**USC** Viterbi School of Engineering

**CSCI 548: Information Integration on the Web**

**Units: 4**

**Fall 2015**

**Section 29980D**

Class: MW – 4:00-5:50pm, ZHS 163

Final Exam: Wednesday, Dec 9th, 4:30-6:30pm

**Section 30398D**

Class: MW – 6:00-7:50pm, GFS 101

Final Exam: Wednesday, Dec 9th, 7:00-9:00pm

**Instructor:** Pedro Szekely

**Office Hours:** Monday & Wednesday 3:30-4:00pm in front of ZHS when I teach, by appointment otherwise

**Contact Info**:  [[pszekely@isi.edu](mailto:pszekely@isi.edu),](mailto:ambite@isi.edu)  310-448-8641.

**Instructor:** Craig Knoblock

**Office:** AHF B55 Small Conference Room (Spatial Sciences)

**Office Hours:** Wednesdays 3-4pm or by appointment

**Contact Info:** [knoblock@usc.edu](mailto:knoblock@usc.edu), 310-448-8786.

**Teaching Assistant:** Chung Ming Cheung

**Office:** EEB 242

**Office Hours:** Thursdays 3:30-5:30pm

**Contact Info:** [chungmin@usc.edu](mailto:chungmin@usc.edu)

**Teaching Assistant:** Mohsen Taheriyan

**Office:** TBD (temporarily at SAL Computer Lab)

**Office Hours:** Mondays 2-4pm

**Contact Info:** [taheriya@usc.edu](mailto:taheriya@usc.edu)

**Prerequisite(s):** CSCI 561

**Co-Requisite (s):** none

**Concurrent Enrollment:** none

**Recommended Preparation**: CSCI 585 and some programming experience

**Online materials**: <http://bit.ly/cs548-2015-fall>

**Catalogue Course Description**

Foundations, techniques, and algorithms for information integration. Topics include Semantic Web, linked data, data integration, entity linkage, source modeling, and information extraction.

**Expanded Course Description**

This course focuses on foundations, techniques, and algorithms for information extraction, modeling and integration. Topics covered include semantic web (RDF, OWL, SPARQL), linked data and services, mash-ups, theory of data integration, schema mappings, record/entity linkage, data cleaning, source modeling, and information extraction. The class will be run as a lecture course with lots of student participation and significant hands-on experience. As an integral part of the course each student will do a project using the research and tools covered in the class.

**Learning Objectives**

The learning objectives for this course are:

* Understand the foundations and techniques of the Semantic Web, including RDF, OWL, SPARQL, linked data, and linked services
* Understand the theory and techniques of traditional data integration, including view integration, schema mapping, record linkage
* Understand the algorithms and techniques for data cleaning, source modeling, building mashups, semi-structured extraction, and information extraction
* Understand the theory and application of the state-of-the-art software and tools for information extraction
* For any given integration problem, be able to select and apply the most relevant information integration techniques to solve that problem

**Course Notes**

The course will be run as a lecture class with student participation strongly encouraged. There are weekly readings and students are encouraged to do the readings prior to the discussion in class. All of the course materials, including the readings, lecture slides, homeworks will be posted online at http://bit.ly/cs548. The class project is a significant aspect of this course and at the end of the semester, students will present their projects in class.

**Technological Proficiency and Hardware/Software Required**

Students are expected to know how to program in a language such as Java, C++, or Python. Students are also expected to have their own laptop or desktop computer where they can install and run software to do the weekly homework assignments.

**Required Readings and Supplementary Materials**

Required Textbook: Principles of Data Integration by Doan, Halevy, & Ives, Morgan Kaufmann, 2012

The book is available online from the USC library and is available for purchase.

All of the required readings are listed in the course schedule.

**Description and Assessment of Assignments**

**Homework Assignments**

There will be weekly homework assignments for the first 12 weeks of class. The assignments must be done individually. The homework assignments are expected to take 6-8 hours per week. Each assignment is graded on a scale of 0-100 and the specific rubric for each assignment is given in the assignment. The homework topics are listed in the Course Schedule.

**Course Project**

An integral part of this course is the course project, which builds on the topics and techniques covered in the class. Students can work in teams of up to two people on this project. They will present a project proposal in class, conduct the project, create a video demonstration of the work, and present the project in class.

*Sample project:*  “Geotagging Ansel Adams’ Photographs” Ansel Adams was one of the most famous American photographers. However, there is no single coherent source, which provides a structured collection of all of Ansel’s photographs. Also most of his photos are not geotagged and thus there is no easy way to visualize Ansel’s journey across the globe. This project extracted Ansel’s photos along with their metadata from various sources into a single coherent consolidated schema and geotagged them by extracting and identifying location entities from each photo’s metadata. The end result is a web application, which used the Google Maps API and Timeline JS to visualize his geotagged and time-stamped photos in a very impressive manner.

