HSS 611: Programming for Digital Humanities and Computational Social Sciences

Fall 2023
Mon Wed 9:00–10:15AM
N4 1309, School of Digital Humanities and Computational Social Sciences

Instructor: Taegyoon Kim, Ph.D. in Political Science and Social Data Analytics

- Email: taegyoon.academic@gmail.com
- Office hours: Mon Wed 10:15–11:00AM & By appointment
- Personal webpage: https://taegyoon-kim.github.io

Teaching Assistant: Minsang Namgoong

- Email: whovian@kaist.ac.kr

Course Overview: Python has gained tremendous popularity as a programming language owing to its versatility in handling various data types, robust data analysis libraries, widespread adoption in machine/deep learning communities, and reliable support for tasks like web scraping and online data extraction. This introductory Python course is designed to cater to the needs of researchers in digital humanities and computational social sciences. It equips students with the fundamentals of Python and key skills for applied computational and quantitative research. The course sessions will feature a blend of theory and practical application. The instructor will deliver lectures and guide students through hands-on coding exercises during class. In addition, students will have ample opportunities to apply and adapt their newly acquired knowledge to their individual research.

Readings: Students are not required to purchase any books for this course as it will not follow any textbook strictly. However, here are useful, freely available books:

- McKinney, W., 2022. Python for data analysis. "O'Reilly Media, Inc." [link]
- VanderPlas, Jake. 2016. Python data science handbook: Essential tools for working with data. "O'Reilly Media, Inc." [link]

Major Tasks: Students are expected to complete the following tasks.

- *Attendance:* Students are required to attend all lectures unless they have exceptional circumstances that they have discussed with the instructor beforehand. Worth 10% of the final grade.

- *Tutorial presentation:* Students will deliver three tutorial presentations, each lasting approximately 15–20 minutes. There are a total of 12 substantive themes in the course. Students will choose one (or two) theme from the first 6 weeks (covering fundamentals) and one (or two) from the next 6 weeks (covering advanced usages). During the presentations, students will showcase the implementation of techniques covered in that week using their own data, preferably related to their research projects. They will guide the entire class through their script, explaining each code line in detail. The presentations serve a dual purpose: firstly, to provide non-presenting students with a deeper understanding of the techniques covered in that week, and secondly, to offer presenting students an opportunity to apply and adapt those techniques in their own data analysis tasks. Students will present their tutorials in Wednesday classes. Please read the instruction and sign up here. Worth 30% of the final grade. Be sure to provide the class with access to your script and data so that the whole class can follow as you walk through your tutorial.

- *Final paper:* Students must complete an original research paper that involves a significant amount of data work, and present it during the final week of the course. The primary objective is to guide students in constructing a comprehensive data-driven research pipeline, covering data collection, pre-processing, analysis, and visualization. Although students have the freedom to choose their preferred topic, they are required to work with data applying a broad range of skills learned in the course. By applying and adapting these skills for their own research, students will not only demonstrate their mastery of programming fundamentals but also gain a practical understanding of how skills learned in class can be effectively utilized in real-world data-driven research scenarios. The research paper should focus on detailed methodological accounts of data collection, cleaning, manipulation, analysis, and visualization, minimizing discussions of theories or prior literature. Students will deliver a presentation (10–15 minutes) in the final week of the course and email their papers to the instructor, along with their scripts and data, by the end of the semester. The paper and presentation are worth 30% and 10% of the final grade, respectively.

- *Problem set:* There will be several problem sets assigned throughout the course. Worth 20% of the final grade.

Grading Scale: Grade values will not be rounded. That is, any grade value that is greater than or equal to 'Lower' and less than 'Upper' will receive the respective grade.

