Proposal for a New Course

Department of Computer Science and Engineering
Indian Institute of Technology, Kanpur

Course number: ESC202
Course title: Fundamentals of Data Engineering
Course prerequisites: ESC111 and ESC112 (waived for PG students)
Course credits: [12] (3-0-3-0)
Course duration: Full semester
Course type: Institute Core Course
Proposing instructor: Purushottam Kar
Other faculty members interested in teaching the course: Nisheeth Srivastava, Sruti Srinivasa Ragavan, Piyush Rai
Other departments interested in the proposed course: None

Course description:
a. *Objectives:* Proper handling of data is an essential prerequisite for the disciplines of AI, ML, and data science. This course will introduce data collection, curation and comprehension techniques using Python libraries. The course will also offer hands-on experience in conducting small-scale experiments involving digital data sources.
b. *Logistics:* The course will serve as a DC for IS UG students (BT, double major, digital IS minor), and as a DE for PG students (MT, MS, PhD) of the CSE and IS departments and CSE UG students (BT, double major, ML minor). The course may be offered one or more times every academic year depending on demand and availability of resources.
c. *Content:* There will be an equivalent of 39 lectures of 50 minutes each and 13 labs of 3 hours each. A weekly breakup of lecture and lab content is given below
d. *Evaluation:* Evaluation will use a combination of graded lab exercises, lab exams, take- home assignments and projects, and traditional sit-down quizzes and exams.

Weekly breakup of content: Numbers in square brackets [] against each topic indicate the number of lectures/labs for that topic.
Lecture content (39 lectures):
1. Data programming basics [15]:
a. Code hygiene: debugging, comments and documentation, writing requirements
b. Python basics: name-object vs variable-value, mutable and immutable types – tuples, lists, dictionaries, sets, list comprehension, indexing, broadcasting
c. Resource usage: API docs, community fora (Stack Overflow, GitHub), IDEs and dev envs (VSCode, Jupyter, Colab, Binder), AI assistants (ChatGPT, Copilot)

2. Data storage, processing, visualization, and analysis [15]:
a. Formats: introduction (e.g. CSV, Excel, parquet, SQLite, JSON, sockets)

b. Processing: ETL (extract, transform, load), data scraping, data preprocessing
c. Visualization: plot types (line, scatter, bar, stacked), error bars and confidence intervals, aesthetics, accessibility (colorblind-friendly color schemes, alt-text)
d. Analysis: ML primitives (e.g. classification, regression, clustering, dimensionality reduction, anomaly detection) -- brief overview only Use of popular Python libraries such as numpy, sklearn, scipy, pandas, openpyxl, sqlite3, matplotlib, seaborn, plotly, bokeh to implement the above
3. Data-based app dev [9]: pipelines, data transformers, trigger/event-based programming, FastAPI and JS tools to build webapps for data visualization, data-enabled services etc.

Lab content (13 labs):
1. Python basics [5]
a. Familiarization with libraries and IDEs
b. Creative use of list comprehension, datatypes
2. Data storage, processing, visualization, and analysis [5]
a. Perform data cleanup, missing data imputation, visualization on static data from a spreadsheet or SQL database, or live data from an online source
b. Implement supervised and unsupervised data analysis on cleaned data
3. App development [3]
a. Build a web app offering ML-as-a-service
b. Build a live face detection system piping camera feed into a pretrained AI model

Lab equipment: PCs with internet connectivity, attached/integrated webcams, and appropriate software (browsers, Python runtime with libraries) will be needed.

Short summary for inclusion in the Courses of Study booklet: the course will introduce Python-based techniques used to collect, curate, and comprehend data.

Textbook: There will be no textbook for this course.

Course proposer: Prof Purushottam Kar
Date:

Convener DPGC:
Date:

The course is approved/not approved

Chairperson, SPGC
Date: