Course Proposal

STA141B Data & Web Technologies for Data Analysis

Units: 4

Format: 3 hours of lecture per week.

Discussion: 1 hour per week.

Prerequisites: STA141A or ECS145

Catalog Description: Essentials of using relational databases and SQL. Processing data in blocks. Scraping Web pages and using Web services/APIs. Basics of text mining. Interactive data visualization with Web technologies. Computational data workflow and best practices. Statistical methods.

Course Topics & Content:

- Relational databases, algebra and SQL (4 lectures)
- Web Scraping concepts & technologies - HTML, XPath (3 lectures)
- Web Services/APIs concepts & technologies - REST, HTTP, XML, JSON, OAuth (4 lectures)
- Text mining - applications, statistical methods, stop words, stemming, natural language processing (5 lectures)
    - Optional: NoSQL databases, ElasticSearch/Lucene text search engine.
- Advanced visualization & mashups - (5 lectures)
    - Concepts of advanced visualization (dynamic, interactive, animated).
    - Computational models.
    - R (grid, ggplot2)
    - Interactive, animated, dynamic plots (with R, JavaScript, SVG, D3, HTML, shiny)
    - Dashboards
- Concepts of block/on-line/streaming algorithms for statistical methods (3 lectures)
- Intermediate/advanced aspects of R:
    - object-oriented programming, classes and methods
    - R packages & basic software design principles
    - Unit testing of code
    - Portability issues for developing code.(4 lectures)
- Version control (git) (1 lecture)
- Reproducibility & dynamic documents (knitr, iPython) (1 lecture).

Statistical Methods explored may include

- Naive Bayes.

- Recommendation systems.
- Clustering.
- Latent Semantic Analysis.
- Sentiment analysis.

Grading: A series of 4 - 6 computer assignments. The instructor may choose to use participation (in class, discussion section and online forum). Similarly, the instructor may include a mid-term and/or final exam. Given the practical computer-based nature of this course, an instructor may deem written exams inappropriate.

Learning Goals/Outcomes: Students learn the concepts and gain experience in using fundamental technologies for data sciences. They learn to access data from new sources and how to convey results using rich technologies. They also learn to analyze text data in qualitatively different ways. They also see statistical methods not taught in other courses.

Illustrative Reading:

- Introduction to Data Technologies, Murrell
- XML and Web Technologies for Data Sciences with R, Nolan and Temple Lang
- Interactive Data Visualization for the Web, Murray.

Course Overlaps: The topics in this course overlap with more specialized courses. A key difference between this course and those is a) we focus on practical use of these technologies, rather than a deep understanding of them or implementing them, and accordingly, b) we cover the material in 2 weeks rather than 10 weeks. The relational database component shares concepts with ECS 165A & B but covers only 1/5th of the material. The visualization component is a much briefer and differently focused treatment of material in ECS163.