (Japan), MetaLogic (Belgium), KODAK (USA), and General Electric (Schenectady, USA). He holds secondary appointments as Associate Professor of Environmental and Energy Engineering at Rensselaer, and in the School of Information Technology at Rensselaer. He was Guest Professor of Materials Science at the University of Leuven in 1990 (Belgium) and spent six months on sabbatical leave in the Center for Economical Studies at the University of Leuven.