USC Marshall
School of Business

DSO 699: Acquiring and Managing Unstructured Data for Applied Research

Instructor: Mohammed Alyakoob
Contact Info: alyakoob@marshall.usc.edu
Office Hours: TBD
Course Units: 3
Hours/week: 3 hours

Course Description

The increase in availability and accessibility of data has provided researchers from various disciplines (i.e., Information Systems, Operations Management, Statistics, Finance, Marketing, Economics, Computer Science, among others) the capability to incorporate large-scale and often unstructured data to further their research agendas. This course introduces students to the often-overlooked process of acquiring and managing large-scale unstructured datasets in applied research settings. Students will obtain hands-on experience acquiring and managing unstructured data from various sources, including web scraping, API's, geospatial data from the Census Bureau, crowdsourcing marketplaces, as well as other sources. The course will focus on providing students a framework to manage these different data types while utilizing cloud computing services to scale their research projects. During the course, each student will be tasked with a semester long project where they obtain and manage an unstructured dataset that can be incorporated in subsequent research projects.

Learning Objectives

The objective of this course is to provide students the experience and technical skills required to acquire and manage large and unstructured data sets for the purpose of incorporating these data sets in multiple research projects. Students will not be expected to contribute new methodologies on the acquisition and use of unstructured datasets. Rather, students will be able to use modern tools available to them, utilizing Python libraries and/or cloud services to obtain and manage these data sets. Students will obtain hands-on experience obtaining and managing the storage and access of different data types, including text, image, and geospatial data. Students will use cloud computing service to scale these efforts and to convert data to a more structured format. This course focuses on acquiring and managing unstructured data sets and utilizing readily available cloud tools to convert the data to a more structured format. Students that wish to obtain a deeper understanding of the theory underlying converting unstructured data to structured data (i.e., natural language and image processing) should seek more focused courses.

Course Prerequisites

Students are expected to code in Python. This course is a graduate level course and will be intensive. If you have concerns about whether this course is suitable for you, please reach out to the instructor to discuss before registration.

Required Readings, Supplementary Materials, and Laptops

This course does not have a required textbook. The class notes will be exhaustive and supplementary materials will be provided for students that want to delve deeper into specific topics. All classes will have a hands-on component and will necessitate that students have computer access.

Course Notes

Prior to each class, students will be provided a slide deck and detailed commented code that they can use to follow along in class. In cases where students will need to install software before class, installation instructions will be provided.

Grading Breakdown

Assignment	**% of Grade**
Homework (4)	50
Relevant Research Presentation	10
Code Contribution	10
Final Project	30
TOTAL	100

A. Homework

There will be 4 individual homework assignments that will be assigned throughout the semester. Homework assignments must be completed individually. Late submissions will be deducted 25% if they are submitted within 24 hours of the deadline. After the 24 hours, they will not be accepted. The homework assignments will provide students an opportunity to obtain hands-on experience using the tools learned in class with real-world unstructured data. Homework will be submitted via Blackboard. For a general guideline regarding the homework topics and their tentative dates, please refer to the tentative schedule below.

B. Research Presentation

Each student is responsible for presenting a summary of a research paper that has utilized text data, image data, and/or another form of unstructured data in an applied setting. The presentation will summarize the overall research objectives of the research project and outline the source and use of unstructured data in the paper. The instructor will be provided a list of potential papers, but students are encouraged to identify alternative research papers that incorporate unstructured data in an applied setting and that they find interesting. Students must obtain the approval of the instructor for any alternative papers they would like to present. The purpose of this presentation is to provide students with exposure to research across various disciplines that has utilized the tools discussed throughout the course. The order of presentations will be decided in the first week of class. Students will present throughout the semester.

C. Code Contribution

Each student will choose a tool (i.e., a Python library) that extends one of the topics covered during the course. Each student must submit code that walks through a specific use case of the tool. Students will present a summary of the tool and potential use cases to the class as well as a walk through of the coded example that the student submitted. The associated code will be made available to all other students. The goal of this assignment is to provide the class with a collection of scripts that can streamline the use of these tools should a member of the class choose to use them in the future. Students will be provided with a list of potential topics but are also encouraged to identify topics that they find interesting beyond the list provided. Code contributions will be submitted via Blackboard.

