

Lecture #0 | Course announcement

SE377 Introduction to Big Data Analysis and Visualization (2017)

Prof. Min-gyu Cho

Target audience

- Junior and senior
- Prerequisite courses
 - Programming
 - Probability and Statistics
- Recommended courses (past or future)
 - Database
 - Machine learning
 - Algorithms
- Who has interest in
 - Obtaining (hidden) insights from data
 - Working at any organization which make decision based on data (incl. startups)
 - Learning LOTS of things

Objectives

- Data science
 - Understand what is data science
 - Learn data science processes and methodologies
 - Build insights through analysis of real data
- Data engineering (to be determined)
 - Understand concept of data engineering
 - Practice small example of data engineering (e.g., map-reduce with hadoop)

Topics which will NOT be covered in this class

- Theories statistics and machine learning
 - We will use statistics and machine learning algorithms, but will not cover the theories/proofs/etc
- Basic python programming skills
 - Although numpy, matplotlib, etc. will be reviewed
- Most of data engineering
 - ETL processes (extract, transfer and loading)
 - Data engineering tools (e.g., hadoop, spark, kafka)
- Interactive Data Visualization (using javascript libraries such as d3.js)

Information on lecture rooms, instructors and TAs

- Hours: M 1630-1800, Th 1030-1200
- Lecture Room: E7 L31
- Instructor: Min-gyu Cho, mingyu.cho@dgist
- TA: Hyunlim Yang, lims1@dgist.ac.kr
- Office hours will be announced later

Course materials

- Course home: https://github.com/chomg/se377_2017
- Q&A: bulletin board on LMS (Unofficial)
- Textbooks
 - Python for Data Analysis
 - Data Science from Scratch
 - Doing Data Science
- Reading list
 - Lecture notes (and supplementary materials)
 - LOTS of blogs/tutorials/books/...
- Environment for data analysis
 - python: python + jupyter notebook + pandas + matplotlib (recommended*)
 - R: R studio + dplyr + ggplot
 - Or any other data analytics tools/libraries of your own choice

* anaconda includes all these packages and is recommend

Assessment

Breakdown*

Area	Weight
Assignments	30%
Proposal	15%
Final Project	30%
Review/comments	15%
Class participation	10%

Letter grade*

Letter grade	Score
A	≥ 85
B	≥ 70
C	≥ 50
D	< 50
F	If necessary

* Subject to change; see the course homepage for the final breakdown/letter grade policies

Term project

- Team project: 3 students/team
 - Team formation?
- Each team will present twice
 - Proposal (3-5 min)
 - Problem definition (i.e., objective of data analysis)
 - Introduction of data sets
 - Proposed approach for data analysis
 - Final presentation (5-7 min)
 - Analytic methodology
 - Insights from analysis
 - Lessons from data analysis



ANY QUESTIONS?