Web Data Science

INFO 4871; Fall 2024

Tuesdays and Thursdays; 9:30 am – 10:45 am
Classroom: Eaton Humanities 135

Brian Keegan, Ph.D.
Assistant Professor, Information Science
E-mail: brian.keegan@colorado.edu
Office: INFO 129
Office hours: Fridays, 1:30 pm – 2:30 pm

Course Description

The internet makes many kinds of information easy to access. The ability to retrieve, parse, and analyze this information is a valuable skill for data scientists. This course will provide an overview of computational tools and practices for transforming web documents and APIs into data for common research designs.

Learning objectives

- Understand the legal and ethical contours of web data access
- Navigate and parse common web data formats like XML and JSON for data
- Retrieve and automate data extraction from HTML and PDF documents
- Access popular APIs to collect data for common research designs

Course Design

Class will meet twice per week on Tuesdays and Thursdays from 9:30 am to 10:45 am in Eaton Humanities 135. The format of each class will vary between lectures, exercises, discussions, and presentations. Student performance will be evaluated through a combination of Exercises, Assignments, and a Final Project (see *Evaluation* below). There is no final exam.

The class is split up into three modules: (1) *Fundamentals* introduces ethical and legal considerations of web data access, common file formats, and fundamental protocols; (2) *Documents* covers methods and tools for retrieving data from documents like web pages and PDFs; and (3) *APIs* provides an overview of tools for accessing popular web APIs and corresponding research designs. See Table 1 for more details.

Each week will cover a new topic. The Tuesday class will consist of a lecture and computational notebook introducing and implementing these concepts. The Thursday class will provide time for students to work on and share the results the notebook exercises for their Weekly Assignments.

Prerequisites

There are no formal pre-requisites for the class, but prior programming experience in Python is strongly recommended. If you have questions about these prerequisites, please email the instructor.

Course Website and Materials

There is no textbook required for class, but there will be required readings and assignments which will be made available and submitted through Canvas:

https://canvas.colorado.edu/courses/109074

This PDF version of the syllabus is a work-in-progress and will be revised as we proceed through the semester. Any revised requirements will be posted as announcements and updated course schedule to Canvas. The instructor reserves the right to make changes to the course's schedule, evaluation criteria, policies, *etc.* through announcements in class and on Canvas. Students should email the instructor if there are any discrepancies or questions.

Computing

Students will use programming languages for data analysis and visualization. Jupyter notebooks written in Python 3 will be used for all in-class examples and assignments. The Anaconda distribution of Python 3.10 (or above) is *strongly* recommended to provide all of these programs and other libraries. We will be using the Matplotlib and Seaborn libraries for data visualization. Lectures will include exercises and presentations with the expectation that students participate with their own laptop computers. If students cannot bring a laptop to class, they should email the instructor to work out an alternative arrangement.

Evaluation

Students will be evaluated through three mechanisms. The class has no midterm or final exam.

Attendance (15%). Attendance in class is required. We will be covering technical methods that are cumulative and require sustained effort. If there are personal, professional, medical, or other circumstances that will prevent your attendance, please notify me via e-mail and we can develop an accommodation plan together. There will be no opportunities to make up missed attendance. If you need to miss multiple classes, please make an effort to come to office hours so we can check in about course material and progress.

Module Assignments (60%). Module Assignments are intended to (1) develop students' confidence with retrieving, reshaping, and analyzing web data and (2) engaging with scholarship using web data. The format and evaluation criteria of each Module Assignment will vary by module and graduate versus undergraduate level. Each Module Assignment is worth 20% of the final grade (60% cumulative across the first three modules) and will be due at 11:59 p.m. on the Friday of the subsequent week. In the absence of an approved excuse, late submissions will be docked 1% of their maxmimum value for every day elapsed since the deadline.

Final Project (25%). The Final Project is intended to be a portfolio piece highlighting a student's skills matching the web data with a research design. Further details about the Final Project will be collaboratively developed and detailed later in the course. In the absence of an approved excuse, late submissions will be docked 1% of their maxmimum value for every day elapsed since the deadline.

Course Policies

In-Class Confidentiality

The success of this class depends on students feeling comfortable sharing questions, ideas, concerns, and confusions about assignments, work-in-progress, and their personal experiences. Students may read, comment, and run classmates' writing, code, and other class-related content for use within this class. However, students may not use, run, copy, perform, display, distribute, modify, translate, or create derivative works of another student's work outside of this class without that student's expressed written consent or formal license. Students may not create any audio, video, or other records of lectures without the instructor's permission nor may students share comments made in class attributable to another person's identity without permission.

Classroom Behavior

Both students and faculty are responsible for maintaining an appropriate learning environment in all instructional settings, whether in person, remote or online. Those who fail to adhere to such behavioral standards may be subject to discipline. Professional courtesy and sensitivity are especially important with respect to individuals and topics dealing with race, color, national origin, sex, pregnancy, age, disability, creed, religion, sexual orientation, gender identity, gender expression, veteran status, political affiliation or political philosophy. For more information, see the policies on class behavior, the Student Code of Conduct, and the Office of Institutional Equity and Compliance.

