Z604 – **Big Data Analytics for Web and Text**

Instructor:	Xiaozhong Liu	E-Mail:	liu237@indiana.edu
	Miao Chen		miaochen@indiana.edu

Data science is a rising discipline that uses data to effectively characterize, interpret or predict complex real-world problems. The importance of data science also lies in its huge potential of changing our current way of doing science and social sciences. Big data, featuring in its high heterogeneity and volume, calls for our new understanding and skills towards data and data operations. Due to these innovative features, new methods, algorithms and applications are needed to index, process and analyze large-scale data. This class focuses on analytics of two types of important big data: *web and text data*.

This course introduces the fundamentals of data science and big data analysis by focusing on: theoretical aspects, such as their philosophical grounds and implications, and methodological aspects, such as numerical and textual data processing, basic statistical analysis and machine learning, data retrieval and recommendation, data representation and semantics, big data storage, along with several important case studies. In addition, this course will introduce and demonstrate open source data-operation framework and tools, e.g., *R, Hadoop, Lucene*, and *NoSQL* (*MongoDB*), which enable students to design class projects with provided real-world data sets.

This course has two goals: to develop students' conceptual understanding of how data science is revolutionizing classical information management and scientific inquiry, and second, to help students acquire hands-on experience and basic implementation capabilities to grasp data science skills and apply to problems of other scientific fields.

In this course, students will learn:

- The most important and current grounding philosophies, theories, and models for data science
- How to view real-world problems from lens of these theories and models and solve these problems using the data science perspective (case studies)
- Basic data processing and statistical analysis methods
- Basic machine learning, data retrieval, ranking, and recommendation algorithms
- Basics of R, Lucene, Hadoop and NoSQL (MongoDB), in lab sessions

Lab session is an important and integral part in the course because it provides valuable chances for students to practice hands-on skills and also enhances students' understanding of the concepts. This course designs a lab session for each theme of data science (as shown in the schedule). Students will learn to use important tools for addressing real-world problems via the instructors' demonstration and one-to-one interaction.

Class Datasets

Since this is a data intensive class, data sets are essential to the class for understanding and analyzing big data. Here are some candidate large-scale data sets that can be used for assignments:

1) **Wikipedia dump**: 386,413,5 Wikipedia articles (in English) with their categorical relationships
2) **Netflix movie**: Large number of movies from Netflix with user watch and rating history information
3) **Twitter**: Large number of Twitter data for textual data analysis and its related social network information
4) **HathiTrust**: a subset of book collections digitalized in the HathiTrust Digital Library (a.k.a. Google Book subset collection)

Schedule

Week	**Date**	**Topic**	**Due**
		Introduction	
1	Jan 12	Introduction to Data Science	Project signup
2	Jan 19	Data science, data storage, data sampling, and processing *Lab: Getting started with R, R data sampling and handling*	Group mini presentation
3	Jan 26	Text indexing, regression *Lab: Linear regression with R*	
4	Feb 2	Advanced regression analysis *Lab: Logistic regression with R, principal component analysis with R, t-test and ANOVA with R*	Group presentation
5	Feb 9	Latent variable and Sentiment analysis *Lab: Sentiment analysis via GraphLab (Dato) and R*	
6	Feb 16	SQL and NoSQL *Lab: MongoDB*	
7	Feb 23	Graph NoSQL and Graph Mining *Lab: Neo4J*	
8	Mar 1	Mid-term presentation	Group presentation
9	Mar 8	Text index, storage and retrieval *Lab: text preprocessing with R*	
10	Mar 15	**Spring Break – No Class**	
11	Mar 22	Ranking, Machine learning	
12	Mar 29	Machine learning *Lab: machine learning, text mining with R and GraphLab (Dato)*	
13	Apr 5	MapReduce and Advanced data processing *Lab: Hadoop*	
14	Apr 12	Information Recommendation *Lab: Music Recommendation via GraphLab (Dato)*	

15	Apr 19	Semantic Web and Ontologies for Data Science	
16	Apr 26	Final Class Presentation	Group presentation
17	May 3	Project Submission	Final Report due

Assignments

Students should complete the weekly assigned readings before each class.

Student will work on an individual or group midterm project for mid term.

During the semester, students will form groups and prepare to give one presentation (final) to the class on the chosen class project, based on one dataset. Students need to email presentations to instructors before the presentation day. At end of the semester, students will submit final report of their project and code. Project code can be deposited to code repositories such like Github or IU Github etc.

Course Grading

Course grading will be based on class participation, class presentations, homework assignments and final projects.

	Points
Class participation	10
Midterm	30
Final presentation (adjusted*)	30
Final report, code (adjusted*)	20
Teamwork	10
Total	100

Ownership of Student Work

This course may use course participation and documents created by students for educational purposes. In compliance with the Federal Family Educational Rights and Privacy Act, works in all media produced by students as part of their course participation at Indiana University may be used for educational purposes, provided that the course syllabus makes clear that such use may occur. It is understood that registration for and continued enrollment in a course where such use of student work is announced constitutes permission by the student. After such a course has been completed, any further use of student work will meet one of the following conditions: (1) the work will be rendered anonymous through the removal of all personal identification of the work's creator/originator(s); or (2) the creator/originator(s)' written permission will be secured.