CSE 398/498  BIG DATA ANALYTICS
Fall 2021  •  2:05 pm – 3:20 pm MW  •  Mountaintop Building C 115

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

Texts


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 Spark framework. The Spark 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, we will focus on its 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.

Principles of Our Equitable Community
Lehigh University endorses The Principles of Our Equitable Community. 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.

---

1 Students may take this hands-on project course in-person on the Lehigh campus or remotely using live (synchronous) Zoom video conferencing, depending on your declared location. 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.

2 http://www.lehigh.edu/~inprv/initiatives/PrinciplesEquity_Sheet_v2_032212.pdf
Academic Integrity

The work you submit in CSE 398/498 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. 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. It is far better to be safe than sorry when your academic career may be on the line.

Tentative Course Schedule

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td>Course Intro</td>
</tr>
<tr>
<td>Week 2</td>
<td>Hands-on Intro to Scala and Spark</td>
</tr>
<tr>
<td>Week 3</td>
<td>Hands-on Intro to Weka</td>
</tr>
<tr>
<td>Week 4</td>
<td>Recommending Music and the Audioscrobbler Data Set</td>
</tr>
<tr>
<td>Week 5</td>
<td>Predicting Forest Cover with Decision Trees</td>
</tr>
<tr>
<td>Week 6</td>
<td>Anomaly Detection in Network Traffic with K-means Clustering</td>
</tr>
<tr>
<td>Week 7</td>
<td>Understanding Wikipedia with Latent Semantic Analysis</td>
</tr>
<tr>
<td>Week 8</td>
<td>Instance and Linear Model Classification using Weka</td>
</tr>
<tr>
<td>Week 9</td>
<td>Analyzing Co-occurrence Networks with GraphX</td>
</tr>
<tr>
<td>Week 10</td>
<td>Geospatial and Temporal Data Analysis on the NYC Taxi Trip Data</td>
</tr>
<tr>
<td>Week 11</td>
<td>Final Project Proposals</td>
</tr>
<tr>
<td>Week 12</td>
<td>Estimating Financial Risk through Monte Carrol Simulation</td>
</tr>
<tr>
<td>Week 13</td>
<td>Deep Learning using Weka</td>
</tr>
<tr>
<td>Week 14</td>
<td>Ethics of Big Data Analytics</td>
</tr>
<tr>
<td>Week 15</td>
<td>Final Project Presentations; Course Wrap-Up</td>
</tr>
</tbody>
</table>

* Readings are taken from our two textooks, *Advanced Analytics with Spark (AAS)* 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.