Special Topics in Text Mining (CS 493)

Spring 2011 Syllabus

Instructor:
Manuel Montes y Gómez
http://ccc.inaoep.mx/~mmontesg/
mmontesg@inaoep.mx
Office hours: 9:00 – 11:00

Lecture sections:
Monday and Wednesday
CH 145; 04:00 pm-05:15 pm

Class web site:
http://ccc.inaoep.mx/~mmontesg/ → *teaching* → *Special Topics in Text Mining*

Course Description:
This course provides an advanced overview to some Text Mining tasks. It mainly focuses on presenting some state-of-the art approaches for text classification and document clustering. In addition, it introduces some important subtasks and applications, such as authorship attribution, sentiment classification, plagiarism detection, text summarization and visualization of search results. This course is intended for students that already have some background on data mining or machine learning.

Reference material:
- Michael W. Berry (Ed). *Survey of Text Mining: Clustering, Classification, and Retrieval*. Springer-Verlag, 2004.
- Ronen Feldman and James Sanger. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2006.
- Ashok Srivastava and Mehran Sahami. *Text Mining: Classification, Clustering, and applications*. Chapman and Hall, 2009.

Topics:

Text classification
- *Introduction to the task*
  1. Feature selection
  2. Classification algorithms
- *Semi-supervised methods for text classification*
- *Problems of non-thematic text classification*
  1. Sentiment classification
  2. Authorship Attribution
  3. Plagiarism detection
- *Crosslingual text classification*
- *Link-based classification*

Document clustering

- *Introduction to the task*
  1. Clustering methods
  2. Clustering evaluation
- *Improving clustering performance*
  1. Unsupervised feature selection
  2. Cluster ensembles
- *Methods for multilingual clustering*
- *Applications of document clustering*
  1. Using clustering in information retrieval
  2. Using clustering in document summarization
  3. Using clustering in text classification

Grading:

Final grades will be based on a combination of homework assignments, in-class attendance and performance, quizzes, and a final project that includes a write up and a presentation. There will be one assignment every two weeks related to the reading and analysis of research papers.

The approximate percentages are as follows:

45% - Assignments (bi-weekly)
25% - Final Project
10% - In-class participation
20% - Quizzes

Additionally, any one of the following will result on a final grade of F, even if the overall average is greater than 60%.

- Obtaining an average of less than 60% on the assignments and project
- Missing more than five lectures

The nominal percentage-score-to-letter-grade conversion is as follows:

- 90% or higher is an A
- 80-89% is a B
- 70-79% is a C
- 60-69% is a D
- below 60% is an F

Assignments:

Homework assignments will be posted at the webpage and discussed in class. All assignments will consider the writing of a *brief report about a selected research paper*. Written reports must include the description of the problem, the proposed solution and achieved results, and some ideas for work improvement.

Projects:

Projects will be related to the *extension of the work presented in one of the selected papers*. Projects will be proposed by students and validated by the instructor.

Late assignments:

All assignments up to two days late will receive up to 70% percent of full credit, and more than three days late will receive no credit. All class assignments are due at the beginning of the class period. The final project report/presentation will NOT be accepted after the due date.

Standards of Conduct and Academic Dishonesty:

Any form of academic dishonesty such as cheating, plagiarism, or the deliberate misrepresentation of fact will be dealt with severely. Students are free to discuss assignments, however, the work you turn in should be solely your own. Academic dishonesty will result in a failing course grade for those involved, and disciplinary actions by the University. Each student is responsible for the correctness of his or her own behavior; ignorance is not an excuse. If you are in doubt about something, ask me about it.