**Grading Breakdown**

Quizzes 20%

Homeworks 20%

Final: 20%

Class Project 40%

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Total 100%

**Quizzes:** There will be weekly quizzes based on the material from the week before. There will be no make-up quizzes for any reason, but we will drop the worst quiz grade.

**Homework:** There will be weekly homeworks based on the topics of the class each week. All homeworks must be done individually even when they are part of the group project. All homeworks must be submitted in Blackboard before 11:59pm PT on the date they are due in order to receive full credit. You may submit any homework up to one week late with for 75% of the grade. Homeworks more than one week late will receive no credit.

**Midterm**: There is no mid-term for this class.

**Final Exam:** There is a final exam at the end of the semester covering all of the material covered in the class.

**Class Project:** Each student will do a group project based on the topics covered in the class. Students will propose their own project, do the research and build a proof-of-concept, present the work in class, and create a video demonstration of the work. The grades of the project are based on:

* Proposal: 5%
* Project video: 30%
* Presentation: 35%
* Overall project innovation: 30%

Course grades will range from A through F. The following is the breakdown for grading:

94 - 100 = A 74 - 76 = C

90 – 93 = A- 70 - 73 = C-

87 – 89 = B+ 67 - 69 = D+

84 – 86 = B 64 - 66 = D

80 – 83 = B- 60 - 63 = D-

77 – 79 =C+ Below 60 is an F

**Assignment Submission Policy**

Homework assignments are due at 11:59pm on the due date and should be submitted in Blackboard. You can submit homework up to one week late, but you will loose 20% of the possible points for the assignment. After one week, the assignment cannot be submitted.

