STAT 5870 Big Data Analysis Using Python

(Summer I 2021)

Instructor:

Name: Kevin H. Lee
Office Location: 5504 Everett Tower
Office Hours: Set up an appointment via email
Email: k.lee@wmich.edu

Course Description:

This course has three main goals: (i) students to learn the basics of Python; (ii) students to learn how to use Python as a tool to effectively store, manipulate, and gain insight from data; (iii) students to learn recent machine learning techniques. The usefulness of Python for data science stems mainly from the large and active ecosystem of third-party packages. Therefore, students will also learn how to use popular packages: NumPy, Pandas, Matplotlib, Seaborn and Scikit-Learn.

In this course, we will use Spyder which is a free integrated development environment (IDE) that is included in Anaconda.

Anaconda can be downloaded from `https://www.anaconda.com/download`.

Prerequisite:

STAT 5850 or CS 5821 with a grade of "B" or better or instructor approval and a suitable laptop.

Course Schedule (Asynchronous Online):

May 10, 2021 – June 30, 2021

Textbook (Recommended):

Python Data Science Handbook, Jake VanderPlas, O'Reilly Media, Inc., 2017.

Reference Books:

Illustrated Guide To Python 3, Matt Harrison, Treading on Python Series, 2017.
Hands-On Machine Learning with Scikit-Learn & TensorFlow, Aurelien Geron, O'Reilly Media, Inc., 2017.

Grading

Lab Assignment: 20%
Homework: 25%
Midterm Exam: 30%
Final Project: 25%

Letter Grades

90% - 100%	A
85% - 89%	BA
80% - 84%	B
75% - 79%	CB
70% - 74%	C
60% - 69%	DC
50% - 59%	D
0% - 49%	E

Course Materials:

Course materials will be uploaded to Elearning every week on Tuesday, 2:00 PM (ET).

Lab Assignment:

Lab assignment will be uploaded to Elearning every week on Tuesday, 2:00 PM (ET) until the fourth week of the semester. Hence, there will be total four lab assignments. The due date for the lab assignment is always same week Sunday, 11:59 PM (ET). You are encouraged to work in groups, but copying the answer is not allowed. That is, each student is expected to write up the answers in his/her own words, even if the problems have been solved in collaboration with others. Late lab assignment will not be accepted.

Homework Assignment:

Homework will be uploaded to Elearning every week on Tuesday, 2:00 PM (ET) except the midterm week. Hence, there will be total six homework. The due date for the homework is always same week Sunday, 11:59 PM (ET). You are encouraged to work in groups, but copying the solution is not allowed. That is, each student is expected to write up the solutions in his/her own words, even if the problems have been solved in collaboration with others. Late homework will not be accepted.

Midterm Exam:

Midterm exam will be given online on Tuesday, June 8, 2:00 PM – 4:30 PM (ET). If, due to sickness, family emergency, time zone difference etc, you need to miss midterm exam, you should inform me (e.g., by email) BEFORE the midterm date in order to make alternate arrangements. There will be NO make-up midterm for those who miss it without pre-notification.

Final Project:

The goal of the final project is to learn more about text analytics applications (sentiment analysis, text classification, topic modeling, etc.). Pick a real text data set for which you believe there are interesting questions to answer. Choose one or more text analytics applications, and study the methods or tools people use and apply to your data set using Python. More details will be announced in the midterm week.

Course Rules:

Students are responsible for all email announcements and notices posted in Elearning.

Elearning:

Lecture materials (e.g. lecture slide, lecture video, Python code and data file), lab assignment, and homework will be uploaded to Elearning. Moreover, Elearning will be used to collect both lab assignment and homework. Your grade will also be updated in Elearning.

Disabilities:

If you have any sort of disability such that you require accommodations to participate in class, take exams, etc, let me know so that we can make the appropriate arrangements. For more information, see `http://www.wmich.edu/disabilityservices`.

Incompletes:

Incompletes will be only be given according to University and Departmental policy. An incomplete is not a substitute for a failing grade; they are only given only after completing major portion of the coursework with a passing grade, and circumstances beyond your control prevent you from completing the coursework.

Academic Integrity:

Students are responsible for making themselves aware of and understanding the academic policies and procedures in the Undergraduate and Graduate Catalogs that pertain to Academic Honesty. These policies include cheating, fabrication, falsification and forgery, multiple submission, plagiarism, complicity and computer misuse. [The policies can be found at `http://catalog.wmich.edu` under Academic Policies, Student Rights and Responsibilities.] If there is reason to believe you have been involved in academic dishonesty, you will be referred to the Office of Student Conduct. Students will be given opportunity to review the charge(s). If you believe you are not responsible, you will have the opportunity for a hearing. Students should consult with your instructor if you are uncertain about an issue of academic honesty prior to the submission of an assignment or test.

Students and instructors are responsible for making themselves aware of and abiding by the "Western Michigan University Sexual and Gender-Based Harassment and Violence, Intimate Partner Violence, and Stalking Policy and Procedures" related to prohibited sexual misconduct under Title IX, the Clery Act and the Violence Against Women Act (VAWA) and Campus Safe. Under this policy, responsible employees (including instructors) are required to report claims of sexual misconduct to the Title IX Coordinator or designee (located in the Office of Institutional Equity). Responsible employees are not confidential resources. For a complete list of resources and more information about the policy see (`http://www.wmich.edu/sexualmisconduct`).

In addition, students are encouraged to access the Code of Conduct, as well as resources and general academic policies on such issues as diversity, religious observance, and student disabilities:

- Office of Student Conduct (`http://www.wmich.edu/conduct`)
- Division of Student Affairs (`http://www.wmich.edu/students/diversity`)

- University Relations Office (`https://wmich.edu/universityrelations`)
- Disability Services for Students (`http://www.wmich.edu/disabilityservices`)

Tentative Course Schedule:

Throughout the semester, this schedule may be adjusted slightly according to the pace of the class.

Date	Topics
Week 1 (05/11 – 05/16)	Introduction, Python Basics
Week 2 (05/18 – 05/23)	Introduction to `Numpy`
Week 3 (05/25 – 05/30)	Data manipulation with `Pandas`
Week 4 (06/01 – 06/06)	Visualization with `Matplotlib` and `Seaborn`
Week 5 (06/08)	Midterm Exam, Introduction to Text Analytics
Week 6 (06/15 – 06/20)	Fundamentals of Machine learning with `Scikit-Learn`
Week 7 (06/22 – 06/27)	Fundamentals of Machine learning with `Scikit-Learn`
Week 8 (06/29)	Final Project Due