IST 418: Big Data Analytics

Course Syllabus

Instructor: Chris Dunham; cndunham@syr.edu

Class Meetings: [LOCATION], [DAY] [TIME]

Office Hours: Office hours are held each [WEEKDAY] during weeks with class meetings, [TIME] in Hinds 309, and by appointment via email.

Communications. I prioritize communication by email. Please do not use Teams or messages within Blackboard because I am not likely to see them in a timely manner. When communicating by email, the subject line should read **IST 418: Brief description of question/issue.** If you have questions regarding code, please attach your code file (ipynb) to the email and provide a detailed explanation of your issue in the body of the email. Note that I generally do not correspond outside of standard working hours, including weekends, so plan accordingly.

Course Description: Learn to develop actionable insights from big data using open-source tools (Python and Spark). This course prepares students to build scalable data analytics pipelines and apply advanced machine learning techniques, culminating in a hands-on project tackling real-world challenges.

Course Prerequisites: IST 387.

Credits: 3

Course Fees and/or Costs: There are no course fees or additional costs beyond tuition that the student should expect to incur.

Learning Objectives: After taking this course, students will be able to:

1. Translate a business challenge into an analytics challenge.
2. Analyze big data, create statistical models, and identify insights that can lead to actionable results.
3. Use Python and Apache Spark to build big data analytics pipelines.
4. Learn classic and state-of-the-art machine learning (ML) techniques.
5. Explain how advanced analytics can be leveraged to create competitive advantages.

Shared Competencies: In IST 418, labs, homework assignments, and the final project ensure that students meet the expectations of the following Syracuse university shared competencies:

- **Critical and Creative Thinking:** Exploration and synthesis of ideas, artifacts, issues, and events to inform and evaluate arguments, develop new insights, and produce creative work. Reflection on, and application of divergent modes of inquiry, analysis, and innovation to research, knowledge, and artistic creation.

IST 418: Big Data Analytics

Course Syllabus

- **Scientific Inquiry and Research Skills:** Application of scientific inquiry and problem-solving in various contexts. Analysis of theories, replication of procedures, and rethinking existing frameworks. Supporting arguments through research, data, and quantitative and qualitative evidence that can generate new knowledge.
- **Information Literacy and Technological Agility:** Identification, collection, evaluation, and responsible use of information. Effective, ethical, and critical application of various technologies and media in academic, creative, personal, and professional endeavors.

Recommended Text: Essential PySpark for Scalable Data Analytics. Sreeram Nudurupati, 2021.

Software: We will use the Python programming language and PySpark with Google Colab.

Course Requirements and Expectations:

Assessment	Notes	Percent of Grade
In-class lab assignments	Exercises and questions in Jupyter notebooks; due the day they are used in class; no late submissions	10 * 1% = 10%
In-class quizzes	Covers concepts; no make-ups; lowest grade dropped	5 * 5% = 25%
Homework	Based on class materials; late submissions are discounted	1 * 10% + 3 * 5% = 25%
Group Project	Groups of 3-4; you choose your teammates or request to be assigned to a group	Total = 40% Proposal – 5% Presentation – 10% Report – 20% Code – 5%

Labs: Labs take the form of Jupyter notebooks designed to be run in Google Colab. They demonstrate how to perform analytics in PySpark, allow you to practice new skills through exercises, and reinforce concepts from reading and lecture materials. They will contain questions and exercises that shall be completed in class and submitted to Blackboard for credit the same day they are used in class. Grading is based on a *good-faith effort* of the exercise. Late labs will not be accepted.

IST 418: Big Data Analytics

Course Syllabus

Quizzes: Quizzes are individual effort, in-class short tests which measure your understanding of the concepts and terminology covered in class and the assigned readings. There will be no make-up quizzes, though I will count only the best 4 out of 5 quizzes toward your final quiz average. Quizzes will cover all material presented in class, up to the point when the quiz is issued.

Homework: Homework assignments are designed to check that you are keeping pace with class concepts. While you are encouraged to discuss homework with your classmates as part of the learning process, homework programming and writing is an individual effort. All work submitted must be your own. Assignments are due by 11:59 PM Sundays, the week they are due (see due dates in Blackboard). Any assignment handed in after the deadline (not counting a 1-hour grace period) is automatically discounted by 1 point of the final grade per day late.

Group Project: The group project is your chance to demonstrate what you've learned in the course and apply it to a new scenario. It is expected that each group's project will be novel. Your group will be responsible for developing an idea and obtaining data on your own.

The group project consists of four elements:

1. **Project proposal**: 2-page preliminary description of the project, including problem statement, solution proposed, techniques, datasets, and expected results. Groups will present their proposals informally in class for discussion and feedback.
2. **Final presentation:** A live presentation during the last week of class. Students will get a zero for their individual grade related to the project presentation if the student arrives late, leaves early, or doesn't participate in the presentation. Each student on the project team is expected to present a portion of the presentation.
3. **Project report:** A comprehensive report due at the end of the semester.
4. **Project code:** A well-organized, fully commented, Jupyter notebook which runs without error. Project code is due at the same time as the project report.

Grading: Below is the formula for number-to-letter grade conversion.

Grade	Points	Grade	Points	Grade	Points	Grade	Points
A	93-100	B+	87-89	C+	77-79	D	60-69
A-	90-92	B	83-86	C	73-76	F	0-59
		B-	80-82	C-	70-72		

IST 418: Big Data Analytics

Course Syllabus

Preliminary Course Schedule (subject to change):

Unit/Dates	Topics	Readings	Events
01 Jan 13, 15	Introductions and Course Overview		
01 Jan 22	Python Review & Exercises		
02 Jan 27, 29	Big Data & Distributed Computing	Ch. 1	
03 Feb 3, 5	Data Ingestion, Cleansing & Integration	Ch. 2 & 3	HW 1 due Sun
04 Feb 10, 12	Spark SQL & Spark DataFrames	Ch. 12	Quiz 1 Mon (units 1 and 2) Project groups due Sun
05 Feb 17, 19	ML & Feature Engineering Concepts	Ch. 5 & 6	Quiz 2 Mon (units 3 and 4) HW 2 due Sun
06 Feb 24, 26	Supervised Learning Part I	Ch. 7	
07 Mar 3, 5	Supervised Learning Part II		HW 3 due Sun
SPRING BREAK — NO CLASSES			
08 Mar 17, 19	Unsupervised Learning Part I	Ch. 8	Quiz 3 Mon (units 6 and 7) Project proposals due Sun
Mar 24, 26	Group Project: Proposal Presentations		
09 Mar 31, Apr 2	Unsupervised Learning Part II		
10 Apr 7, 9	Special Topic: TBD		Quiz 4 Mon (units 8 and 9) HW 4 due Sun
Apr 14, 16	Group Project: Working sessions/check-ins		
Apr 21, 23	Quiz 5 Mon (comprehensive) and Group Project: Working sessions/check-in		
Apr 25 (Fri.)	Group Project: Final Presentations at Poster Day		
Apr 28	No class meeting, work on your final project reports		
May 2	Project report and code due to Blackboard		