USC Viterbi School of Engineering

DSCI 558/CSCI 563: Building Knowledge Graphs

Units: 4
Term—Day—Time:
Spring 2021 – Monday/Wednesday – 4pm - 5:50pm

Location: Online (see Blackboard for Zoom link)

Instructor: Craig Knoblock
Office: Zoom
Office Hours: After each class on Zoom, or by appointment
Contact Info: knoblock@isi.edu, 310-448-8786.
For Appointments contact: Karen Rawlins (krawlins@isi.edu)

Instructor: Jay Pujara
Office: Zoom
Office Hours: After each class on Zoom, or by appointment
Contact Info: jpujara@usc.edu, 310-448-8482.

Teaching Assistant: Minh Pham
Office: https://usc.zoom.us/j/7348692989
Office Hours: Thursdays, 2:00-3:00pm or by appointment
Contact Info: minhpham@usc.edu.

Grader: TBD
Contact Info: TBD

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 websites, 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): DSCI 551 or CSCI 585
DSCI 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. The first 4-5 weeks of the course are structured as a quickstart to provide a shallow primer on the end-to-end process of knowledge graph construction, followed by deeper presentations and more technical material for the remainder of the course. 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, and homeworks will be posted online:
(https://drive.google.com/drive/folders/1PKBpjKYtNJITwvDESRuxqCat56sMtvgw?usp=sharing).
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 10-11 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 0-10 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 write a project proposal, present the proposal in class, conduct the project, and then create a video demonstration of the work, present the project in class and write a final report of their work.

Project Timeline:

- Week 7: Project proposals presented in class (team members, topic)
- Week 9: Project status (i) update due (online form status report)
- Week 12: Project status (ii) update due (online form status report)
- Week 15: 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 websites. The best projects build on many of the topics covered in the class. The homeworks have 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 graph database such as elasticsearch, and then build a simple user interface to show the results by executing queries against the graph database.

Grading breakdown of the course project:

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

Grading Breakdown

Quizzes: There will be weekly quizzes at the start of class based on the material from the week before. The lowest three quiz grades will be dropped. Missed quizzes will receive a zero grade, and there will be no make-up quizzes for any reason.

Midterm: There is no midterm exam for this class.

Homework: There will be weekly homework based on the topics of the class each week.

Final Exam: There is a final exam at the end of the semester covering all of the material covered in the class. The final exam will be on the date designated by USC in https://classes.usc.edu/term-20211/finals/

Class Project: Each student will do a group class project based on the topics covered in the class. Students will propose their own project, write a 1-page proposal, present the proposal in class, do the research, build a proof-of-concept, create a video demonstration of the proof-of-concept, write a final report and present the project in class.

Grading Schema:

Quizzes	20%
Homework	25%
Final	15%
Class Project	40%
Total	100%

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

94 - 100 = A	74 – 76.9 = C
90 – 93.9 = A-	70 – 73.9 = C-
87 – 89.9 = B+	67 – 69.9 = D+
84 – 86.9 = B	64 – 66.9 = D
80 – 83.9 = B-	60 – 63.9 = D-
77 – 79.9 =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 lose 20% of the possible points for the assignment. After one week, the assignment cannot be submitted.

