USC Viterbi School
of Engineering

DSCI 558/CSCI 563: Building Knowledge Graphs

Units: 4
Term—Day—Time:
Spring 2025 – Tuesday/Thursday – 2:00pm - 3:50pm

Location: THH 102

Instructor: Jay Pujara
Office: ISI 936
Office Hours: By appointment
Contact Info: jpujara@usc.edu, 310-448-8482.
For Appointments contact: ckg-admin@isi.edu

Teaching Assistant: Bahareh Harandizadeh
Office:
Office Hours:
Contact Info:

Grader:
Contact Info:

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/1s2wXehQ6onC-KzZvGHNyhJKH8lo89GzO) -- you must be logged into a USC Google Account to view the materials. 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 8 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 five 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-20223/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:

Letter grade	Corresponding numerical point range
A	95-100
A-	90-94
B+	87-89
B	83-86
B-	80-82
C+	77-79
C	73-76
C-	70-72
D+	67-69
D	63-66
D-	60-62
F	59 and below

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
1/14	Intro	Hogan, Aidan, et al. Knowledge graphs. *ACM Computing Surveys (CSUR)* 54.4 (2021), Section 1. 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.	
1/16	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. Optional: Knowledge Graphs: Fundamentals, Techniques, and Applications, Chapter 3, Kejriwal, Knoblock, and Szekely, 2021	Quiz 1

1/21	Information Extraction	Hogan, Aidan, et al. Knowledge graphs. *ACM Computing Surveys (CSUR)* 54.4 (2021), Section 6. 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 2
1/23	Knowledge Representation	Hogan, Aidan, et al. Knowledge graphs. *ACM Computing Surveys (CSUR)* 54.4 (2021), Section 2-3. A. Barr and J. Davidson. Representation of Knowledge, in Handbook of AI, volume 1, Chapter 3A-B, pages 141–160. Frank Manola and Eric Miller. Rdf primer. Technical report, W3C, February 2004.	Quiz 3 Homework 1: Crawling / IE
1/28	Large KGs + Grounding	Heist, N., Hertling, S., Ringler, D., & Paulheim, H. (2020). Knowledge Graphs on the Web-An Overview. Shen, Wei, Jianyong Wang, and Jiawei Han. "Entity linking with a knowledge base: Issues, techniques, and solutions." *IEEE TKDE 2014*.	Quiz 4
1/30	Entity Resolution	Christophides, Vassilis, Vasilis Efthymiou, and Kostas Stefanidis. Entity resolution in the web of data. *Synthesis Lectures on the Semantic Web* 5.3 (2015): 1-122 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 5 Homework 2: KR / Entity Resolution
2/4	Structured Data and Semantic Typing	Limaye, G., Sarawagi, S., Chakrabarti, S.: Annotating and Searching Web Tables using Entities, Types and Relationships. Proc. VLDB Endow. 3(1-2), 1338-1347 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 6 Project Proposal instructions
2/6	Querying	https://www.w3.org/TR/sparql11-query/ Francis, N., Green, A., Guagliardo, P., Libkin, L., Lindaaker, T., Marsault, V., ... & Taylor, A. (2018, May). Cypher: An evolving query language for property graphs. In Proceedings	Quiz 7 Homework 3: Queries, KGs, & Structured Data

