USC Viterbi School of Engineering

INF 558/CSCI 563: Building Knowledge Graphs

Units: 4
Term—Day—Time:
Fall 2017 – Tuesday – 3:30-6:50pm

Location: VKC 156

Instructor: Wensheng Wu
Office: GER 204
Office Hours: 9-9:45 Mon & Wed
Contact Info: wenshenw@usc.edu

Teaching Assistant: Minh Pham
Office: TBD
Office Hours: TBD
Contact Info: minhpham@usc.edu

Catalogue Course Description

Foundations, techniques, and algorithms for building knowledge graphs and doing so at scale. Topics include information extraction, data alignment, entity linking, and the Semantic Web.

Expanded Course Description

This course focuses on foundations, techniques, and algorithms for building knowledge graphs. Students will learn the theory and applications of the techniques needed to build and query massive knowledge graphs. Topics include crawling web sites, wrapper learning, information extraction, source alignment, string matching, entity linking, graph databases, querying knowledge graphs, data cleaning, Semantic Web, linked data, graph analytics, and intellectual property. 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 algorithms and techniques for crawling web sites, structured data extraction, and information extraction from unstructured text.
- Understand the theory and techniques for cleaning, aligning, matching, and linking data.
- Understand the foundations and techniques of the Semantic Web, including RDF, ontologies, SPARQL, and linked data.
- Understand how to work with graph databases, including how to load massive datasets into such databases, how to organize the data for efficient access, and how to efficiently query the contents.
- Understand the entire process of how to design, construct, and query a knowledge graph to solve real-world problems.
- Understand how to apply the big data tools and infrastructure (e.g., Spark) to build and query knowledge graphs.

Required Preparation:

Prerequisite(s): INF 551 or CSCI 585
INF 552 or CSCI 567
Recommended Background: Experience programming in Python

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, home works will be posted online. The class project is a significant aspect of this course and at the end of the semester students will present their projects in class.

Required Readings and Supplementary Materials

Required Textbook: none

We use a set of technical papers and book chapters that are all available online. 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 10 weeks of class. The assignments must be done individually. The homework assignments are expected to take 8-10 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 their project proposals in class, conduct the project, and then create a video demonstration of the work and present the project in class.

Project Timeline:

- Week 2-4: Project proposals presented in class (team members, topic)
- Week 10: Project status update due (1 page status report)
- Week 16: Project presentation in class (short talk and video demonstration)

Project description: Each project team will build a knowledge graph for a topic of their choice. The knowledge graph must combine data from at least 3 different sources and at least 2 of those sites must be from online web sites. The best projects build on many of the topics covered in the class. The homework has been designed so that you can work on your projects in the process of doing your homework.

An example project would be to build a knowledge graph of used bicycles that could be purchased near the USC campus. This project would combine data from used sources, such as Craig's List, new bike sources such as BikeNashbar, and bicycle review sites, such as bicycling.com. The project would collect the data from each of these sources using wrapper techniques, extract the details of the used bicycle ads from Craig's List using information extraction techniques, align the data across these various sources to a domain ontology, link the entities across sources to combine the used data with the reviews from bicycling.com and prices from BikeNashbar, store all of the data into a search engine, e.g., elasticsearch, and then build a simple user interface to show the results by executing queries on the search engine.

Grading breakdown of the course project:

- Proposal: 10%
- Project video: 30%
- Presentation: 30%
- Overall project: 30%

Grading Breakdown

Quizzes: There will be weekly quizzes based on the material from the week before.
Homework: There will be weekly homework based on the topics of the class each week.
Exams: There will be a midterm exam, covering the first half of the course, and a final exam, covering the second half.
Class Project: Each student will do a group class project based on the topics covered in the class. Students will propose their own project, do the research and build a proof-of-concept, create a video demonstration of the proof-of-concept, and present the project in class.
Paper presentation: Students will be required to present assigned research papers on the course subjects.

Grading Schema:

Homework	25%
Class Project	20%
Paper presentation	5%
Quizzes	20%
Midterm	15%
Final	15%
Total	100%

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

[93, 100] = A	[73, 76) = C
[90, 93) = A-	[70, 73) = C-
[87, 90) = B+	[67, 70) = D+
[83, 87) = B	[63, 67) = D
[80, 83) = B-	[60, 63) = D-
[77, 80) = C+	Below 60 is an F

Note that [90, 93) means that your score is greater than or equal to 90 but less than 93. Note that every point in your coursework counts. We will strictly follow the above cut-off and NO roundup will be performed. Note that grades are NOT negotiable!

Assignment Submission Policy

Homework assignments are due at 11:59pm on the due date and should be submitted in Blackboard. Late homework will be deducted 10% of its points for every 24 hours that it is late. No credit will be given after 72 hours of its due time.

