Lecture 1: Setting our environment and parsing text files

Instructor: Michela Taufer

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

Associate Instructors:

Danny Rorabaugh

Mike Wyatt

Dylan Chapp

GRA:

Joe Teague

Course goals

- Build and use environments in which research on data can be designed, performed, and shared
  - GitHub, Jupyter Notebook, XSEDE Jetstream cloud
- Use distributed programming models and associated framework to analyze the data
  - MapReduce → