**Web Analytics (BIA 660)**

**Tuesdays, 6:15-8:45pm, BC 310**

**Professor Winter Mason**

**Course Description:**

In this course, students will learn through hands-on experience how to extract data from the web and analyze web-scale data using distributed computing. Students will learn different analysis methods that are widely used across the range of internet companies, from start-ups to online giants like Amazon or Google. At the end of the course, students will apply these methods to answer a real scientific question or to create a useful web application.

**Prerequisites**: Students must have taken Multivariate Data Analytics (BIA 652), Knowledge Discovery in Databases (MIS 637) and Statistical Learning & Analytics (BIA 656). It is also highly recommended for the students to have programming experience.

**Grading Percentages**: HW: 20%; Class work: 20%; Mid-term Project: 25%; Final Project: 35%

**Mid-term project:** *Combine methods from first four weeks of class to generate a novel data set (e.g., by combining different sources of data) and write a report exploring data with at least 3 key insights.*

**Final project:** *Option 1: Choose a scientific question from a relevant literature and write a report describing findings, to be submitted to a scientific journal or conference.* *Option 2: Combine data sources from the web to generate a useful and interesting web application, to be published online.*

**Textbook(s) or References**:

[Programming Collective Intelligence: Building Smart Web 2.0 Applications](http://www.amazon.com/Programming-Collective-Intelligence-Building-Applications/dp/0596529325/ref=sr_1_1?s=books&ie=UTF8&qid=1310054165&sr=1-1)

Toby Segaran / Paperback / 2007

[Hadoop: The Definitive Guide](http://www.amazon.com/Hadoop-The-Definitive-Guide-ebook/dp/B0043D2ECC/ref=pd_sim_kinc_4?ie=UTF8&m=AG56TWVU5XWC2)

Tom White / Paperback / 2010

**Recommended Reading:**

[The Visual Display of Quantitative Information](http://www.amazon.com/Visual-Display-Quantitative-Information/dp/0961392142/ref=sr_1_1?s=books&ie=UTF8&qid=1310053258&sr=1-1)   
Edward Tufte / Hardcover / 2001

[Visualize This: The FlowingData Guide to Design, Visualization, and Statistics](http://www.amazon.com/Visualize-This-FlowingData-Visualization-Statistics/dp/0470944889/ref=sr_1_1?ie=UTF8&qid=1310053219&sr=8-1)

Nathan Yau / Paperback / 2011

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Topic(s)** | **Reading(s)** | **HW** |
| 8/28 | Introduction to the course & Overview; Review of relevant programming languages and web technologies |  |  |
| 9/4 | Obtaining Data:  Scraping Web Data  Web Crawlers  APIs  Crowd Sourcing | Ch. 4 in Programming Collective IQ | Assignment 1: Build a web crawler & collect data |
| 9/11 | Obtaining data, part 2 |  |  |
| 9/18 | Data cleaning  Data types  Regex  Databases |  | Assignment 2: Clean data for later analysis |
| 9/25 | Visualizing data for exploration | Visual display of quantitative information, Ch. 2 & 4 | Assignment 3: Create 3 visualizations that describe data from Assignment 1 |
| 10/2 | Cloud computing:  Map Reduce  Hadoop  PIG  Streaming Algorithms | Chapter 2,3,5,6 in Hadoop: The Definitive Guide |  |
| 10/16 | Integrating Week 1-5  Introduce mid-term projects |  | Begin Mid-term project |
| 10/23 | Social Network Analysis Community detection | Ch. 3 in Programming Collective IQ  Newman (2001)  Yang & Leskovec (2011) | Continue mid-term project |
| 10/30 | **No class**  **Hurricane Sandy** | | |
| 11/6 | **No class**  **Hurricane Sandy** | | |
| 11/13 | **Presentation of Mid-Term Project Results** |  |  |
| 11/20 | Visualizing data for exposition  Building mashups | Visual display of quantitative information, Ch. 5 & 9  Visualize This | Begin work on Final project |
| 11/27 | Final Project Introductions |  | Continue work on Final project |
| 12/4 | Final Project Presentations |  |  |