INST 447 – Data Sources and Manipulation

Pre-requisite: INST 326 or CMSC 131; INST 327

Catalog Description

Examines approaches to locating, acquiring, manipulating, and disseminating data. Imperfection, biases, and other problems in data are examined, and methods for identifying and correcting such problems are introduced. The course covers other topics such as automated collection of large data sets, and extracting, transforming, and reformatting a variety of data and file types.

Extended Course Description

This course will introduce methods and tools for developing application layers that include both front-end and back-end of a web-based system. This course will cover acquiring, installing and running database servers, web servers, modules, and web applications. This course will also cover methods, skills, and processes for developing and maintaining application layers that allow end-users to interact with underlying databases through dynamic web interfaces.

Student Learning Outcomes: After successfully completing this course you will be able to:

• Identify imperfections, biases, and other problems in data sets

• Clean up, standardize, and normalize data to prepare for data analysis

• Extract data from a variety of data types and formats

• Collect large data sets through scalable,

automated means, such as web scrapers

• Transform data among a variety of formats and standards

• Explain ethical and equity issues with the collection and use of data

Textbooks and Readings

Optional: Python for Everybody (free online) - https://www.py4e.com/book
(Optional Print Version of Above: $10)
Python for Everybody
Paperback: 242 pages
Publisher: CreateSpace Independent Publishing Platform (April 9, 2016)
Language: English
ISBN-10: 1530051126
ISBN-13: 978-1530051120

Optional: Python Data Science Handbook
(free online) - https://jakevdp.github.io/PythonDataScienceHandbook/index.html
(Optional Print Version of Above: ~$30)
Python Data Science Handbook: Essential Tools for Working with Data
Paperback: 548 pages
Publisher: O'reilly Media; 1st Edition (December 10, 2016)
Language: English
ISBN-10: 9781491912058
ISBN-13: 978-1491912058

Learn Python the Hard Way (Recommended if new to Python)

https://learnpythonthehardway.org/

Other course materials will include class notes and slides provided by the instructor on course webpage.

Required Technology and Background

- **Laptop –** We will make extensive use of computer software for this course. All course materials will be made available via ELMS on the course page. It is imperative for all students to have access to a reliable computer/laptop.

- **Python Software**. Freely available at:

  https://conda.io/docs/user-guide/install/index.html#regular-installation

Course Activities, Learning Assessments, & Expectations for Students

Before class you are expected to be prepared by:

- Reading the assigned texts or watching assigned videos
- Performing other activities, as assigned.

During class you will be assigned a variety of activities including, but not limited to:

- Completing "worksheets"(labs) comprised of programming exercises
- Participating in discussions
- Writing short reflections
- Performing other activities, as assigned.

Lab activities are graded and there will be a 12 graded activities. The lowest 2 grades will be dropped.

Homework Assignments

There will be a total of 4 homework assignments. These are your opportunity to apply concepts learned in class to real problems and data sets. These assignments will be approximately 2 page reports composed using Jupyter notebooks. You are expected to work *individually* to answer the specific problems that are assigned. Completed assignments will be submitted via Canvas/ELMS. For each assignment you must turn in a copy of your Jupyter notebook as both a .ipynb and a .html file. These act as mini-data science project reports.

You may work with your classmates to figure out the underlying concepts but are expected to work *individually* to answer the specific problems that are assigned. Timely submission of the completed assignments is essential. The due date of each assignment will be stated clearly in the assignment description. If an assignment due date is a religious holiday for you, please let the instructor know at least one week in advance, so an alternate due date can be set.

Late assignments will be penalized with a letter grade reduction, per 24 hour period. Assignments more than two days late may not be graded.

Group Project

Over the course of the semester you will also define and complete your own data science project. You will identify data set(s) and research question(s); follow the steps in the data science pipeline to extract insights from data; and report on your results. The only requirements for this project is that 1) it must be centered on data, 2) it

must tell us something interesting, and 3) to get an A you must go beyond what we learn in class. You will work on this project as a group of 3 to 4 members. Early in the semester you will turn in a project proposal that outlines your goals for the project. All projects must be approved by me. At the end of the semester you will present your results in class and write up your results as blog post composed in Jupyter notebook.

Grading

Class Activities 10%

- 12 in class exercises (drop 2)

Homework 40%

- 4 programming assignments

Group Project 50%

- Project proposal (20%)
- Project presentation (10%)
- Project report (20%)

Your final grade for the course is computed as the sum of your scores on the individual elements below (100 possible points total), converted to a letter grade:

A+ 97-100*	**B+ 87-89.99**	**C+ 77-79.99**	**D+ 67-69.99**	**F 0-59.99**
A 93-96.99	**B 83-86.99**	**C 73-76.99**	**D 63-66.99**	
A- 90-92.99	**B- 80-82.99**	**C- 70-72.99**	**D- 60-62.99**	

*** Note: To receive an A+ you must have demonstrated significant contributions to the class in addition to achieving this numeric grade. We reserve the right to curve grades upward (but will not curve grades downward).**

