**Web Analytics (BIA 660)**

**Tuesdays, 6:15-8:45pm, BC 220**

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

**Grading Percentages**: Class work: 20%; Mid-term Project: 30%; Final Project: 50%

**Mid-term project:** *Develop a solution to a web analytics problem, to be explained in class. This will require collecting and cleaning data and doing some analyses to present the results.*

**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

**Recommended Reading:**

[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

[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** |
| 1/15 | Introduction to the course  Install python & packages  Install gitHub | Dive Into Python, Chapter 5-6  CheckIO,, problems 4-6 |  |
| 1/22 | Diving into Python  In-class Python exercises |  |  |
| 1/29 | XML, HTML, CSS  Javascript, JSON  In-class web-page exercises | Ch. 4 in Programming Collective IQ  Dive Into Python, Chapter 8 & 11 |  |
| 2/5 | Obtaining Data:  Scraping Web Data  Web Crawlers  APIs  Introduce mid-term projects | Dive Into Python, Chapter 7 | Decide coordination plan with team |
| 2/12 | Databases  Data cleaning  Regex  Basic data analysis | *Visual display of quantitative information, Ch. 2 & 4* | Work on mid-term project |
| 2/26 | Visualizing data for exploration  Google AppEngine & AppEngine Data Store |  | Prepare presentation |
| 3/5 | **Presentation of Mid-Term Project Results** | *Chapter 2,3,5,6 in Hadoop: The Definitive Guide* |  |
| 3/19 | Cloud computing:  Map Reduce  Hadoop  Amazon AWS  Word count example |  |  |
| 3/26 | Cloud computing:  Streaming Algorithms  PIG  BFS example | *Visual display of quantitative information, Ch. 5 & 9*  *Visualize This* |  |
| 4/2 | Final Project Introductions  Visualizing data for exposition  d3, bootstrap, processing | Ch. 3 in Programming Collective IQ  Newman (2001) | Decide coordination plan with team |
| 4/9 | Social Network Analysis | Ch. 6 in Programming Collective IQ |  |
| 4/16 | Text Mining  Topic Modeling | Ch. 7 & 9 in Programming Collective Intelligence | Begin work on Final project |
| 4/23 | Statistical Learning / Classification |  | Continue work on Final project |
| 4/30 | **Final Project Presentations** |  |  |