COMP SCI 3001: Introduction to Data Science

Instructor: Dr. Yanjie Fu, Assistant Professor of Computer Science
E-mail: fuyan@mst.edu
Phone: (573)341-4991
Blackboard: CS3001@Canvas
Office: 313, Computer Science Building

Text Book: "Introduction to Data Mining", by Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Addison Wesley, ISBN: 0-321-32136-7, 2005.

Recommended Readings

- "Introduction to Information Retrieval", by Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, Cambridge University Press, ISBN: 0-521-86571-9, 2008
- "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython", by Wes McKinney, O'Reilly Media, ISBN: 1-449-31979-3, 2012.
- "Data Science from Scratch: First Principles with Python 1st Edition", by Joel Grus, O'Reilly Media, ISBN: 149190142X, 2015.
- "Learning scikit-learn: Machine Learning in Python", Guillermo Moncecchi and Raul Garreta, Packt Publishing, ISBN: 1-783-28193-6, 2013.
- "Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems", by Aurélien Géron, O'Reilly Media, ISBN: 1-491-96229-1, 2017.

Course Description and Objectives:

The objectives of this course are to introduce the fundamental concepts and lifecycle of data science, including (1) python programming for data science, (2) web crawler for data collection, (3) data indexing, retrieval models, and search engine for complex datasets, (4) data mining techniques including linear regression, regression with regularizations, classification, clustering, (5) advance data mining techniques such as learning to rank and recommender systems, and (5) data science system development.

Prerequisites:

A "C" or better grade in COMP SCI 2300, in COMP SCI 2500, and in one of STAT 3113/3115/3117/5643

Programming Languages:

R and Python are two major languages for data science. To learn Python, check out: https://www.programiz.com/python-programming/first-program To learn R, check out:

https://www.statmethods.net You are suggested to use R or/and Python to complete the homework assignments.

Grading Policy:

4-10 homework assignments:	40%
Exam	20%
Attendance and participation in discussion:	10%
Project/Presentation/Paper	30%

All grading will be determined based on a scale of: A: [90-100%], B: [80-90%), C: [70-80%), D: [60- 70%), F: [0-60%). Final grades in the course may be curved at the instructor's discretion.

Schedule (non-restrictive):

1. Introduction
2. Understanding Data
3. Python Programming for Data Science
4. Crawling and Parsing Data
5. Inverted Index, Query Processing, and Search Engine
6. Data Preprocessing
7. Classification
   - Decision Trees
   - Nearest Neighbor Based Classifiers
   - Naïve Bayesian Classifiers
   - Evaluation of Classification
8. Clustering
   - K-Means Clustering
   - Hierarchical Clustering
   - Density-based Clustering
   - Evaluation of Clustering
9. Regression
   - Linear Regression
   - Logistic Regression
   - Lasso Regularization (L1)
   - Ridge Regularization (L2)
10. Recommender Systems
    - Content-based Recommender Systems
    - User-based Collaborative Filtering
    - Item-based Collaborative Filtering
    - Matrix Factorization
    - Evaluation of Recommender Systems
11. Ranking
    - Point-wise Learning To Rank

- Pairwise-wise Learning To Rank
- Listwise-wise Learning To Rank
- Evaluation of Ranking