Final Exam
Date: May 11 Thursday
Location: CSB130
Time: 7:30AM ~ 9:30AM (120 minutes)
Total: 100 points
  40 points: 20 True/False questions: 2 points each
  60 points: 4 groups of questions with 5-6 sub questions per group

20% of the total score will be directly related to the quizzes.

Closed book, NO computer, No cheat sheets

Topics covered (Week 9 ~ Week 15)

1. Validation Methods
Analysis
  • k-fold cross validation
  • Leave-one-out cross validation
  • Confusion matrix
  • F1 score, sensitivity, miss rate, accuracy

2. Large scale analysis
  • Linear regression with stochastic gradient descent algorithm and running with MapReduce
  • k-Means clustering using MapReduce
  • Canopy clustering algorithm
  • Input and Output patterns in MapReduce
  • Recommendation Systems
    o Collaborative filtering using Jaccard/Cosine similarity
    o Normalizing rating
    o Computational complexity analysis of recommendation systems
    o Amazon’s item-to-item collaborative filtering

3. NoSQL Storage
  • Polyglot persistence
  • Column family based storage: Big Table
4. Dataflow management system: Pig Latin
   - Data types and cast
   - Relational operations
   - Skew reducing for order
   - Replicated, skewed, and merge join

5. Data exchange model (RESTful Web services)
   - 4 major HTTP methods for REST CRUD
   - Idempotent request
   - Managing errors