Requirements and Contingencies for COVID-19

As a matter of public health and safety, all members of the CU Boulder community and all visitors to campus must follow university, department and building requirements and all public health orders in place to reduce the risk of spreading infectious disease. CU Boulder currently requires COVID-19 vaccination and boosters for all faculty, staff and students. Students, faculty and staff must upload proof of vaccination and boosters or file for an exemption based on medical, ethical or moral grounds through the MyCUHealth portal.

If you feel ill and think you might have COVID-19, if you have tested positive for COVID-19, or if you are unvaccinated or partially vaccinated and have been in close contact with someone who has COVID-19, you should stay home and follow the further guidance of the Public Health Office (contacttracing@colorado.edu). If you are fully vaccinated and have been in close contact with someone who has COVID-19, you should self-monitor for symptoms and follow the further guidance of the Public Health Office.

Students who fail to adhere to these requirements will be asked to leave class, and students who do not leave class when asked or who refuse to comply with these requirements will be referred to Student Conduct and Conflict Resolution. For more information, see the policy on classroom behavior and the Student Code of Conduct. If you require accommodation because a disability prevents you from fulfilling these safety measures, please see the "Accommodation for Disabilities" statement.

Should a student contract any illness (COVID-19 or otherwise) that requires mandatory sequestration, intensive medical treatment, or extended convalescence and disrupts their ability to participate in class and complete assignments, the instructor will try to accommodate their condition without penalty with extensions and incompletes. This also applies if the student has a family member whose

diagnosis, treatment, and recovery will affect their ability to participate. *Please do not ghost me*: students should notify the instructor as soon as possible of events that will impact their engagement with the class so that we can triage and develop an accommodation plan rather than scrambling to solve problems near deadlines at the end of the semester.

Accommodations for Disabilities
If you qualify for accommodations because of a disability, please submit your accommodation letter from Disability Services to your faculty member in a timely manner so that your needs can be addressed. Disability Services determines accommodations based on documented disabilities in the academic environment. Information on requesting accommodations is located on the Disability Services website. Contact Disability Services at 303-492-8671 or dsinfo@colorado.edu for further assistance. If you have a temporary medical condition, see Temporary Medical Conditions on the Disability Services website.

Preferred Student Names and Pronouns
CU Boulder recognizes that students' legal information does not always align with how they identify. Students may update their preferred names and pronouns via the student portal; those preferred names and pronouns are listed on the instructor's class rosters. In the absence of such updates, the name that appears on the class roster is the student's legal name.

Honor Code
All students enrolled in a University of Colorado Boulder course are responsible for knowing and adhering to the Honor Code. Violations of the Honor Code may include, but are not limited to: plagiarism, cheating, fabrication, lying, bribery, threat, unauthorized access to academic materials, clicker fraud, submitting the same or similar work in more than one course without permission from all course instructors involved, and aiding academic dishonesty. All incidents of academic misconduct will be reported to the Honor Code (honor@colorado.edu; 303-492-5550). Students found responsible for violating the academic integrity policy will be subject to nonacademic sanctions from the Honor Code as well as academic sanctions from the faculty member. Additional information regarding the Honor Code academic integrity policy can be found on the Honor Code website.

Harassment and Discrimination
CU Boulder is committed to fostering an inclusive and welcoming learning, working, and living environment. CU's discrimination and harassment policies prohibit acts of sexual misconduct (harassment, exploitation, and assault), intimate partner violence (dating or domestic violence), stalking, or protected-class discrimination or harassment by or against members of our community. Individuals who believe they have been subject to misconduct or retaliatory actions for reporting a concern should contact the Office of Institutional Equity and Compliance (OIEC) at 303-492-2127 or email cureport@colorado.edu. Information about university policies, reporting options, and the support resources can be found on the OIEC website.

Please know that faculty and graduate instructors have a responsibility to inform OIEC when they are made aware of incidents of sexual misconduct, dating and domestic violence, stalking, discrimination, harassment and/or related retaliation, to ensure that individuals impacted receive informa-

tion about their rights, support resources, and reporting options. To learn more about reporting and support options for a variety of concerns, visit Don't Ignore It.

Religious Observance
Campus policy regarding religious observances requires that faculty make every effort to deal reasonably and fairly with all students who, because of religious obligations, have conflicts with scheduled exams, assignments or required assignments/attendance. If this applies to you, please email the instructor as soon as possible to make the appropriate accommodations.

CMCI Diversity, Equity, and Inclusion
CMCI strives to be a community whose excellence depends on diversity, equity, and inclusion. We aim to understand and challenge systems of privilege and disadvantage in higher education, such as those based on class, race, ethnicity, gender, sexuality, and dis/ability. We seek to reach across social and political divides and to make space for voices historically underrepresented in higher education and marginalized in society. In other words, diversity is not just a future reality for which we try to prepare students. It is a priority we want to put into practice here, now, and together, in order to foster places of learning where all members can thrive. Please contact the CMCI diversity team (email Lisa Flores or visit the CMCI Diversity, Inclusion, and Equity Staff page) if:

- you need support or other resources but don't know where to turn
- any aspect of your educational experience with CMCI does not reflect the commitment expressed here, or if you want to share a positive instance of this commitment in action
- you have any questions, concerns, or ideas related to diversity

Acknowledgements

This syllabus was typeset in LaTeX using Overleaf with the fbb/Bembo font and is derived from the `memoir` styles adapted by Kieran Healy and Benjamin 'Mako' Hill.