**Course Schedule: A Weekly Breakdown**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Topics** | **Readings** | **Homeworks** | **Instructor** |
| Aug 24 | Course Introduction |  |  | Professor Szekely |
| Aug 26 | RDF / Graph data model | Frank Manola and Eric Miller. RDF primer. Technical report, W3C, February 2004. <http://www.w3.org/TR/2004/REC-rdf-primer-20040210/>.  Tim Berners-Lee. Why rdf model is different from the xml model. Technical report, W3C, 1998. <http://www.w3.org/DesignIssues/RDF-XML.html>. |  | Professor Szekely |
| Aug 31 | Information Extraction 1 | Andrew McCallum. Information Extraction: Distilling Structured Data from Unstructured Text. ACM Queue, volume 3, Number 9, November 2005. <http://people.cs.umass.edu/~mccallum/papers/acm-queue-ie.pdf>  Matthew Michelson and Craig A. Knoblock. Semantic Annotation of Unstructured and Ungrammatical Text. In Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-2005), Edinburgh, Scotland, 2005. <http://www.isi.edu/integration/papers/michelson05-ijcai.pdf>  Hanna M. Wallach , Conditional Random Fields: An Introduction, 2004. <http://repository.upenn.edu/cgi/viewcontent.cgi?article=1011&context=cis_reports>  Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead and Oren Etzioni, Open Information Extraction from the Web, 2007. <https://homes.cs.washington.edu/~soderlan/OpenIE-ijcai07.pdf> | HW1: Crawling | Professor Knoblock |
| Sep 2 | Information Extraction 2 |  | Professor Knoblock |
| Sep 9 | Semi-Automatic Modeling (Karma) | Craig A. Knoblock, Pedro Szekely. Exploiting Semantics for Big Data Integration. AI Magazine, 2015. <http://usc-isi-i2.github.io/papers/knoblock15-aimagazine.pdf>  Mohsen Taheriyan, Craig A. Knoblock, Pedro Szekely, Jose ́ Luis Ambite. Learning the Semantics of Structured Data Sources. Journal of Web Semantics Special Issue on Knowledge Graphs, 2015. Available at [Blackboard](https://blackboard.usc.edu). | HW2:Information Extraction | Mohsen Taheriyan |
| Sep 14 | RDF Schema / Inference | RDF vocabulary description language 1.0: Rdf schema. Technical report, W3C, February 2004. <http://www.w3.org/TR/2004/REC-rdf-schema-20040210/>.  Ben Adida, Ivan Herman, Manu Sporny, and Mark Birbeck. Rdfa 1.1 primer rich structured data markup for web documents. Technical report, W3C, June 2012. <http://www.w3.org/TR/2012/NOTE-rdfa-primer-20120607/>. | HW3: RDF | Professor Szekely |
| Sep 16 | SPARQL Query Language | Steve Harris and Andy Seaborne. Sparql 1.1 query language. Technical report, W3C, January 2012. <http://www.w3.org/TR/2012/PR-sparql11-query-20121108/>. |  | Professor Szekely |
| Sep 21 | RDF Mapping Tools | Franck Michel, Johan Montagnat and Catherine Faron-Zucker. A survey of RDB to RDF translation approaches and tools. https://hal.archives-ouvertes.fr/hal-00903568 | HW4: SPARQL | Professor Szekely |
| Sep 23 | Wrapper Generation 1 | Ion Muslea, Steve Minton, and Craig A. Knoblock. A hierarchical approach to wrapper induction. In Proceedings of the 3rd International Conference on Autonomous Agents, Seattle, WA, 1999. <http://www.isi.edu/integration/papers/muslea99-agents.pdf>.  AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. Principles of Data Integration, chapter 9. Morgan Kaufmann, 2012. <http://www.sciencedirect.com/science/book/9780124160446>  W. Crescenzi, G. Mecca, and P. Merialdo. RoadRunner. Towards automatic data extraction from large web sites. 2001. <http://www.vldb.org/conf/2001/P109.pdf>.  B. Cenk Gazen and Steven Minton. Overview of autofeed: An unsupervised learning system for generating webfeeds. In Proceedings of AAAI, 2006. <http://www.isi.edu/integration/courses/csci548/Papers/gazen06-aaai.pdf>. |  | Professor Knoblock |
| Sep 28 | Wrapper Generation 2 | HW5:Wrapper | Professor Knoblock |
| Sep 30 | Linked Data | Linked Data: Evolving the Web into a Global Data Space  [Synthesis Lectures on the Semantic Web: Theory and Technology](http://www.morganclaypool.com/toc/wbe.1/1/1)  February 2011, 136 pages, (doi:10.2200/S00334ED1V01Y201102WBE001)  Tom Heath and Christian Bizer  Chapters 1 to and including section 4.2.  <https://west.uni-koblenz.de/files/ws1213/seminar-web-science/linked-data.pdf>://west.uni-koblenz.de/filews1213/seminar-web- |  | Professor Szekely |
| Oct 5 | REST / JSON-LD | M. Vaqqas, RESTful Web Services: A Tutorial. <http://www.drdobbs.com/web-development/restful-web-services-a-tutorial/240169069>  Roy T. Fielding, REST APIs must be hypertext-driven. <http://roy.gbiv.com/untangled/2008/rest-apis-must-be-hypertext-driven>  Markus Lanthaler and Christian Gütl. 2012. On using JSON-LD to create evolvable RESTful services. <http://smtp.markus-lanthaler.com/research/on-using-json-ld-to-create-evolvable-restful-services.pdf> | HW6:Karma | Professor Szekely |
| Oct 7 | String Matching 1 | AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. Principles of Data Integration, chapter 4. Morgan Kaufmann, 2012. <http://www.sciencedirect.com/science/book/9780124160446> |  | Professor Szekely |
| Oct 12 | String Matching 2 | HW7:String Matching | Majid Ghasemi |
| Oct 14 | Data Cleaning | Wrangler: Interactive visual specification of data transformation scripts. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2011. <http://vis.stanford.edu/papers/wrangler>.  Bo Wu, Pedro Szekely, and Craig A. Knoblock. Minimizing user effort in transforming data by example. In Proceedings of the International Conference on Intelligent User Interface, 2014. <http://www.isi.edu/integration/papers/wu14-iui.pdf>.  Open Refine, Explore data. <http://youtu.be/B70J_H_zAWM>.  Open Refine, Clean and transform data. <http://youtu.be/cO8NVCs_Ba0>.  Open Refine, Reconcile and match data. <http://youtu.be/5tsyz3ibYzk>. |  | Bo Wu |
| Oct 19 | Automatic Source Modeling | Mark James Carman and Craig A. Knoblock. Learning semantic descriptions of web information sources. In Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI), January 2007. <http://www.isi.edu/integration/papers/carman07-ijcai.pdf>.  Jose ́ Luis Ambite, Sirish Darbha, Aman Goel, Craig A. Knoblock, Kristina Lerman, Rahul Parundekar, and Thomas Russ. Automatically constructing semantic web services from online sources. In Proceedings of the 8th International Semantic Web Conference (ISWC 2009), 2009. <http://www.isi.edu/integration/papers/ambite09-iswc.pdf>. | HW8:Data Cleaning | Professor Knoblock |
| Oct 21 | Data Integration Basics | AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. Principles of Data Integration, chapter 2.1, 2.2, 2.3 and 2.4. Morgan Kaufmann, 2012.  <http://www.sciencedirect.com/science/book/9780124160446> |  | Professor Knoblock |
| Oct 26 | Logical Data Integration | HW9:Data Integration | Professor Knoblock |
| Oct 28 | NO CLASS |  | This class will be made up as a review class at the end of the semester. | No Class |
| Nov 2 | Record Matching 1 | AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. Principles of Data Integration, chapter 7. Morgan Kaufmann, 2012. <http://www.sciencedirect.com/science/book/9780124160446> | HW10:Record Linkage | Professor Szekely |
| Nov 4 | Record Matching 2 |  | Professor Szekely |
| Nov 9 | Querying Knowledge Graphs | TBD | HW11:Mashups | Jason Slepicka |
| Nov 11 | Mashups 1 | Shubham Gupta and Craig A. Knoblock. Building geospatial mashups to visualize information for crisis management. In Proceedings of the 7th International Conference on Information Systems for Crisis Response and Management, 2010. <http://www.isi.edu/integration/papers/gupta10-iscram.pdf>.  Jeffrey Wong and Jason I. Hong. Making mashups with marmite: towards end-user programming for the web. In ACM SIGMOD Record, 2007. <http://repository.cmu.edu/cgi/viewcontent.cgi?article=1063&context=hcii>.  Rob Ennals, Eric Brewer, Minos Garofalakis, Michael Shadle, and Prashant Gandhi. Intel mash maker: join the web. 2007.  <http://23.30.224.201/publications/intel-mash-maker-join-web>.  Huynh David, Mazzocchi Stefano, and Karger David. Piggy bank: Experience the semantic web inside your web browser. 2007. <http://simile.mit.edu/papers/iswc05.pdf>. |  | Professor Knoblock |
| Nov 16 | Mashups 2 |  | HW12:OWL | Professor Knoblock |
| Nov 18 | OWL 2 | Krtzsch Markus, Simancik Frantisek, and Horrocks Ian. A description logic primer. 2012. <http://arxiv.org/pdf/1201.4089.pdf>.  Hector Prez-Urbina, Ian Horrocks, and Boris Motik. Efficient query answering for owl 2. In International Semantic Web Conference, 2009. Efficient Query Answering for OWL 2. <https://www.cs.ox.ac.uk/boris.motik/pubs/puhm09query-OWL2.pdf>. |  | Professor Szekely |
| Nov 23 | Intellectual Property | Thomas P. Vartanian and Robert H. Ledig. Scrape it, scrub it and show it: The battle over data aggregation. [http://web.archive.org/web/20070818130311/http:/www.ffhsj.com/bancmail/bmarts/aba art.html](http://web.archive.org/web/20070818130311/http:/www.ffhsj.com/bancmail/bmarts/aba%20art.html).  Kembrew McLeod. Intellectual property law, freedom of expression, and the web, 2003. <http://www.electronicbookreview.com/thread/technocapitalism/proprietary>.  Electronic frontier foundation. <http://www.eff.org/issues/intellectual-property>. |  | Professor Knoblock |
| Nov 30 | Project Presentations |  |  |  |
| Dec 2 | Project Presentations |  |  |  |

