Instructor Name: Weijia Xu

Course Name: Introduction to Methods and Tools for Large Scale Data Driven Analysis

Course Description: This course introduces students to the basic concepts and methods in data mining and software tools that can be used for practical data analysis. Three types of data mining tasks will be discussed: association analysis, cluster analysis, and classification. The purpose of this course is to teach students broad and applicable knowledge about data mining and how to use open source tools to carry out those analyses in practice. The class will focus on discussing the pros and cons among different methods and focus on how existing tools can be used rather than teaching students how to implement a particular method from scratch. The class will introduce open source data mining tools with WEKA and R for in-class demonstration as well as introduce large scale analysis framework, Hadoop and Spark, for conducting data mining tasks at large scale. Examples of using tools will be shown during the class. Basic knowledge about operating systems is required. The student should be capable of running applications and commands through command line interface. Programming experiences with Java/R is highly recommended to better understand some content of the course. The course will start with an overall introduction on data mining methods and open source tools in the first day. Then we will cover concepts and methods can be used in association, cluster and classification analysis in the following two sessions. In the last day, we will focus on introducing Apache Hadoop project and how it can be used for large-scale data mining.

Day 1:
- Course Overview
- Basic concepts and Definitions
- Types of data mining methods
- Data Analysis support tools with R/Weka (depends on student background)
- What happens when data keeps getting bigger

Day 2:
- Association rule mining using Apriori principle
- Collaborative filtering
- K-means clustering method
- Hierarchical clustering method
- Density based clustering
Day 3:
- Overview, concepts and goal
- Decision tree
- Naïve Bayes classifier
- Nearest neighbor classification
- Support Vector Machine classification
- Issues and Evaluations for classification methods

Day 4:
- Introduction to tools for large scale data analysis, Hadoop and Spark.
- Basic Concept in Hadoop How to get started
- Programming in Hadoop
- Introduction on Mahout library for data mining with Hadoop
- Difference between Hadoop and Spark.
- Introduction of MLlib with Spark for data analysis.