COURSE SCHEDULE

Week Schedule

Week 1: 8/26	Introduction and Overview	Install Python Software
Week 2: 9/2	**Thinking about data:** Tidy Data Python and Pandas	**Readings:** Wickham, H. (2014). Tidy Data. Journal of Statistical Software, 59(10). VanderPlas, J. (2016). Introducing Pandas Objects. In Python Data Science Handbook. O'reilly Media.
Week 3: 9/9	**Working with Pandas:** Indexing & Slicing DataFrames Combining & Merging DataFrames	**Readings:** VanderPlas, J. (2016). Data Indexing and Selection. In Python Data Science Handbook. O'reilly Media. VanderPlas, J. (2016). Combining Datasets: Merge and Join. In Python Data Science Handbook. O'reilly Media.
Week 4: 9/16	**Exploring and Visualizing Data:** Aggregations with Pandas Visualizations with Matplotlib	**Readings:** VanderPlas, J. (2016). Aggregation and Grouping. In Python Data Science Handbook. O'reilly Media. VanderPlas, J. (2016). Data Indexing and Selection. In Python Data Science Handbook. O'reilly Media.
Week 5: 9/23	**Metadata and Missing Data:** Missing data Bad data Meta data	**Readings:** Quartz Guide to Bad Data: https://github.com/Quartz/baddata-guide VanderPlas, J. (2016). Handling Missing Data. In Python Data Science Handbook. O'reilly Media. DataONE Education Module: Metadata One Page Handout. DataONE.
Week 6: 9/30	**Data Pipelines and SQL:** Using databases in Python Using databases with Pandas	**Readings:** Severance, C. (2016). Chapter 15: Python and Databases. In Python for everybody : Exploring data using python 3. Ann Arbor, MI: Charles Severance.

Week 7: 10/7	**Structured Data - XML:** Understanding XML Navigating trees Ingesting XML data	**Readings:** XML Tutorial w3schools.com Severance, C. R. (2016). Chapter 13: Using Web Services. In Python for everybody: exploring data using Python 3 (pp. 155-158). Ann Arbor, MI: Charles Severance. Helland, Pat. "XML and JSON Are Like Cardboard." Communications of the ACM 60, no. 12 (December 2017): 46–47.
Week 8: 10/14	**Structured Data - JSON:** Ingesting JSON data	**Readings:** Severance, C. R. (2016). Chapter 13: Using Web Services. In Python for everybody: exploring data using Python 3 (pp. 155-158). Ann Arbor, MI: Charles Severance. JSON Tutorial w3schools.com
Week 9: 10/21	**Advanced Data Ingestion: APIs** Using APIs Geocoding	**Readings:** Severance, C. R. (2016). Chapter 13: Using Web Services. In Python for everybody: exploring data using Python 3 (pp. 155-158). Ann Arbor, MI: Charles Severance.
Week 10: 10/28	**Advanced Data Ingestion: Web Scraping** Web scraping Data science ethics	**Readings:** Python, Real. "Practical Introduction to Web Scraping in Python – Real Python." Fiesler, Casey. "Law & Ethics of Scraping: What HiQ v LinkedIn Could Mean for Researchers Violating TOS." Medium (blog), August 15, 2017.
Week 11: 11/4	**Text Processing: Strings** String manipulations String Operations	**Readings:** Severance, C. R. (2016). Chapter 11: Regex. In Python for everybody: exploring data using Python 3 (pp. 155-158). Ann Arbor, MI: Charles Severance.
Week 12: 11/11	**Text Processing: Regular Expressions** Regular Expressions Unstructured data	**Readings:** Severance, C. R. (2016). Chapter 6: Strings. In Python for everybody: exploring data using Python 3 (pp. 155-158). Ann Arbor, MI: Charles Severance.
Week 13: 11/18	Special Topics	TBD
Week 14: 11/25	Fall break	
Week 15: 12/2	Data Science Team Projects	
Week 16: 12/9	Data Science Team Projects	

This schedule is for planning purposes and may change. See course webpage for current information and deadlines.

Policy on Academic Misconduct Cases of academic misconduct will be referred to the Office of Student Conduct irrespective of scope and circumstances, as required by university rules and regulations. It is crucial to understand that the instructors do not have a choice of following other courses of actions in handling these cases. There are severe consequences of academic misconduct, some of which are permanent and reflected on the student's transcript. For details about procedures governing such referrals and possible consequences for the student please visit http://osc.umd.edu/OSC/Default.aspx.

It is very important that you complete your own assignments, and do not share any Excel or SPSS files or other work. The best course of action to take when a student is having problems with an assignment question is to contact the instructor. The instructor will be happy to work with students while they work on the assignments.

University of Maryland Code of Academic Integrity

"The University of Maryland, College Park has a nationally recognized Code of Academic Integrity, administered by the Student Honor Council. This Code sets standards for academic integrity at Maryland for all undergraduate and graduate students. As a student you are responsible for upholding these standards for this course. It is very important for you to be aware of the consequences of cheating, fabrication, facilitation, and plagiarism. For more information on the Code of Academic Integrity or the Student Honor Council, please visit http://shc.umd.edu/SHC/Default.aspx.

Special Needs

Students with disabilities should inform the instructor of their needs at the beginning of the semester. Please also contact the Disability Support Services (301-314-7682 or http://www.counseling.umd.edu/DSS/). DSS will make arrangements with the student and the instructor to determine and implement appropriate academic accommodations. Students encountering psychological problems that hamper their course work are referred to the Counseling Center (301-314-7651 or http://www.counseling.umd.edu/) for expert help.