D. Final Project

Each student is required to complete a final project. This final project is in lieu of a final exam. The exact details will be outlined in class. The project requires that each student employ one of the applied tools discussed in class to obtain and convert unstructured data into a format that can be utilized in a research setting. Specifically, provide the research objective that this data will support and the full outline of the obtained data, the methods of storing and converting the data to a usable format, and the applied research context in which the data will be used. Students will be required to acquire unstructured data and convert it using one of the methods discussed in class. Ideally, this collected data set will be used in one of the student’s research projects after the course is completed. Importantly, all project ideas must be approved by the instructor, so please reach out to me early in the course with your ideas. Students will submit a written report (approximately 10 pages) as well as present their findings to the class.

Supplementary Lab Session

To provide students with additional opportunities to obtain hands-on experience using the tools outlined in class, biweekly (once every two weeks) supplementary labs will be available to students. These labs will provide students the opportunity to work on assignments and/or final projects with the assistance of either a TA or the instructor. The nature of acquiring and managing unstructured data is such that unforeseen problems are likely to arise. These sessions provide students an opportunity to obtain assistance while preparing final project, homework, and code contribution. I highly recommend that students utilize these opportunities whenever possible.

AI Policy

In this course, I encourage you to use artificial intelligence (AI)-powered programs to help you with assignments that indicate the permitted use of AI. You should also be aware that AI text generation tools may present incorrect information, biased responses, and incomplete analyses; thus they are not yet prepared to produce text that meets the standards of this course. 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. Please review the instructions in each assignment for more details on how and when to use AI Generators for your submissions.

Tentative Course Outline

Subject to change, note that each week has a hands-on component and will necessitate that students have computer access. Material will be provided by the instructor prior to each meeting.

WEEK(S)	TOPIC	SUBMISSION
Week 1/2	Introduction and Python Basics	
Week 3	Web Scraping (1) - Web basics and introduction to HTML. - Parsing HTML using the Beautiful Soup library and regular expressions. - Crawling sites with links, forms, logins, and/or multiple pages. - Introduction to JSON.	
Week 4	Web Scraping (2) - Scraping dynamic web pages (i.e., JavaScript). - Crawling through API's. - Scraping images and basic image processing to store. - Scaling web scrapers with Scrapy. - Using crowdsourcing platforms. - Web scraping ethics.	**Homework 1**
Week 5/6	Cloud Computing (1) - Using command line interface to interact with cloud computing platforms. - Creating and using virtual machine instances. - Cluster Computing. - Creating and accessing a virtual database.	
Week 6/7	Cloud Computing (2) - Performance benefits of using cloud computing services. - Storing and accessing unstructured data on virtual databases (i.e., using virtual MongoDB instance). - Using cloud computing to scale web-scraping projects.	**Homework 2**
Week 8/9/10	Acquiring and ManagingGeospatial Data - Introduction to geospatial analysis. - Map Projections. - OpenStreetMap. - GeoPandas. - Storing geospatial data in database.	**Homework 3**

	- Merge geospatial data with other data sources. - Conduct operations on geospatial data.	
Week 11	Natural Language Processing (1) - Obtain text data using web-scraping techniques introduced in Weeks 3/4. - Utilizing pretrained models (i.e., Python libraries) to evaluate sentiment. - Using cloud computing services to perform sentiment analysis and comparing results.	
Week 12	Natural Language Processing (2) - Using pretrained models for topic modelling. - spaCy. - Scaling with cloud computing. - Using Elasticsearch to perform text queries.	**Homework 4**
Week 13/14	Image Processing and Classification - Obtain images using web-scraping techniques introduced in Weeks 3/4. - Reading, processing, and storing images using Python. - Using cloud-based image processing services (i.e., identify objects, people, text, etc. in images and videos). - Image classification using pre-trained models .	**Code Contribution Submission**
Week 15	Final Project Presentations	

Statement on Academic Conduct and Support Systems

Academic Integrity:
The University of Southern California is a learning community committed to developing successful scholars and researchers dedicated to the pursuit of knowledge and the dissemination of ideas. Academic misconduct, which includes any act of dishonesty in the production or submission of academic work, compromises the integrity of the person who commits the act and can impugn the perceived integrity of the entire university community. It stands in opposition to the university's mission to research, educate, and contribute productively to our community and the world.

All students are expected to submit assignments that represent their own original work, and that have been prepared specifically for the course or section for which they have been submitted. You may not submit work written by others or "recycle" work prepared for other courses without obtaining written permission from the instructor(s).

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

The impact of academic dishonesty is far-reaching and is considered a serious offense against the university. All incidences of academic misconduct will be reported to the Office of Academic Integrity and could result in outcomes such as failure on the assignment, failure in the course, suspension, or even expulsion from the university.

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.

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.

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 is comprised 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) 740-5086 or (213) 821-8298
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.

Revised 07/2023