Grade	Lower	Upper
A+	90	101
A_0	87	90
A-	84	87
B+	81	84
B_0	78	81
B-	75	78
C+	72	75
C_0	69	72
C-	66	69
D+	63	66
D_0	60	63
D-	57	60
F	0	57

Course Schedule:

Week 1. Introduction (Aug 28 & 30)

- Course Overview, Logistics, etc.
- Installing Python, Python packages
- Learning about IDEs
- Reading Python documentation
- Mathematical operations, basic data types and variables

Week 2. Branching and iteration (Sep 4 & 6)

- Conditional Statements
- Loops
- Comparison and logical operators

Week 3. String manipulation, tuples, and lists (Sep 11 & 13)

- String and list methods
- Aliasing
- Mutability
- Cloning
- List Comprehensions

Week 4. Functions and modules (Sep 18 & 20)

- Defining and calling functions
- Parameters and arguments
- Introduction to modules, importing
- Scopes, abstraction
- Raising Exceptions, warnings

Week 5. Dictionaries, sets, and recursion (Sep 25 & 27)

- Dictionaries, sets
- Defaultdict
- Dictionary comprehensions
- Set comprehensions
- Recursion

Week 6. NumPy and Pandas I (Oct 2 & 4)

- NumPy ndarrays
- NumPy indexing/slicing
- NumPy Mathematical and Statistical Methods
- Getting started with Pandas

Week 7. Proposal presentation (Oct 9 & 11)

- Tue: No class on Oct 9 (Hangeul Day)
- Thu: Proposal presentation

Week 8. No class (Oct 16 & 18)

- Mid-term examination period

Week 9. NumPy and Pandas II (Oct 23 & 25)

- Cleaning
- Join, combine, reshaping
- Grouping, aggregation

Week 10. Visualization (Oct 30 & Nov 1)

- Basic plot types
- Customization and styling
- Subplots and layouts
- Advanced plot types
- Saving and exporting

Week 11. Interacting with API (Nov 6 & 8)

- The requests library
- GET requests
- JSON files
- XML files

Week 12. Web scraping I (Nov 13 & 15)

- Understanding the HTML structure
- Scraping static web pages
- The BeautifulSoup library

Week 13. Web scraping II (Nov 20 & 22)

- Understanding dynamic web pages
- The Selenium library

Week 14. NLP / Text-as-data (Nov 27 & 29)

- Text preprocessing

- Representation of text
- Regular expressions

Week 15. Machine learning (Dec 4 & 6)

- Linear Regression
- Logistic Regression

Week 16. Final presentation (Dec 11 & 13)

- Tue: Final presentation I
- Thu: Final presentation II

Instruction Mode: The instruction mode is in-person. However, depending on the public health challenges caused by the COVID-19 pandemic, some classes might be offered remotely. Any change to the mode of instruction will be announced in advance.

Attendance: Consistent attendance is essential for this course. You are permitted to miss a maximum of two classes without any impact on your grade. However, should you exceed this limit and miss additional classes, a deduction of two points will be applied to your attendance score for each absence. There is no obligation to notify the instructor in advance if you cannot attend a class unless you are presenting in that class. Please be aware that arriving more than fifteen minutes late will be considered as an absence for that particular class.

Email Policy: I try to respond to emails promptly, typically within two business days. If you have complex questions or need an in-depth discussion, I encourage you to attend my office hours.

Office Hours: I welcome all students to attend my office hours for discussions related to course content and learning strategies. If you need to set up a meeting outside my office hours, send me an email with your availability, and we will arrange a mutually convenient time to meet.

Late Submission Policy: Late submissions will incur a penalty of 20% for each day (rounded up) beyond the due date

Academic Integrity: As students at KAIST, you are entrusted with upholding the utmost standards of academic integrity. Academic honesty is paramount, and any form of misconduct is strictly prohibited. In the event of suspected misconduct, our class adheres to the established policy of KAIST. All such incidents are promptly reported to the dean of the Department of Humanities and Social Sciences to ensure a fair and transparent resolution.

Grade Dispute Memo: Should any student want to contest a grade received on the response note, theme presentation, and team project, they have the option to do so by submitting a written memo. The memo must specifically outline the reasons why the assignment warrants a different grade. The memo must then be submitted within seven days of receiving the grade. Submissions should be sent via email for consideration.

Syllabus Change Policy: This syllabus is a guide, and every attempt will be made to provide an accurate overview of the course. However, circumstances and events may make it necessary for the instructor to modify the syllabus during the semester and may depend, in part, on the progress, needs, and experiences of the students.