

CSCI 548

Information Integration on the Web

Spring 2011

Instructors: Kristina Lerman (lerman@isi.edu) and

Jose-Luis Ambite (ambite@isi.edu)

Meeting Time: Mondays and Wednesdays 3:30-4:50pm

Location: ZHS352

Office Hours:

Professor Lerman

- Immediately BEFORE class
- Or by appointment (ISI 932 or by phone: 310-448-8714)

Professor Ambite

- Immediate before or after class
- Or by appointment (ISI 935 or by phone: 310-448-8472)

Teaching Assistant/Grader:

Rumi Ghosh (rumig@usc.edu)

Course Web Page: USC Blackboard (blackboard.usc.edu)

This course will focus on the basic foundations and techniques in Information Extraction and Integration. There has been a great deal of interest and research on this topic and the course will cover the research and tools for addressing the technical problems. The topics covered will include data integration techniques, machine learning techniques for information extraction and wrapper construction, high-performance query execution systems based on streaming dataflow, constraint-based integration systems, approaches to record linkage for resolving naming inconsistencies across sites, and the challenges of accessing and integrating information from online social networking sites.

The class will be run as a lecture course with lots student participation and hands-on experience. As an integral part of the course each student will develop and build an integrated Web application using the research and tools covered in the class.

Prerequisites:

CSCI561 -- Introduction to AI
CSCI585 -- Database Systems

Recommended Course:

CSCI571— Web Technologies
CSCI573—Advanced AI

Grading:

Homework -- 20%
Course project -- 30%
Quizzes -- 20%
Final Exam -- 30%

Books: There is no required textbook. We will read technical papers on each topic.

Lab: There is no lab for this course. Students should contact the instructor if they do not have access to a computer where they can install their own software.

Course Syllabus and Schedule

- **January 10**
 - **Topic: Introduction (Professors Lerman and Ambite)**
- **January 12**
 - **Topic: Wrapper Learning (Professor Lerman)**
 - **Homework**
 - Assignment 1 (Due January 21 at midnight)
- **January 17**
 - **Martin Luther King's Birthday**
- **January 19**
 - **Topic: Automatic Wrapper Generation (Professor Lerman)**
 - **Homework**
 - Assignment 2 (Due January 28 at midnight)
- **January 24**
 - **Topic: Data Integration (Professor Ambite)**
 - **Quiz 1: Wrapper Learning/Generation**
- **January 26**
 - **Topic: Data Integration (Professor Ambite)**
 - **Homework**
 - Assignment 3 (Due February 4 at midnight)

- **January 31**
 - **Topic: Schema Mapping (Professor Lerman)**
 - **Quiz 2: Data Integration**

- **February 2**
 - **Topic: Source Modeling (Professor Lerman)**
 - **Homework**
 - Assignment 4 (Due February 11 at midnight)

- **February 7**
 - **Topic: Semantic Web (Professor Ambite)**
 - **Quiz 3: Schema Mapping/Source Modeling**

- **February 9**
 - **Topic: Semantic Web (Professor Ambite)**
 - **Homework**
 - Assignment 5 Due (February 18 at midnight)

- **February 14**
 - **Topic: Information Extraction (Professor Lerman)**
 - **Quiz 4: Semantic Web**

- **February 16**
 - **Topic: Information Extraction (Professor Lerman)**
 - **Homework**
 - Assignment 7 Due (February 25 at midnight)

- **February 21**
 - **President's Day!**

- **February 23**
 - **Topic: Mashup Construction (Guest lecture by Dr. Szekely)**
 - **Quiz 5: Information Extraction**
 - **Homework**
 - Assignment 6 Due (March 4 at midnight)

- **February 28**
 - **Topic: Record Linkage (Professor Ambite)**
 - **Quiz 6: Mashup Construction**

- **March 2**
 - **Topic: Record Linkage (Professor Ambite)**
 - **Project proposals due at 3pm on March 3 (submitted online)**
 - **Homework**
 - **Assignment 8 Due (March 11 at midnight)**

- **March 7**
 - **Topic: Linked Data (Professor Ambite)**
 - **Quiz 7: Record Linkage**

- **March 9**
 - **Topic: Sentiment Analysis (Professor Lerman)**
 -

- **March 14-19**
 - **Spring Break!**

- **March 21**
 - **Social Tagging (Professor Lerman)**
 - **Quiz 8: Linked Data/Sentiment Analysis**

- **March 23**
 - **Topic: Folksonomies (Professor Lerman)**

- **March 28**
 - **Topic: Topic: Data and Privacy in Social Networks (Professor Lerman)**
 - **Quiz 9: Social Tagging**

- **March 30**
 - **Topic: Ranking in Social Networks (Professor Lerman)**

- **April 4**
 - **Topic: Data Flow and Optimization (Guest lecture by Dr. Barish)**
 - **Quiz 10: Social Networks**

- **April 6**
 - **Topic: Constraint Integration (Professor Ambite)**

- **April 11**
 - **Topic: Advanced Topics in Data Integration (Professor Ambite)**
 - **Quiz 11: Constraint Integration/Data Flow**
- **April 13**
 - **Topic: Intellectual Property (Professor Lerman)**
- **April 18**
 - **Project Presentations**
- **April 20**
 - **Project Presentations**
- **April 25**
 - **Project Presentations**
- **April 27**
 - **Course Review (Professors Lerman and Ambite)**
- **Final Exam (Friday, May ?, 2-4pm)**
 - **Location: Classroom**