DS-210: Programming for Data Science

Lecture 1: Class overview. Survey. Basics of data analysis.

Teaching Staff and Office Hours

Instructor: Krzysztof Onak

- Email: konak@bu.edu
- Office hours: Wed 4:30–6:30pm @ MCS 138N (or adjacent common space)

TA: Vedaant Tiwari

- Email: vedaant@bu.edu
- Office hours: TBD (Mon/Fri) 4–6pm @ TBD

Context

- New course
- Second programming course in the data science major, after DS-110
- Other students are welcome too!

Course Content

Part 1: Python for data science

- Continuation of DS-110
- Many topics in machine learning
    - basic data science pipeline
    - various algorithms: decision trees, linear regression, clustering
- Also: various packages, software documentation

Course Content

Part 1: Python for data science

- Continuation of DS-110
- Many topics in machine learning
    - basic data science pipeline
    - various algorithms: decision trees, linear regression, clustering
- Also: various packages, software documentation

Part 2: Rust

- What is it? Why Rust?
    - relatively new language
    - as efficient as C/C++
    - memory–safe
- Additional concepts:
    - Programming languages
    - Data stuctures and algorithms
- Some highlights:
    - Calling Rust code from Python
    - Basic multithreading
- Small final project: analyze a social network

Prerequisites

- **Python!!!**
- Jupyter Notebook
- NumPy
- Pandas
- Matplotlib

See self study resources!

Prerequisites

- **Python!!!**
- Jupyter Notebook
- NumPy
- Pandas
- Matplotlib

See self study resources!

- Big-Oh notation
- Version Control (git)

Final Grade

- Weekly homeworks (25%)
- Midterm and final (20% each)
- Graph analysis project in Rust
    - Proposal (5%)
    - Final project (20%)
- Class participation (10%)

Meetings

Lectures

- Time: Mon/Wed/Fri 12:20-1:10pm
- Location: MCS B37

Discussion

- Time:
    - Group 1: Wed 2:30–3:20pm
    - Group 2: Wed 3:35–4:25pm
- Location: CGS 111B

Meetings

Lectures

- Time: Mon/Wed/Fri 12:20-1:10pm
- Location: MCS B37

Discussion

- Time:
  - Group 1: Wed 2:30–3:20pm
  - Group 2: Wed 3:35–4:25pm
- Location: CGS 111B

Online

Website: https://onak.pl/ds210

- Class info, lecture slides
- Homework

Piazza: https://piazza.com/bu/spring2022/ds210/home

- Announcements and additional information
- Questions and discussions

Gradescope

- Homework, project, project proposal submissions

Quick survey

- Office hours: Monday or Friday? (apart from Wednesday)
    - Weekly homeworks due Wednesday
- Any prerequisites you are least comfortable with?
- What programming languages do you know? (basic/intermediate/advanced)

Two basic approaches

- Classical programming:
  - write a program
  - run it
  - see the result

Two basic approaches

- Classical programming:
  - write a program
  - run it
  - see the result

- Interactive notebooks:
  - long tradition: Mathematica, Matlab, ...
  - mixing code, description, visualisation
  - Jupyter Notebook + Python can be used as a replacement of all of this
- Presentations are possible: these slides use `RISE`

Types of Data Analysis

Predictive

- want to learn future based on the past

Types of Data Analysis

Predictive

- want to learn future based on the past

Descriptive

- understand the past