Course Syllabus

Instructor:
Dr. Wael Jabr
Email: wjabr@gsu.edu
Office: Robinson College of Business - Room 906
Phone: 404.413.7363
Office Hours: By appointment

Pre- or Co-requisites:
CIS 8040.

Recommended Textbooks & Supplementary Material:
- Book1: Taming Text, by Grant Ingersoll and Thomas Morton, Manning Publications, 2013
- Papers assigned during the term. Check the course site for a complete listing.

Course Description:
This course addresses the unstructured data management skills needed for modern data analysis including those salient to big data and real-time data environments. The focus is on unstructured data and its environment. Unstructured data includes web data (blogs, text), user generated content, social media, location-aware data, and digital media among others. Topics covered include extraction methods for real time audio and video data, data capture, cleaning, representation, storage, queries, manipulation, and real-time data management. Also included as they apply to unstructured data environment are data security, governance, and visualization. Students will learn natural language processing and geo-spatial analytical tools.

Course Objectives:
Upon completion of the course, students should be able to:
- Articulate the similarities and differences between managing structured & unstructured data.
- Apply techniques to manage unstructured data and develop an unstructured data warehouse.
- Integrate data from multiple sources such as social media.
• Prepare unstructured data for analysis including real-time analysis, predictive analytics, data mining, optimization, etc.
• Apply techniques for visualizing unstructured data
• Recommend best practices related to security and governance of unstructured data.

Weekly Mini-Projects:
Each week, students complete a hands-on project that further explores the topic/technique covered in class. This is an individual activity. With these min-projects, students gain proficiency in the various SW and databases assigned for this class.
Check the course site for details.

Final Project:
The project consists of a research report and presentation on a student-selected topic that is relevant to the course. It is group-based. A sample topic might be to track the performance of products listed on Amazon. This requires defining the research problem (e.g., how to quantify performance), identify contributing factors (e.g., internal and competitive pressure), gather data (e.g., review write-ups), and use appropriate analysis techniques. On the last day of class, each group will present their findings to the class.
Check the course site for additional details.

Course Software & Databases:
• XML (freely available online)
• MongoDB (freely available online)
• Neo4j (freely available online)
• SAS Enterprise Miner (provided through the class)

Typical class session:
Class sessions will comprise (1) lectures/discussions of relevant techniques, concepts and features, (2) instructor demonstrations, and (3) student lab sessions with hands-on work.
The purpose of this pedagogical approach is to introduce and reinforce ideas and skill sets so that you can master these on your own after class hours.
To bring this knowledge to a highly proficient, professional level, you will have to spend time and effort outside of class reviewing and practicing the class material.
To ensure that you have the basic knowledge that will allow you to function on your own after class, be sure to ask the instructor questions during class, either during the lecture/discussion, demo, or lab.

Classroom guidelines:
Attendance is not mandatory but highly recommended.
Coming to class fully prepared and contributing to the discussion help deepening the learning.
Individual deliverables are to be submitted individually and group work is collaborative.
Refer to http://www2.gsu.edu/~wwwfhb/sec400.html for additional information on instructional information. Another useful resource is available at http://deanofstudents.gsu.edu/student-conduct/academic-honesty-policy/.
**Grading:**
Deliverables*

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Participation &amp; Quizzes</td>
<td>10%</td>
</tr>
<tr>
<td>Mini-Projects</td>
<td>45%</td>
</tr>
<tr>
<td>Final Project</td>
<td>30%</td>
</tr>
<tr>
<td>Exam</td>
<td>15%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>

*Late submission policy: deliverables submitted after their due date will be penalized 10% per day for the following three days. No submission is accepted after the third day.

**Letter Grade Scale**

<table>
<thead>
<tr>
<th></th>
<th>98</th>
<th>94</th>
<th>90</th>
<th>86</th>
<th>82</th>
<th>79</th>
<th>75</th>
<th>71</th>
<th>68</th>
<th>65</th>
<th>62</th>
<th>Less than 61</th>
</tr>
</thead>
<tbody>
<tr>
<td>A+</td>
<td>A</td>
<td>A-</td>
<td>B+</td>
<td>B</td>
<td>B-</td>
<td>C+</td>
<td>C</td>
<td>C-</td>
<td>D+</td>
<td>D</td>
<td>F</td>
<td></td>
</tr>
</tbody>
</table>
### Class Schedule (modifications may be necessary)

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Readings</th>
<th>Mini-projects (MP)</th>
</tr>
</thead>
</table>
| Class 1 Jan 12 | Unstructured data: Usefulness and Challenges  
- Scaling issues  
- Intro to SW & Databases adopted | - Book1 chapters 1  
- Article “Unstructured Textual Data in the Organization” by B. Inmon & A. Nesavich, Prentice Hall, 2008 | MP 1 assigned |
| Class 2 Jan 13 | Reading unstructured data: well formatted (XML), not so well-formatted (free flowing)  
- Intro to automated data readers | - Book2 chapters 1, 2, 3, 4  
- Book4 chapters 10 | SW installation (No MP 2) |
| Class 3 Jan 19 | Storing & Manipulating Unstructured Data  
- Lessons from NoSQL | - Book3 chapters 1, 2, 9 | MP 1 due  
MP 3 assigned |
| Class 4 Jan 26 | Representing & Scaling text  
- Use of SAS EM | - Book1 chapters 2  
- Book4 chapters 1, 2, 3, 6  
- Article “CAP Twelve Years Later: How the "Rules" Have Changed” by Brewer, Computer, 45(2), Feb 2012 | SW installation |
| Class 5 Feb 2 | Visualizing text (graphs, network analysis)  
- Article “Tapping the power of social networks” by Miller & Christakis, Harvard Business | MP 3 due  
MP 4 assigned |
| Class 6 Feb 9 | Data management (ownership, security, privacy...)  
Guest Speaker: John Holt, Sr. Architect at LexisNexis | - Book5 chapters 14, 18, 19, 20 and 21  
- Article “The Cross-Atlantic Tussle over Financial Data and Privacy Rights” by Selnier & Frasher, Business Horizons, 56(6), Nov 2013 | |
| Class 8 Feb 23 | Project presentation & report submission  
Final Exam | | |