CSE 349/449 BIG DATA ANALYTICS

Fall 2023 • 2:05 pm – 3:20 pm MW • Mountaintop Building C 115[1]

Instructor **Professor Daniel Lopresti**
Email dal9@lehigh.edu ~ Ext 85782 ~ Zoom Office Hours TBD

Texts *Advanced Analytics with PySpark: Patterns for Learning from Data at Scale Using Python and Spark (1st Edition)* by Akash Tandon, Sandy Ryza, Uri Laserson, Sean Owen, and Josh Wills, O'Reilly Media, 2022, ISBN-13: 978-1098103651.

Also freely available via Lehigh University Library using your Lehigh login: https://learning.oreilly.com/library/view/advanced-analytics-with/9781098103644/

Data Mining: Practical Machine Learning Tools and Techniques (4th Edition) by Ian Witten, Eibe Frank, Mark A. Hall, and Christopher Pal, Morgan Kaufmann, 2016, ISBN 978-0128042915.

Also freely available via Lehigh University Library using your Lehigh login: https://learning.oreilly.com/library/view/data-mining-4th/9780128043578/

CourseSite Materials and discussion forums will be available @ http://coursesite.lehigh.edu/

Grading

• Homework assignments	50 points (25%)
• Class participation	50 points (25%)
• Final project, presentation, and write-up	100 points (50%)

Notes This 3-credit project course gives a practical working knowledge of large-scale data analysis using the popular open-source Apache PySpark framework. The PySpark programming model elegantly supports patterns that are commonly employed in big data analytics, including classification, collaborative filtering, and anomaly detection, among others. We will also employ Weka, a well-known open-source platform for developing machine learning applications. Although Weka is typically used as a GUI-based tool, it also provides Knowledge Flow and programmatic interfaces which are more efficient for big data applications.

Enrollment in this course is limited and requires permission of the instructor. Please note that this is not a basic course on data mining, machine learning, or distributed computing; it assumes you already know something about these topics and/or you can pick up the necessary details on your own. The course also assumes you already have substantial programming experience in one or more high-level languages.

Accommodations for Students with Disabilities If you have a disability for which you are or may be requesting accommodations, please contact both your instructor and the Office of Academic Support Services, Williams Hall, Suite 301 (610-758-4152) as early as possible in the semester. You must have documentation from the Academic Support Services office before accommodations can be granted.

---

[1] This hands-on project course will be taught in-person on the Lehigh campus. Students who might have a documented need to connect remotely should contact the instructor in advance for further details. We will employ online resources that are accessible remotely from anywhere in the world, including access to the CSE Department computing network and Lehigh's CourseSite system. We will always follow all federal, state, local, and university safety guidelines.

CSE 349/449 BIG DATA ANALYTICS

Fall 2023 • 2:05 pm – 3:20 pm MW • Mountaintop Building C 115

Principles of Our Equitable Community

Lehigh University endorses The Principles of Our Equitable Community[2]. We expect each member of this class to acknowledge and practice these Principles. Respect for each other and for differing viewpoints is a vital component of the learning environment inside and outside the classroom.

Academic Integrity

The work you submit in CSE 349/449 must be entirely your own. While I encourage you to discuss basic concepts with others, plagiarism is never acceptable. Neither is reusing work you did for another purpose, or employing generative AI tools such as ChatGPT without my previous permission. Such cases will be referred to the University Committee on Discipline and, if you are found guilty, you may be given the failing grade WF in the course. If you have questions about this policy at any point, ask me. Be safe, not sorry.

Tentative Course Schedule

Date	Topic
Week 1	Course Intro
Week 2	Hands-on Intro to PySpark
Week 3	Hands-on Intro to Weka
Week 4	Recommending Music and the Audioscrobbler Data Set
Week 5	Predicting Forest Cover with Decision Trees
Week 6	Anomaly Detection in Network Traffic with K-means Clustering
Week 7	Understanding Wikipedia with Latent Dirichlet Analysis
Week 8	Text Sentiment Analysis with Weka
Week 9	Geospatial and Temporal Data Analysis on the NYC Taxi Trip Data
Week 10	TBD
Week 11	Final Project Proposals
Week 12	Multilayer Perceptrons with Weka
Week 13	Image Similarity Detection with Deep Learning and PySpark LSH
Week 14	Ethics of Big Data Analytics
Week 15	Final Project Presentations; Course Wrap-Up

* Readings are taken from our two textooks, *Advanced Analytics with PySpark (AAP)* and *Data Mining (DM)*. In addition to the specified readings along with the associated supplemental materials assigned throughout the course, you will work through big data programming exercises on your own. After you have completed the basic functionality, you will implement your own extensions and enhancements, which you will be asked to demonstrate for the class. Creativity will be rewarded! Class participation will constitute a significant portion of your grade.

[2] https://www2.lehigh.edu/diversity-inclusion-equity/principles-equitable-community