

## **IS: 688 Web Mining**

**Course Description:** This course will cover data mining techniques to mine the useful patterns from the Web hyperlink structure, page contents and usage logs. It can be applied to e-commerce, Web analytics, information retrieval/filtering, personalization, and recommender systems. The covered topics include machine learning, data mining, text mining, and the usage of databases. Through hands-on experience in lab and homework, you will be able to apply technical and analytic skills to develop a practical web application (or create a new web mining algorithm for publication).

**Term:** Fall 2017

**Pre-requisites:**

- Basic:
  - a. Knowledge of statistics and data structure.
- Advanced:
  - a. Data mining related courses.
  - b. Basic knowledge of database design and programming if you are shooting for publication.

**Instructor:** Lin Lin

**Office:** GITC 5600A

**Phone:** 973-596-5212

**Email:** [llin@njit.edu](mailto:llin@njit.edu).

**Office Hours:** Friday 4 - 5 pm

**Text:**

***Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data***, 2nd edition, Bing Liu, Springer, 2011, ISBN-10: 3642194591.

**Recommended Text:**

***Data Smart*** by John W. Foreman & ***Mining the Social Web*** by Matthew Russell

**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.

**Schedule:** The following is a tentative schedule and subject to change. Refer to class web page for most recent information.

Date	Topics	Reading
Week 1	Introduction and project idea	
Week 2	Data collection and preprocessing	Chapter 8.1-8.3
Week 3	Association rules	Chapter 2
Week 4	Supervised learning	Chapter 3
Week 5	Unsupervised learning	Chapter 4
Week 6	Evaluation	
Week 7	Information retrieval	Chapter 6
Week 8	Graph mining	Chapter 7
Week 9	Partially supervised learning	Chapter 5
Week 10	Web usage mining	Chapter 12
Week 11	TBD	
Week 12	TBD	
Week 13	Project presentation	
Week 14	Project presentation	
Week 15	TBD (final)	

**Laboratory Sessions:** This course does not have a separate laboratory session. However, some class meeting time throughout the semester will be dedicated to hands-on laboratory assignments. This work will be done using the computers in the classroom. If necessary, laboratory assignments should be worked on outside the class time.

**Credit:** 3

**Grade:** Final Grades will be based on:

Assignments: 30%

Labs: 20%

Presentation: 10%

Final Exam: 20%

Project: 20%

## **POLICIES:**

### **Assignments (Homework and Project)**

Homework for this class consists of 6 homework assignments. They are usually due about one week after being issued. Their purpose is to help you keep up with the material and assess your readiness for the midterm and final.

Homework is due before midnight (11:55pm) on the due date specified on the schedule. It 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), unless there is a reason beyond your control.

### **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.

### **Team Work Grading**

Normally, all the team members get the same points for their team project. However, when submitting the team project, each team member is also required to submit a document describing the percentage of contribution and the duty for each team member like the following.

Name	Percentage of contribution	Duty
A	30	Collecting data
B	40	Select the model and optimize the parameter
C	30	Evaluation

Upon request, I might assign different points to different team members.

### **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.