Other resources

- vanden Broucke, S. and Baesens, B. (2018). *Practical Web Scraping for Data Science: Best Practices and Examples with Python*. Apress, New York, 1st edition
- Bradshaw, P. (2012). *Scraping for Journalists (2nd Edition)*. Leanpub, 2nd edition
- Chapagain, A. (2019). *Hands-On Web Scraping with Python*. Packt Publishing, Birmingham
- Heydt, M. (2018). *Python Web Scraping Cookbook*. Packt Publishing, 1st edition
- Lawson, R. (2015). *Web Scraping with Python*. Packt Publishing Ltd
- Mitchell, R. (2018). *Web Scraping with Python: Collecting More Data from the Modern Web*. O'Reilly
- Munzert, S., Rubba, C., Meißner, P., and Nyhuis, D. (2014). *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. John Wiley & Sons
- Rogers, R. (2019). *Doing Digital Methods*. SAGE Publications
- Russell, M. and Klassen, M. (2019). *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Instagram, GitHub, and More*. O'Reilly, Boston, 3rd edition
- Salganik, M. J. (2019). *Bit by Bit: Social Research in the Digital Age*. Princeton University Press

Course Outline

This is an outline of modules and skills.

Module	Week	Dates	Topics
Fundamentals	1	Aug 27; Aug 29	Law & ethics
	2	Sep 3; Sep 5	Post-A(P)I Age
	3	Sep 10; Sep 12	XML & JSON
	4	Sep 17; Sep 19	IP & HTTP
	Module Assignment 1 due September 23		
Documents	5	Sep 24; Sep 26	Static web pages
	6	Oct 1; Oct 3	Archived web pages
	7	Oct 8; Oct 10	Dynamic web pages
	8	Oct 15; Oct 17	PDFs
	Module Assignment 2 due October 21		
APIs	9	Oct 22; Oct 24	Wikipedia
	10	Oct 29; Oct 31	Government
	11	Nov 5; Nov 7	Social
	12	Nov 12; Nov 14	AI
	13	Nov 19; Nov 21	Automation
	Module Assignment 3 due December 2		
	14	Nov 26; Nov 28	**No Class: Fall Break**
Final projects	15	Dec 3; Dec 5	Final projects
	16	Dec 10; Dec 12	

Table 1: Course outline by week.

Course Schedule

The schedule will evolve throughout the semester, so please consult the schedule online at Canvas for the most up-to-date information.

Week 1 – Fundamentals: Law and ethics

Wednesday, August 23; Friday, August 25

Administrivia; computing environment setup; law and ethics of web data retrieval.

Week 2 – Fundamentals: Post-A(P)I Age

Wednesday, August 30; Friday, September 1

Changing contexts of data access and transparency.

Week 3 – Fundamentals: XML and JSON

Wednesday, September 6; Friday, September 8

JSON and XML data formats and supporting libraries like `BeautifulSoup`.

Week 4 – Fundamentals: IP & HTTP

Wednesday, September 13; Friday, September 15

Web protocols like IP and HTTP using libraries like `urllib` and `requests`.

Week 5 – Documents: Static web pages

Wednesday, September 20; Friday, September 22

HTML elements and document parsing on static web pages.

Week 6 – Documents: Archived web pages

Wednesday, September 27; Friday, September 29

Internet Archive's Wayback Machine to analyze archived web pages.

Week 7 – Documents: Dynamic web pages

Wednesday, October 4; Friday, October 6

Dynamic web pages with `selenium`.

Week 8 – Documents: PDFs

Wednesday, October 11; Friday, October 13

Accessing and parsing PDFs with `PyPDF2`.

Week 9 – APIs: Wikipedia

Wednesday, October 18; Friday, October 20

Retrieving data from Wikipedia APIs with `wikifunctions`.

Week 10 – APIs: Government

Wednesday, October 25; Friday, October 27

Retrieving and filtering data from government APIs like the Census and Federal Reserve.

Week 11 – APIs: Social

Wednesday, November 1; Friday, November 3

Retrieving content, social graphs, and activity streams from Mastodon with `mastodon.py`.

Week 12 – APIs: AI

Wednesday, November 8; Friday, November 10

Accessing and using AI models through their APIs.

Week 13 – APIs: Automation

Wednesday, November 15; Friday, November 17

Automating web data retrieval with GitHub Actions.

Week 14 – Fall Break

Wednesday, November 22; Friday, November 24

No class.

Week 15 – Final Projects

Wednesday, November 29; Friday, December 1

Workshopping and presenting final projects.

Week 16 – Final Projects

Wednesday, December 6; Friday, December 8

Workshopping and presenting final projects.