		of the 2018 International Conference on Management of Data (pp. 1433-1445).	
2/11	Representation Learning	Hogan, Aidan, et al. Knowledge graphs. *ACM Computing Surveys (CSUR)* 54.4 (2021), Section 5.2-5.3. Q. Wang, Z. Mao, B. Wang and L. Guo, Knowledge Graph Embedding: A Survey of Approaches and Applications. IEEE TKDE. 2020	Quiz 8
2/13	Neo4J and Knowledge Graph Toolkit	Ilievski, F., Garijo, D., Chalupsky, H., Divvala, N. T., Yao, Y., Rogers, C., ... & Szekely, P. (2020, November). KGTK: a toolkit for large knowledge graph manipulation and analysis. In International Semantic Web Conference (pp. 278-293). Springer, Cham. https://usc-isi-i2.github.io/kgtk-tutorial-iswc-2021/	Quiz 9
2/18	Knowledge Graph Quality	Hogan, Aidan, et al. Knowledge graphs. *ACM Computing Surveys (CSUR)* 54.4 (2021), Section 7. M. Farber, B. Ell, A. Rettinger, F. Bartscherer. Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. The Semantic Web, 2016 Shenoy, K., Ilievski, F., Garijo, D., Schwabe, D., & Szekely, P. (2022). A Study of the Quality of Wikidata. Journal of Web Semantics, 72, 100679.	Quiz 10
2/20	Project Proposals		Homework 4: KGTK and Embeddings
2/25	Probabilistic Soft Logic & Collective Entity Resolution	J. Pujara, H. Miao, L. Getoor, and W. Cohen. Using Semantics & Statistics to Turn Data into Knowledge. AI Magazine, 36(1):65–74, 2015b J. Pujara and L. Getoor. Generic Statistical Relational Entity Resolution in Knowledge Graphs. StaRAI 2016.	
2/27	Knowledge Graph Application Domains: Commonsense Reasoning and Explainability	Ilievski, F., Oltramari, A., Ma, K., Zhang, B., McGuinness, D. L., & Szekely, P. (2021). Dimensions of commonsense knowledge. *Knowledge-Based Systems*, *229*, 107347. Gilpin, L. H., & Kagal, L. A Self-Monitoring Framework for Opaque Machines.	Quiz 11

3/4	Advanced Information Extraction Techniques	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	Quiz 12
3/6	Advanced Knowledge Representation: ontologies and OWL	McGuinness, D. L., & Van Harmelen, F. (2004). OWL web ontology language overview. W3C recommendation, 10(10), 2004. Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology.	Quiz 13 Homework 5: OWL and PSL
3/11	Knowledge Graph Application Domains: Scientific Research	Building Spatio-Temporal Knowledge Graphs from Vectorized Topographic Historical Maps. Shbita, B.; Knoblock, C. A; Duan, W.; Chiang, Y.; Uhl, J. H; and Leyk, S. Semantic Web, (Preprint): 1–23. 2022.	Quiz 14
3/13	Advanced Representation Learning: Augmentation and Graph Neural Networks	Wang, J., Ilievski, F., Szekely, P., & Yao, K. T. (2022). Augmenting Knowledge Graphs for Better Link Prediction. arXiv preprint arXiv:2203.13965. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., ... & Sun, M. (2020). Graph neural networks: A review of methods and applications. AI Open, 1, 57-81.	Quiz 15
3/18 3/20	No class - Spring Recess		
3/25	Knowledge Graphs & LLMs		Quiz 16 Project Status Report 1
3/27	Neurosymbolic Reasoning	Wang, P., Peng, N., Ilievski, F., Szekely, P., & Ren, X. (2020). Connecting the dots: A knowledgeable path generator for commonsense question answering. arXiv preprint arXiv:2005.00691. West, P., Bhagavatula, C., Hessel, J., Hwang, J. D., Jiang, L., Bras, R. L., ... & Choi, Y. (2021). Symbolic knowledge distillation: from general language models to commonsense models. arXiv preprint arXiv:2110.07178.	Quiz 17 Homework 6: Neurosymbolic Approaches and Representation Learning
4/1	Question Answering	High, R. (2012). The era of cognitive systems: An inside look at IBM Watson and how it works. IBM Corporation, Redbooks, 1, 16. Bakhshi, M., Nematbakhsh, M., Mohsenzadeh, M., & Rahmani, A. M. (2020).	Quiz 18

