Lecture 1: Syllabus and Motivation

COSC 526: Introduction to Data Mining

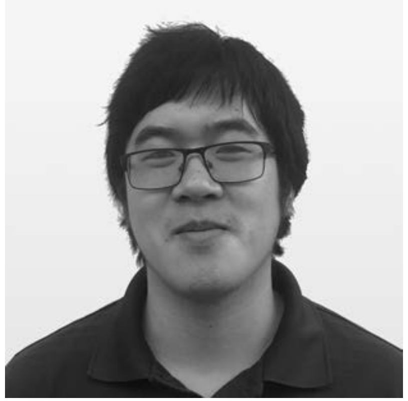
THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

Instructors:

Michela Taufer

Assistants to the Instructor:

Ian Lumsden



Nigel Tan

Kae Suarez

Paula Olaya

Leo Valera

Who we are

- Slide deck at: https://bit.ly/3pbtXNV

Course goals

- Build and use environments in which research on data can be designed, performed, and shared
  - GitHub, Jupyter Notebook, XSEDE Jetstream cloud, HPC systems
- Use distributed programming models and associated framework to analyze the data
  - MapReduce and Spark
  - Several ML methods
- Work with Cloud resources
  - XSEDE JetStream cloud (free of charge)
- Challenge yourself in a 4-week Hackathon in which you will work on a project
  - You can use your dataset

Lecture structure

- Short lecture (~60 minutes) to introduce a topic and define one or multiple practical problems related to that topic
- Work on the practical problems
- Group discussion and assessment of achievements
- Push of results (e.g., solutions and comments) in your private GitHub
- What if you need some more time to solve your problems?
    - Complete the unfinished work during the week and submit before the next lecture on Friday at 8AM (hard deadline)

Assignments

- Complete unfinished work during the week and submit before the next lecture on Friday at 8AM ET (hard deadline)

Course requirements

- Students have to bring their own laptop to the lecture
- No book is required
- Python programming skills requested
  - If you feel Python is not your forte, you are welcome to stay in the course but you will need to catch up with the programming skills in the next three weeks by yourself
- Weekly submissions are mandatory

Grades

- Participation and submission of practical problems: 50%
- Project with poster and 2-page paper: 50%

Office hours

- Instructor: Friday 3:15PM – 4:45PM or by appointment (sent email to taufer@utk.edu) – zoom room
- GRA: TBD

Lecture

- Friday 4:45PM – 7:45PM – zoom room
- *Are you available to anticipate the lecture at 3:00PM or 3:30PM ET?*

Outline of today’s lecture

- Live talk:
  - Genevieve Bell (Intel) on the origin of data analytics
- Establish a collaborative environment
  - Install git, GitHub, Jupyter, and learn how to use the tools
- Establish familiarly with text parsing
  - Handling files in different formats and different text formats
  - Code developed today will be used to get familiar with GitHub next week
- Learn to share solutions and discuss ideas

Code of Conduct (inspired by Bruce Elgort notes)

- Test **audio equipment**
- Join the lecture **on time**
- **Mute** your audio when you are not speaking
- Turn on the **webcam** during the lecture
- Be thoughtful of **when** you speak
  - Be courteous, and don’t interrupt the speaker
  - Use nodding, thumbs up, hand-raising to communicate, etch
- Minimize distractions and be present
  - Put away phones, close unrelated work, close the door

Before the Zoom session (From UTK)

- Join the Zoom session before the start time and test your microphone and webcam.
  - Your instructor will share the Zoom session link with you.
- Plan to be in a quiet room without potential interference and interruptions. Silence your phone.
- Find a sitting area with a plain, non-distracting background.
- Position the webcam so that your face is bright.
- Avoid back-lighting, such as sitting with your back to a window with bright light.
- Practice speaking to the camera and not to the screen.

During the Zoom session (From UTK)

- Mute your microphone when you are not talking.
- Have your webcam on.
- Use chat to communicate technical issues to the instructor and ask questions.
- Watch your actions on camera.
- Everyone can see that big, wide-mouth yawn!

The History of Big Data

Building our motivation

- Intel's Genevieve Bell shows that we have been dealing with big data for millennia, and that approaching big data problems with the right frame of reference is the key addressing many of the problems we face today from the keynote of Supercomputing 2013:

  https://youtu.be/CNoi-XqwJnA

- Your task:
  - List three key concepts you learned by watching the video