Course Schedule: A Weekly Breakdown

	Topics/Daily Activities	Readings	Quizzes & Homeworks	Instructor
1/20	Intro	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.		Pujara Knoblock
1/25	QS: Crawling the Web	The Anatomy of a Large Scale Hypertextual Web Search Engine Sergey Brin and Lawrence Page, Seventh International World Wide Web Conference, 1998.	Homework 1: Crawling	Knoblock
1/27	QS: Information Extraction	D. C. Wimalasuriya and D. Dou. Ontology-based Information Extraction: An Introduction and a Survey of Current Approaches. J. Information Science, 36(3), 2010.	Quiz 1 Homework 2: IE	Pujara
2/1	QS: Knowledge Representation	A. Barr and J. Davidson. Representation of Knowledge, in Handbook of AI, volume 1, Chapter 3A-B, pages 141–160.	Quiz 2	Knoblock
2/3	QS: Entity Resolution	W. E. Winkler. The state of record linkage and current research problems. In Statistical Research Division, US Census Bureau. Citeseer, 1999.	Quiz 3 Homework 3: KR & ER	Pujara
2/8	QS: Large KGs	M. Farber, B. Ell, A. Rettinger, F. Bartscherer. Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. The Semantic Web, 2016	Quiz 4	Knoblock
2/10	QS: Queries and KGs	SPARQL 1.1 Query Language.	Quiz 5 Homework 4: SPARQL	Pujara
2/15	No Class President's Day			
2/17	Large KGs Entity Linking		Quiz 6	Knoblock
2/22	String Similarity	W. Cohen, P. Ravikumar, and S. Fienberg. A Comparison of String Distance Metrics for Name-matching Tasks. Conference on Information Integration on the Web, 2003.	Quiz 7	Knoblock
2/24	Information Extraction	Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: rapid training data creation with weak supervision. *Proc. VLDB Endow.* 11, 3, 269-282.	Quiz 8	Pujara
3/1	Project Proposals			Pujara Knoblock

3/3	ER & Probabilistic Soft Logic (PSL)	J. Pujara and L. Getoor. Generic Statistical Relational Entity Resolution in Knowledge Graphs. StaRAI 2016.	Homework 5: IE II	Pujara
3/8	Ontologies and RDF	Frank Manola and Eric Miller. Rdf primer. Technical report, W3C, February 2004.	Quiz 9	Knoblock
3/10	Structured Data	Limaye, G., Sarawagi, S., Chakrabarti, S.: Annotating and Searching Web Tables using Entities, Types and Relationships. Proc. VLDB Endow. 3(1-2), 1338-1347	Quiz 10 Homework 6: PSL & OWL	Pujara
3/15	Semantic Typing / Semantic Models	Pham, M.; Alse, S.; Knoblock, C.; and Szekely, P, Semantic labeling: A domain-independent approach. In *ISWC* 2016. Taheriyan, M., Knoblock, C.A., Szekely, P. and Ambite, J.L., 2016. Learning the semantics of structured data sources. Journal of Web Semantics.	Quiz 11	Knoblock
3/17	Knowledge Graph Embeddings	Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning Entity and Relation Embeddings for Knowledge Graph Completion. In AAAI	Quiz 12	Muhao Chen
3/22	Data Cleaning & Transformation	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.	Quiz 13	Minh Pham
3/24	Blocking and Relational ER	G. Papadakis D. Skoutas, E. Thanos and T. Palpanas. Blocking and Filtering Techniques for Entity Resolution: A Survey. ACM Computing Surveys, 53(2): 1-42. 2020 J. Pujara, H. Miao, L. Getoor, and W. Cohen. Using Semantics & Statistics to Turn Data into Knowledge. AI Magazine, 36(1):65–74, 2015b	Quiz 14	Pujara
3/29	Intellectual Property	Kembrew McLeod. Intellectual property law, freedom of expression, and the web, 2003.	Quiz 15	Knoblock
3/31	Graph Analytics	A Comprehensive Guide to Graph Algorithms in Neo4J	Quiz 16	Pujara
4/5	Linked Data & Semantic Web, Scientific Data	DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia	Quiz 17	Knoblock
4/7	No Class Wellness Day			
4/12	Common Sense Knowledge		Quiz 18 Homework 7: KGEs & Blocking	Filip Ilievski
4/14	Question Answering		Quiz 19	Pujara

4/19	Special Topics: GraphDBs SciUnits LinkedGeoData		Quiz 20	Vu/ Shbita
4/21	Course Review			Pujara Knoblock
4/26	Project Presentations			Pujara Knoblock
4/28	Project Presentations			Pujara Knoblock
5/5	**Final Exam: 4:30pm-6:30pm**			

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.