		Data-driven construction of SPARQL queries by approximate question graph alignment in question answering over knowledge graphs. Expert Systems with Applications, 146, 113205.	
4/3	Network Analytics for Knowledge Graphs	Hogan, Aidan, et al. Knowledge graphs. *ACM Computing Surveys (CSUR)* 54.4 (2021), Section 5.1. A Comprehensive Guide to Graph Algorithms in Neo4J Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., ... & Zhang, W. (2014, August). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 601-610).	Quiz 19
4/8	Bias in Knowledge Graphs	Mehrabi, N., Zhou, P., Morstatter, F., Pujara, J., Ren, X., & Galstyan, A. (2021). Lawyers are dishonest? quantifying representational harms in commonsense knowledge resources. arXiv preprint arXiv:2103.11320. Shaik, Z., Ilievski, F., & Morstatter, F. (2021). Analyzing Race and Country of Citizenship Bias in Wikidata. arXiv preprint arXiv:2108.05412.	Quiz 20 Homework 7: Network Analytics
4/10	Temporal Knowledge Graphs	Galkin, M., Trivedi, P., Maheshwari, G., Usbeck, R., & Lehmann, J. (2020). Message passing for hyper-relational knowledge graphs. arXiv preprint arXiv:2009.10847.	Quiz 21 Project Status Report 2
4/15	Intellectual Property	Kembrew McLeod. Intellectual property law, freedom of expression, and the web, 2003.	Quiz 22
4/17	Knowledge Graph Application Domains: Geospatial KGs		Quiz 23
4/22	Knowledge Graph Application Domains: Health		Quiz 24
4/24	Project Presentations		
4/29	Project Presentations		
5/1	Course Review		

5/8	**Final Exam: 2-4p**

Statement on Academic Conduct and Support Systems

Academic Integrity

The University of Southern California is foremost a learning community committed to fostering successful scholars and researchers dedicated to the pursuit of knowledge and the transmission of ideas. Academic misconduct is in contrast to the university's mission to educate students through a broad array of first-rank academic, professional, and extracurricular programs and includes any act of dishonesty in the submission of academic work (either in draft or final form).

This course will follow the expectations for academic integrity as stated in the USC Student Handbook. All students are expected to submit assignments that are original work and prepared specifically for the course/section in this academic term. You may not submit work written by others or "recycle" work prepared for other courses without obtaining written permission from the instructor(s). Students suspected of engaging in academic misconduct will be reported to the Office of Academic Integrity.

Other violations of academic misconduct include, but are not limited to, cheating, plagiarism, fabrication (e.g., falsifying data), knowingly assisting others in acts of academic dishonesty, and any act that gains or is intended to gain an unfair academic advantage.

Academic dishonesty has a far-reaching impact and is considered a serious offense against the university. Violations will result in a grade penalty, such as a failing grade on the assignment or in the course, and disciplinary action from the university itself, such as suspension or even expulsion.

For more information about academic integrity see the student handbook or the Office of Academic Integrity's website, and university policies on Research and Scholarship Misconduct.

Please ask your instructor if you are unsure what constitutes unauthorized assistance on an exam or assignment or what information requires citation and/or attribution.

Use of Generative AI

This course aims to develop creative, analytical, and critical thinking skills. Therefore, assignments should be primarily prepared by the student working individually or in groups. Students may not have another person or entity complete any substantive portion of the assignment. Developing strong competencies in these areas will prepare you for a competitive workplace. If you choose to use artificial intelligence (AI)-powered programs in the process of completing assignments, you must clearly disclose such usage. To adhere to our university values, you must cite any AI-generated material (e.g., text, images, etc.) included or referenced in your work and provide the prompts used to generate the content. Using an AI tool to generate content without proper attribution will be treated as plagiarism and reported to the Office of Academic Integrity. You should also be aware that

AI text generation tools may present incorrect information, biased responses, and incomplete analyses; thus they are not prepared to produce text that meets the standards of this course.

Course Content Distribution and Synchronous Session Recordings Policies

USC has policies that prohibit recording and distribution of any synchronous and asynchronous course content outside of the learning environment.

