IS:688 Web Mining

Instructor: Christopher Markson, PhD
Class Location: Online
Office: GITC 5601
Email: crm23@njit.edu
Office Hours: by appointment

Course Description: Web mining aims to discover useful information and knowledge from the Web hyperlink structure, page contents and usage logs. It has direct applications in e-commerce, Web analytics, information retrieval/filtering, personalization, and recommender systems. Employees knowledgeable about Web mining techniques and their applications are highly sought by major Web companies such as Google, Amazon, Yahoo, MSN and others who need to understand user behavior and utilize discovered patterns from terabytes of user profile data to design more intelligent applications. The primary focus of this course is on Web usage mining and its applications to business intelligence and biomedical domains. We learn techniques from machine learning, data mining, text mining, and databases to extract useful knowledge from the Web and other unstructured/semi-structured, hyper-textual, distributed information repositories. This data could be used for site management, automatic personalization, recommendation, and user profiling. Topics covered include crawling, indexing, ranking and filtering algorithms using text and link analysis, applications to search, classification, tracking, monitoring, and Web intelligence. Programming assignments give hands-on experience. A group project highlights class topics.
(http://catalog.njit.edu/search/?P=is+688)

Required Background:

- Prerequisite Courses: IS 665
- Basic:
    a. Knowledge of statistics and data structure.
    b. Basic knowledge of database design and programming
- Advanced:
    a. Data mining related courses

Course Website: Moodle.com

Textbook:
Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2nd edition, Bing Liu, Springer, 2011, ISBN-10: 3642194591.
Reference: Introduction to Data Mining, by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Pearson/Addison Wesley, 2006, ISBN-10: 0321321367 (Not required but highly recommended)

Moodle: Additional material and resources will be found on the class website on Moodle, (http://moodle.njit.edu). It will be modified and updated as the course progresses and will contain the most recent information.

Special Arrangement: There is an increasing demand from employers for the graduating students with the transferrable learning skills, including self-regulation, resilient, openness, communication skills, etc. These skills enable learners to continually upgrade the knowledge and all the learning related skills through their own self-motivated learning, which are more likely to happen in the working environment. However, improving these transferrable learning skills requires tremendous extra time and energy. In order to both encourage the learning activities which improve the transferrable learning skills and satisfy the need for just getting the credit to graduate, the full score (100%) is solely determined by the traditional learning through lecture for basic knowledge and the extra credit is determined by the unconventional learning through problem-based learning for advanced knowledge. To get the full score

related to the lecture part, there is no requirement for the programming skill. All the related topics can be implemented in R, a popular data mining platform.

Credit: 3

Grade: Final grades will be based on:

Group Assignments:	25%
Project:	55%
Participation*:	20%

*Your participation grade will be computed based on how many video lectures you've watch and video lecture discussions you've participated in. Because students will be working in groups, it's important that everyone watches the lecture material at the same pace. Therefore, you will be required to watch the lecture videos within the posted week on Moodle in order to receive participation points. Unfortunately, extensions to this deadline will not be granted. You have a few options for meeting the 1 post requirement per lecture set.
To receive credit after watching the lecture video you can either:
1) Write a small post describing a topic in the lecture you think is particularly interesting or important
2) Ask a question for clarification on a topic, or
3) Answer another student's posted question in the forum.
The posts don't have to be long but should have some meaningful content in them.

The final letter grades for the semester are based solely on the points you earn (no curve).

Grade	Points
A	90+
B+	86-89
B	80-85
C+	76-79
C	70-75
F	0-69

Lecture Schedule: The following is a tentative schedule and subject to change. Refer to the class web page for most recent information. All of the readings are from the main textbook for the course. For most topics, there is a laboratory part to apply the related algorithms to the given sample dataset to examine the output in R.

Week	Topics
Week 1/2	Review Data Mining - Supervised, Unsupervised, and their Algorithms (Linear Regression, SVM, K-means)
Week 3	Introduction to R and important packages
Week 4	Web Crawling/Data Gathering
Week 5	Text Processing - Cleaning, Data Representations/DocTerms
Week 6	Text Transformation - Topic Modeling/LDA, Webpage Classifier
Week 7	Project Work
Week 8	NLP - Sentence splitting, Tokenization, POS tagging, Lemmas
Week 9	Sentiment Analysis/Opinion Mining
Week 10	Social Network Analysis
Week 11	Usage Mining (Association rules, etc.)
Week 12	Recommender Systems/Collaborative Filtering
Week 13/14	Reserved for topic overflow and project work
Week 15	Reserved for topic overflow and project work

POLICIES:

Assignments (Homework and Project)

Homework will be submitted via Moodle electronically. Late homework will be penalized 10% of the available points (and another 10% will be deducted for every 24-hour period after the original due date). After two days beyond the deadline, I will no longer accept homework submissions (No exceptions).

Makeup Tests

Requests for makeup tests must be made in advance with the instructor and will only be approved if the reason is beyond your control.

Project

The project will consist of a report and Powerpoint slides. This project should be more complicated than the homework assignments. The project should include the use of a web-based dataset, the analysis of the data, and code written in R.

Academic Integrity Policy

The NJIT academic honor code is located at: http://integrity.njit.edu/index.html. This honor code applies in its entirety to this class. Violations will not be tolerated. In addition, students should familiarize themselves with NJIT's "Best Practices related to Academic Integrity" which is developed and published on the Provost's website (on the policies page).

TURNITIN Policy

NJIT uses Turnitin.com, a service that helps prevent plagiarism on student papers. I will be using the Turnitin.com service at my discretion to determine the originality of student papers. If I submit your paper to Turnitin.com, it will be stored by Turnitin.com in their database as long as their service remains in existence. If you object to this storage of your paper, you must let me know no later than two weeks after the start of this class. If you object to the storage of your paper on Turnitin.com, I will utilize other services and techniques to check your work for plagiarism.

Disabilities

If you have a disability that may require some modification of seating, testing, or any other class requirement; please let the Professor know so that appropriate arrangements can be made. Similarly let the Professor know if you have any emergency medical information about which to be aware, or if you need special arrangements in the event of building evacuation. See the Professor after class hours or schedule an appointment. Assistance is available from the Office of Student Disability Services (205 Campbell Hall; 973-596-3420). Be sure and fill out appropriate paperwork with this office during the first week of class.