Course Schedule: A Weekly Breakdown (subject to change when course progresses)

Week	Topics/Daily Activities	Readings	Quizzes & Homeworks
1 (1/9)	**Course Introduction & Use Case** **Crawling**	• Pedro Szekely, et al. Building and using a knowledge graph to combat human trafficking. In Proceedings of the 14th International Semantic Web Conference (ISWC 2015), 2015. http://iswc2015.semanticweb.org/sites/iswc2015.semanticweb.org/files/93670175.pdf • "Chapter 20: Web crawling and indexes" of book "Introduction to Information Retrieval" by Chris Manning, et. al. 2008 (https://nlp.stanford.edu/IR-	Homework 1: Project Ideas

		book/) • [W1-1] The Anatomy of a Large Scale Hypertextual Web Search Engine Sergey Brin and Lawrence Page, Seventh International World Wide Web Conference, 1998. http://ilpubs.stanford.edu:8090/361/1/1998-8.pdf • [W1-2] Searching the Web Arvind Arasu et al., ACM Transactions on Internet Technology, 2001 http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.25.5183	
2 (1/16)	**Wrapper Generation and Learning**	• 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 • 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 • 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 • [W2-1] 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.	Quiz 1 Homework 2: Crawling a website
3 (1/23)	**Information Extraction**	• Information Extraction. S Sarawagi, Foundations and Trends in Databases, 2008. http://www.nowpublishers.com/article/DownloadSummary/DBS-003 • [W3-1] 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 • [W3-2] 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	Quiz 2 Homework 3: Building wrappers
4 (1/30)	**Data Cleaning**	• Data cleaning: problems and current approaches. Erhard Rahm, Hong Hai Do. IEEE Data Engineering Bulletin, 2000.	Quiz 3 Homework 4:

		http://www.academia.edu/download/41858217/A00DEC-CD.pdf#page=5 • Potter's Wheel: An Interactive Data Cleaning System. Vijayshankar Raman and Joseph M. Hellerstein. In Proc. VLDB 2001. http://control.cs.berkeley.edu/pwheel-vldb.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. • [W4-1] 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 • [W4-2] 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	Information Extraction
5 (2/6)	**Graph Representation RDF, RDF Schema, JSON-LD**	• 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. • 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/	Quiz 4 Homework 5: Data Cleaning
6 (2/13)	**Ontologies / RDF Mapping Tools**	• Krtzsch Markus, Simancik Frantisek, and Horrocks Ian. A description logic primer. 2012. http://arxiv.org/pdf/1201.4089.pdf. • R2rml: Rdb to rdf mapping language. http://www.w3.org/TR/r2rml/	Quiz 5 Homework 6: RDF
7 (2/20)	**Midterm**		
8 (2/27)	**String Matching**	• AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. Principles of Data Integration, chapters 4. Morgan Kaufmann, 2012. http://www.sciencedirect.com/science/book/9780124160446	Quiz 6 Homework 7: Ontologies/R2RML

9 (3/6)	**Entity Linking (Data matching)**	• 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	Quiz 7 Homework 8: String Matching
10 (3/13)	**Spring recess**		
11 (3/20)	**Semantic Labeling**	• Ramnandan, S.; Mittal, A.; Knoblock, C. A.; and Szekely, P. Assigning Semantic Labels to Data Sources. In *Proceedings of the 12th ESWC*, 2015. http://usc-isi-i2.github.io/papers/ramnandan15-eswc.pdf • Venetis, P., Halevy, A., Madhavan, J., Pasca, M., Shen, W., Wu, F., Miao, G., Wu, C.: Recovering semantics of tables on the web. Proc. VLDB Endow. 4(9), 528–538 (2011) http://www.vldb.org/pvldb/vol4/p528-venetis.pdf • Syed, Z., Finin, T., Mulwad, V., Joshi, A.: Exploiting a web of semantic data for interpreting tables. In: Proceedings of the Second Web Science Conference (2010) http://ebiquity.umbc.edu/_file_directory_/papers/478.pdf	Quiz 8 Homework 9: Entity Linking
12 (3/27)	**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 • José 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 • Craig A. Knoblock and Pedro Szekely. Exploiting semantics for big data integration. AI Magazine, 2015. http://usc-isi-i2.github.io/papers/knoblock15-aimagazine.pdf	Quiz 9 Homework 10: Source Modeling
13 (4/3)	**Querying Knowledge Graphs / ElasticSearch / SPARQL**	• Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Stenzhorn, H., & Tummarello, G. (2008). Sindice.com: a document-oriented lookup index for open linked data. *International Journal of Metadata, Semantics and Ontologies*, *3*(1), 37-52. http://wtlab.um.ac.ir/images/e-	Quiz 10 Homework 11: SPARQL/ElasticSearch

		library/linked_data/other/Sindice.pdf • Freitas, A., Oliveira, J. G., Curry, E., O'Riain, S., & da Silva, J. C. P. (2011, June). Treo: combining entity-search, spreading activation and semantic relatedness for querying linked data. In *Proc. of 1st Workshop on Question Answering over Linked Data (QALD-1) at the 8th Extended Semantic Web Conference (ESWC 2011)*. https://www.deri.ie/sites/default/files/publications/freitas_qald_2011_0.pdf • 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	
14 (4/10)	**Linked Data/ DBpedia**	• Tom Heath and Christian Bizer. Linked Data: Evolving the Web into a Global Data Space Synthesis Lectures on the Semantic Web: Theory and Technology • Chapters 1 to and including section 4.2. https://west.uni-koblenz.de/files/ws1213/seminar-web-science/linked-data.pdf	Quiz 11 Homework 12: Linked Data
15 (4/17)	**Graph Analytics** **Intellectual Property**	• Rajaraman, J. Leskovec and J. D. Ullman, *Mining of Massive Datasets*, Cambridge University Press, 2012. http://infolab.stanford.edu/~ullman/mmds/ch10.pdf • 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. • 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.	Quiz 12
16 (4/24)	**Student Presentations**		Quiz 13
FINAL	**Final Exam** **Same classroom**	**Tuesday, May 8, 2-4pm**	

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. 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.

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* will provide safety and other updates, including ways in which instruction will be continued by means of blackboard, teleconferencing, and other technology.