**Statement on Academic Conduct and Support Systems**

**Academic Conduct**

Plagiarism – presenting someone else’s ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences.  Please familiarize yourself with the discussion of plagiarism in *SCampus* in Section 11, *Behavior Violating University Standards* [https://scampus.usc.edu/1100-behavior-violating-university-standards-and-appropriate-sanctions](https://scampus.usc.edu/1100-behavior-violating-university-standards-and-appropriate-sanctions/).  Other forms of academic dishonesty are equally unacceptable.  See additional information in *SCampus* and university policies on scientific misconduct, [http://policy.usc.edu/scientific-misconduct](http://policy.usc.edu/scientific-misconduct/).

Discrimination, sexual assault, and harassment are not tolerated by the university.  You are encouraged to report any incidents to the *Office of Equity and Diversity* <http://equity.usc.edu> or to the *Department of Public Safety* <http://capsnet.usc.edu/department/department-public-safety/online-forms/contact-us>.  This is important for the safety of the whole USC community.  Another member of the university community – such as a friend, classmate, advisor, or faculty member – can help initiate the report, or can initiate the report on behalf of another person.  *The Center for Women and Men* http://www.usc.edu/student-affairs/cwm/ provides 24/7 confidential support, and the sexual assault resource center webpage <http://sarc.usc.edu> describes reporting options and other resources.

## **Support Systems**

A number of USC’s schools provide support for students who need help with scholarly writing.  Check with your advisor or program staff to find out more.  Students whose primary language is not English should check with the *American Language Institute* <http://dornsife.usc.edu/ali>, which sponsors courses and workshops specifically for international graduate students.  *The Office of Disability Services and Programs* <http://sait.usc.edu/academicsupport/centerprograms/dsp/home_index.html> provides certification for students with disabilities and helps arrange the relevant accommodations.  If an officially  declared emergency makes travel to campus infeasible, *USC Emergency Information* [*http://emergency.usc.edu*](http://emergency.usc.edu)will provide safety and other updates, including ways in which instruction will be continued by means of blackboard, teleconferencing, and other technology.