DSCI 511: Data Acquisition and Pre-Processing

Course Syllabus

Credits: 3 hour lecture [3 credits]
Spring 2019
Short Title: Acquisition and Pre-Processing

General Information

Course Coordinator(s): Jake Ryland Williams

Instructor Contact Information: Jake.Williams@drexel.edu

Office Hours; Location: Thursdays, 5–6pm; 3675 Market Street, Room 1113

Student Learning Information

Course Description

Introduces the breadth of data science through a project lifecycle perspective. Covers early-stage data-life cycle activities in depth for the development and dissemination data sets. Provides technical experience with data harvesting, acquisition, pre-processing, and curation. Concludes with an open-ended term project where students explore data availability, scale, variability, and reliability.

College/Department: College of Computing & Informatics
Repeat Status: Not repeatable for credit
Restrictions: None
Prerequisites: None

Course Purpose within a Program of Study

This course provides a high-level view of what data is, where it comes from, and how it is used to render insight and support technical products. After introducing the breadth of steps that data scientists take throughout the lifecycle of a project, hands-on and in-depth experience is provided with several early-stage steps that interact closely, and often require iteration before later steps can be approached.

This course is a core course in the Data Science Masters program.

Statement of Expected Learning

The course objectives are to:

- obtain an overview of what data is, where it comes from, and what data science entails;
- understand of the range of activities essential to a data science project's lifecycle;
- apply technical methods in data collection, construction, and curation; and
- execute a data collection/curation project exploring lifecycle stage interdependence and iteration.

As learning outcomes, students completing this course should be able to use their understanding of project lifecycles, data sources and availability, acquisition and harvesting, pre-processing, and data management to curate data sets of high value that are in alignment with down-stream project goals.

Course Materials

Required and Recommended Texts, Readings, and Resources

Note: all text readings are supplemental to the course lecture notes and will be assigned on a weekly basis. Python Data Science Handbook is available free of charge on Github:

- https://jakevdp.github.io/PythonDataScienceHandbook/

and The Python Data Science Handbook and Data Science from Scratch are available to Drexel students through the University Libraries. Specific text information is as follows:

- Data Science from Scratch (DSFS). ISBN: 978-1491901427, O'Riley, April 2015
- Python Data Science Handbook (PDSH). ISBN: 978-1491912058, O'Riley, November 2016
- The Data Science Handbook (TDSH). ISBN:9781119092940, John Wiley & Sons, 2017

Required and Supplemental Materials and Technologies

Note: instructions and discussion of the following materials and technologies are provided in Chapter 0 of the course lecture notes. Students are expected to have the following by the start of the first week:

- A Github account: https://www.github.com
- A command line environment with Python (version 3) installed
- The Jupyter notebooks interactive development environment

Lecture Notes

The primary course materials consist of a collection of interactive Jupyter notebooks, which may be found on the course blackboard website and on the following private Github repository:

- http://github.com/jakerylandwilliams/DSCI511/

For access to the Github repository, all students are required to sign up for an account and post user names on the course discussion board. An invite to the Github repository will follow.

The course lecture notes are broken down into the following topics, which roughly correspond to a week of content each. More information on course scheduling can be found in that section, below.

Chapter 0: System configuration and processing fundamentals
Chapter 1: Introduction, process, and getting started with data
Chapter 2: Data types and structures: different data, different challenges
Chapter 3: Established collections: databases, dumps, and APIs
Chapter 4: Pre-processing considerations: foresight for downstream needs
Chapter 5: Harvesting content from the world wide web
Chapter 6: Data integration and enrichment
Chapter 7: Building and maintaining a robust acquisition stream
Chapter 8: Establishing a database with documentation
Chapter 9: Distribution, accessibility, and data sharing

Assignments, Assessments, and Evaluations

Graded Assignments and Learning Activities

Homework: Structured, individual assignments will be distributed according to four topic areas:

1. Data science programming
2. APIs and pre-processing

3. Web scraping and data integration
4. Disseminating a data processing tool

These assignments will be composed in a modular fashion, with each module/problem worth about 35–45 points apiece. The total assignment value requirement for the term is 400 *target* points and there will be about 450 *possible* points available across all modules. Each module can be completed and submitted separately. This means there will be roughly 12–16 modules, total.