Recording a university class without the express permission of the instructor and announcement to the class, or unless conducted pursuant to an Office of Student Accessibility Services (OSAS) accommodation. Recording can inhibit free discussion in the future, and thus infringe on the academic freedom of other students as well as the instructor. (Living our Unifying Values: The USC Student Handbook, page 13).

Distribution or use of notes, recordings, exams, or other intellectual property, based on university classes or lectures without the express permission of the instructor for purposes other than individual or group study. This includes but is not limited to providing materials for distribution by services publishing course materials. This restriction on unauthorized use also applies to all information, which had been distributed to students or in any way had been displayed for use in relation to the class, whether obtained in class, via email, on the internet, or via any other media. Distributing course material without the instructor's permission will be presumed to be an intentional act to facilitate or enable academic dishonestly and is strictly prohibited. (Living our Unifying Values: The USC Student Handbook, page 13).

Statement on University Academic and Support Systems

Students and Disability Accommodations:

USC welcomes students with disabilities into all of the University's educational programs. The Office of Student Accessibility Services (OSAS) is responsible for the determination of appropriate accommodations for students who encounter disability-related barriers. Once a student has completed the OSAS process (registration, initial appointment, and submitted documentation) and accommodations are determined to be reasonable and appropriate, a Letter of Accommodation (LOA) will be available to generate for each course. The LOA must be given to each course instructor by the student and followed up with a discussion. This should be done as early in the semester as possible as accommodations are not retroactive. More information can be found at osas.usc.edu. You may contact OSAS at (213) 740-0776 or via email at osasfrontdesk@usc.edu.

Student Financial Aid and Satisfactory Academic Progress:

To be eligible for certain kinds of financial aid, students are required to maintain Satisfactory Academic Progress (SAP) toward their degree objectives. Visit the Financial Aid Office webpage for undergraduate- and graduate-level SAP eligibility requirements and the appeals process.

Support Systems:

Counseling and Mental Health - (213) 740-9355 – 24/7 on call
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.
988 Suicide and Crisis Lifeline - 988 for both calls and text messages – 24/7 on call
The 988 Suicide and Crisis Lifeline (formerly known as the National Suicide Prevention Lifeline) provides free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week, across the United States. The Lifeline consists of a national network of over 200 local crisis centers, combining custom local care and resources with national standards and best practices. The new, shorter phone number makes it easier for people to remember and access mental health crisis services (though the previous 1 (800) 273-8255 number will continue to function indefinitely) and represents a continued commitment to those in crisis.
Relationship and Sexual Violence Prevention Services (RSVP) - (213) 740-9355(WELL) – 24/7 on call
Free and confidential therapy services, workshops, and training for situations related to gender- and power-based harm (including sexual assault, intimate partner violence, and stalking).

Office for Equity, Equal Opportunity, and Title IX (EEO-TIX) - (213) 740-5086
Information about how to get help or help someone affected by harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants.

Reporting Incidents of Bias or Harassment - (213) 740-2500
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office for Equity, Equal Opportunity, and Title for appropriate investigation, supportive measures, and response.

The Office of Student Accessibility Services (OSAS) - (213) 740-0776
OSAS ensures equal access for students with disabilities through providing academic accommodations and auxiliary aids in accordance with federal laws and university policy.

USC Campus Support and Intervention - (213) 740-0411
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity, Equity and Inclusion - (213) 740-2101
Information on events, programs and training, the Provost's Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety - UPC: (213) 740-6000, HSC: (323) 442-1200 – 24/7 on call
Non-emergency assistance or information.

Office of the Ombuds - (213) 821-9556 (UPC) / (323-442-0382 (HSC)
A safe and confidential place to share your USC-related issues with a University Ombuds who will work with you to explore options or paths to manage your concern.

Occupational Therapy Faculty Practice - (323) 442-2850 or otfp@med.usc.edu
Confidential Lifestyle Redesign services for USC students to support health promoting habits and routines that enhance quality of life and academic performance.