COURSE SYLLABUS

CSC17104 – PROGRAMMING FOR DATA SCIENCE

(Last update: 25/09/2022)

1 GENERAL INFORMATION

Course name:	Programming for data science
Course name (in Vietnamese):	Lập trình cho khoa học dữ liệu
Course ID:	CSC17104
Knowledge block:	
Number of credits:	4
Credit hours for theory:	45
Credit hours for practice:	30
Credit hours for self-study:	90
Prerequisite:	
Prior-course:	
Instructors:	Bùi Tiến Lên Trần Trung Kiên Phạm Trọng Nghĩa Lê Nhựt Nam

2 COURSE DESCRIPTION

The course is designed to provide students with knowledge and skills about how to use data science tools: Linux commands, Git and Github, Conda, Jupyter Notebook, Markdown, Python, Matplotlib, Numpy, Pandas. We will learn these tools at a quite deep level, and we will learn them in the context of data science process: ask a meaningful question which can be answered with data, collect data, explore data, preprocess data, analyze data, communicate results.

3 COURSE GOALS

At the end of the course, students are able to

ID	Description	Program LOs
G1	Use Linux commands and Conda to set up working environment	1.3.7
G2	Use Python with built-in libraries to do a data science process	1.2.1 1.7.3
G3	Use Python with Numpy library to do a data science process with numerical data	1.2.1 1.7.3
G4	Use Python with Pandas library to do a data science process with tabular data	1.2.1 1.7.3
G5	Use Python with Matplotlib library to visualize data	1.2.1 1.7.3
G6	Use Jupyter Notebook and writing skill to document the whole data science process	1.3.7 2.3
G7	Use Git, Github and teamwork skill to control versions and collaborate with others	1.3.7 2.2

4 COURSE OUTCOMES

CO	Description	I/T/U
G1.1	Use Conda to setup/update tools from a specification file; use Linux commands to organize files and folders for a data science project	T
G2.1	Use Python built-in data structures and Python built-in operations to explore - preprocess - analyze data	UT

G3.1	Use Numpy arrays and Numpy built-int operations to explore - preprocess - analyze numerical data	T
G4.1	Use Pandas dataframe and Pandas built-int operations to explore - preprocess - analyze tabular data	T
G5.1	Use Matplotlib library to visualize data during data science process	T
G6.1	Use Jupyter Notebook markdown cell to organize a data science document into sections, use code cell to code and display results	T
G6.2	Use writing skill to write and code clearly	U
G7.1	Use Git & Github to control versions and collaborate with others	T
G7.2	Use teamwork skill to collaborate with others	U

5 TEACHING PLAN

THEORY

ID	Topic	Course outcomes	Teaching/Learning Activities (samples)	Assessments
1	Introduction to the course; how to use tools in general	G1.1 G2.1 G3.1 G4.1 G5.1 G6.1 G7.1	Lecturing, demo, Q&A	FP

2	Linux commands, Conda	G1.1	Lecturing, demo, Q&A	HW1, HW2, HW3, FP
3	Git & Github	G7.1	Lecturing, demo, Q&A	FP
4	Jupyter Notebook, Markdown	G6.1	Lecturing, demo, Q&A	HW1, HW2, HW3, FP
5	Python	G2.1	Flipped classroom Lecturing, demo, quiz, Q&A	HW1, FP
7	Numpy	G3.1	Lecturing, demo, quiz, Q&A	HW2, FP
8	Pandas, Matplotlib	G4.1 G5.1	Lecturing, demo, quiz, Q&A	HW3, FP

LABORATORY

ID	Topic	Course outcomes	Teaching/Learning Activities (samples)	Assessments
1	Setup		Q&A on Moodle	
2	HW1 guide	G1.1 G2.1 G6.1 G6.2	Q&A on Moodle	HW1
3	HW2 guide	G1.1	Q&A on Moodle	HW2

		G3.1 G6.1 G6.2		
3	HW3 guide	G1.1 G4.1 G5.1 G6.1 G6.2	Q&A on Moodle	HW3

6 ASSESSMENTS

ID	Topic	Description	Course outcomes	Ratio (%)
HW	**Homework**			**50%**
HW1	Python	Use Python (with built-in libraries) to do a data science process	G1.1 G2.1 G6.1 G6.2	50/3%
HW2	Numpy	Use Numpy to do a data science process with numerical data	G1.1 G3.1 G6.1 G6.2	50/3%
HW3	Pandas + Matplotlib	Use Pandas + Matplotlib to do a data science process with tabular data	G1.1 G4.1 G5.1 G6.1 G6.2	50/3%

FP	**Final Project**			**50%**
FP	Using tools to do a data science process to answer your own questions	Find a public dataset, explore data, ask meaningful questions, preprocess + analyze to answer each question, communicate results with Teacher Do it in a team	G1.1 G2.1 G3.1 G4.1 G5.1 G5.2 G6.1 G6.2 G7.1 G7.2	50%

7 RESOURCES

Textbooks

- Jake VanderPlas, Python Data Science Handbook, https://jakevdp.github.io/PythonDataScienceHandbook/

Others

- Brandon Rhodes, Stopping to Sharpen Your Tool, https://www.youtube.com/watch?v=I56oFTm9UlE
- MIT, The Missing Semester, https://missing.csail.mit.edu/
- Berkeley, Computational and Inferential Thinking, https://inferentialthinking.com
- Conda document, https://docs.conda.io/projects/conda/en/latest/index.html
- Git document, https://git-scm.com/doc
- Jupyter Notebook document, https://jupyter-notebook.readthedocs.io/en/stable/
- Python document, https://docs.python.org/3/
- Matplotlib document, https://matplotlib.org/stable/contents.html
- Numpy document, https://numpy.org/doc/

- Pandas document, https://pandas.pydata.org/docs/

8 GENERAL REGULATIONS & POLICIES

- All students are responsible for reading and following strictly the regulations and policies of the school and university.
- Students who are absent for more than 3 theory sessions are not allowed to take the exams.
- For any kind of cheating and plagiarism, students will be graded 0 for the course. The incident is then submitted to the school and university for further review.
- Students are encouraged to form study groups to discuss on the topics. However, individual work must be done and submitted on your own.