Project: One open-ended group assignment will have two phases:

1. Proposal for Data Set Construction and Potential for Use
2. Implementation of Data Set Construction with Documentation and Dissemination

Grading Matrix

Students will not receive letter grades for individual assignments. Grades are calculated as:

Project:	30% (10% Proposal, 20% Implementation)
Homework:	70% (4 x 17.5%)
Total:	100%

Grade Scale

The following scale will be used to convert points to letter grades:

Points	*Grade*	*Points*	*Grade*	*Points*	*Grade*
97-100	A+	82-86.99	B	70-71.99	C-
92-96.99	A	80-81.99	B-	67-69.99	D+
90-91.99	A-	77-79.99	C+	60-66.99	D
87-89.99	B+	72-76.99	C	0-59.99	F

Note that the instructor may revise this conversion if/when necessary.

Course Schedule

The course's schedule follows the lecture notes at roughly one week per chapter with the expectation that students will configure systems and review or work through the processing fundamentals in Chapter 0. Some chapters may extend over multiple weeks, depending the section's pace. Week 10 and the regularly scheduled final exam period (to be determined) are reserved for final project presentations. Please observe the following (tentative) schedule, and be aware that weekly may change depending on the pace of class.

- Week 1:
  - DSFS: 1–2
  - TDSH: 1–2, 3.1–3.2
  - Project: Group formation; begin Phase 1
  - Homework: Begin Assignment Group 1
- Week 2:
  - DSFS: Chapter 9 (All, except pgs. 108–114)
  - TDSH: Ch. 12

    - PDSH: Chapters 2.01–2.02
    - Project: continue Phase 1
    - Homework: continue Assignment Group 1
- Week 3:
    - DSFS: Chapter 9 (114–120)
    - Project: begin Phase 2 (Monday); complete Phase 1 (Friday)
    - Homework: begin Assin. Group 2 (Monday); complete Assn. Group 1 (Friday)
- Week 4:
    - DSFS: Chapter 10 (pgs. 127–133)
    - TDSH: Chapter 4 (All, except 4.5)
    - Project: continue Phase 2
    - Homework: continue Assignment Group 2
- Week 5:
    - DSFS: Chapter 9 (pgs. 108–114)
    - Project: continue Phase 2
    - Homework: begin Assn. Group 3 (Monday); complete Assn. Group 2 (Friday)
- Week 6:
    - PDSH: 3.01–3.04, 3.06–3.08
    - Project: continue Phase 2
    - Homework: continue Assignment Group 3
- Week 7:
    - Project: continue Phase 2
    - Homework: begin Assn. Group 4 (Monday); complete Assn. Group 3 (Friday)
- Week 8:
    - DSFS: Chapter 23 (Supplementary)
    - Project: continue Phase 2
    - Homework: continue Assignment Group 4
- Week 9:
    - DSFS: Chapter 25
    - Project: continue Phase 2
    - Homework: complete Assignment Group 4
- Week 10:
    - DSFS: Chapter 24 (Supplementary)
    - Project: complete Phase 2; Project presentations
- Week 11:
    - Project: Project presentations

Academic Policies

This course follows university, college, and department policies, including but not limited to:

- Academic Honesty: http://www.drexel.edu/provost/policies/academic_dishonesty.asp
- Student Life Honesty Policy from Judicial Affairs: http://www.drexel.edu/provost/policies/academic-integrity
- Students with Disability Statement: http://drexel.edu/oed/disabilityResources/faculty/SyllabusStatement/
- Course Drop Policy: http://www.drexel.edu/provost/policies/course_drop.asp
- Department Academic Integrity Policy: http://drexel.edu/cs/academics/undergrad/policies/academic-integrity/
- Drexel Student Learning Priorities: http://drexel.edu/provost/assessment/outcomes/dslp/
- Office of Disability Resources: http://www.drexel.edu/ods/student_reg.html

The instructor(s) may, at his/her/their discretion, change any part of the course before or during the term, including assignments, grade breakdowns, due dates, and schedule. Such changes will be communicated to students via the course web site. This web site should be checked regularly and frequently for such changes and announcements.

Students requesting accommodations due to a disability at Drexel University need to request a current Accommodations Verification Letter (AVL) in the ClockWork database before accommodations can be made. These requests are received by Disability Resources (DR), who then issues the AVL to the appropriate contacts. For additional information, visit the DR website at drexel.edu/oed/disabilityResources/overview/, or contact DR for more information by phone at 215.895.1401, or by email at disability@drexel.edu.