Department of Computer Science and Software Engineering
Concordia University

COMP 499 — Introduction to Data Analytics
Course Outline
Summer 2018 Semester 1

Web page: `http://users.encs.concordia.ca/~gregb/home/comp499-s2018.html`

Course Calendar Description

COMP 499 Introduction to Data Analytics (4 credits) Data wrangling: Extract-Transform-Load, data cleaning, data integration. Data analytics: description, prescription, modeling, simulation, optimization. Discipline specific datasets, tools, and analysis approaches. Python ecosystem: language; libraries `numpy`, `scipy`, `pandas`; Jupyter environment; `R` data analysis package. A project. *Prerequisite*: COMP 233 or ENGR 371; COMP 352.

Course website: Check frequently the website for announcements, course material, assignments, etc.

IMPORTANT NOTE: In the event of extraordinary circumstances beyond the University's control, the content and/or evaluation scheme in this course is subject to change.

Course Objectives

Big Data and Data Analytics has permeated into every industry, government, and business function. The future will need data-driven approaches for all fields of human endeavour. The challenges in handling massive datasets and performing the computations for analysis to transition from raw data to information to knowledge and to application are many and varied. Data analytics is at the core of interdisciplinary collaboration with the social sciences, humanities, health and life sciences, ecology and the environment, culture and heritage, and engineering.

The aim of this course is to introduce students to Python programming language and related tools for data analytics; and to expose them to a broad range of data analysis problems across a range of disciplines.

Recommended Books:

(1) *Data Science from Scratch: First Principles with Python*, by Joel Grus, O'Reilly, 2015.

(2) *Data Crunching: Solve Everyday Problems using Java, Python and More*, by Greg Wilson, The Pragmatic Bookshelf, 2005. This book is out of print, but can be found in the Library.

(3) *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, by Hadley Wickham, Garrett Grolemund, O'Reilly Media, 2016.

(4) *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, by William McKinney, O'Reilly Media, 2012.

(5) Material on available data resources and data analysis problems from a range of disciplines will be provided.

Grading Scheme

The grade will be determined by the following components:

20% Assignments
30% Course project
15% Midterm
35% Final Examination

Course Components

This course is a 4-credit course with 2.5 class hours of lecture and 2 hours of lab.

This is an accelerated six-and-a-half week semester, so you should expect to **average** a total of 20–24 hours per week on this course.

Lectures

Lecture 1: Introduction and overview
Lecture 2: Python introduction using running examples
Lecture 3: Data wrangling using examples
Lecture 4: Python analytics and visualisation: `numpy`, `scipy`, `matplotlib` using examples
Lecture 5: Statistical analysis: `R` package using running examples
Lecture 6: High-dimensional datasets: `pandas`
Lecture 7: Midterm
Lecture 8: Discipline: supply chain management
Lecture 9: Discipline: recommender systems
Lecture 10: Discipline: geospatial data, smart cities
Lecture 11: Discipline: social networks
Lecture 12: Project presentations
Lecture 13: Project presentations

Laboratory

The lab session for the course will address practical issues with the Python ecosystem, provide a series of running examples and assignments for students to do hands-on data analytics, and provide students with data resources from a range of disciplines.

Attendance at the lab sessions is compulsory.

The Python ecosystem to be used is the anaconda distribution of Jupyter together with Python v3.x with its standard libraries including `numpy`, `scipy`, `matplotlib`, and `pandas`, and R.

Assignments

Assignment 1 is to complete the data wrangling task of **Lab 2** as a Jupyter notebook, show the work to the TA, and submit the notebook to the electronic submission system `https://fis.encs.concordia.ca/eas/` under `programming_assignment 1`.

Assignment 2 is to complete the exploratory data analysis task using Python in **Lab 4** and using R in **Lab 5**, show the work to the TA, and to submit one (combined) Jupyter notebook to the electronic submission system `https://fis.encs.concordia.ca/eas/` under `programming_assignment 2`.

Project

Each student is to investigate a selected large dataset; carry out data wrangling, and exploratory data analysis; and develop models and stories to communicate their findings. The student will present their work in class at the end of the course.

Each student will submit a report as a pdf file, based on their Jupyter notebook, to the electronic submission system `https://fis.encs.concordia.ca/eas/` under `project 1`.

The choice of dataset, and the questions to be explored using the dataset, should be discussed with the course instructor immediately after the midterm examination.

Examinations

The midterm exam will be a one-hour exam covering data wrangling, exploratory data analysis, and the use of Python, R and their libraries. It will cover both theory and practice. It may have a mixture of true-false questions, multiple choice questions, short answer questions, and programming tasks.

The final examination will be a three-hour examination covering all theoretical and practical parts of the course. It may have a mixture of true-false questions, multiple choice questions, short answer questions, and programming tasks.

Academic Honesty

Violation of the Academic Code of Conduct in any form will be severely dealt with. This includes copying (even with modifications) of program segments. You must demonstrate independent thought through your submitted work. The Academic Code of Conduct is available at: http://www.concordia.ca/students/academic-integrity/code.html

Graduate Attributes

As part of both the Computer Science and Software Engineering program curriculum, the content of this course includes material and exercises related to the teaching and evaluation of graduate attributes. Graduate attributes are skills that have been identified by the Canadian Engineering Accreditation Board (CEAB) and the Canadian Information Processing Society (CIPS) as being central to the formation of Engineers, computer scientists and information technology professionals. As such, the accreditation criteria for the Software Engineering and Computer Science programs dictate that graduate attributes are taught and evaluated as part of the courses. This particular course aims at teaching and evaluating several graduate attributes. The following is a description of these attributes, along with a description of how these attributes are incorporated in the course.

(GA1) A Knowledge Base for Engineering: *Demonstrated competence in university level mathematics, natural sciences, engineering fundamentals, and specialized engineering knowledge appropriate to the program.*

(GA2) Problem analysis: *An ability to use appropriate knowledge and skills to identify, analyze, and solve complex engineering problems in order to reach substantiated conclusions.*

(GA5) Use of Engineering tools *is the ability to create, select, apply, adapt, and extend appropriate techniques, resources, and modern engineering tools to a range of engineering activities, from simple to complex, with an understanding of the associated limitations.*

(GA6) Individual and Team Work: *An ability to work independently and as a member and leader in diverse teams and in multi-disciplinary settings.*

(GA7) Communication Skills: *An ability to communicate complex engineering concepts within the profession and with society at large. Such abilities include reading, writing, speaking and listening, and the ability to comprehend and write effective reports and design documentation, and to give and effectively respond to clear instructions.*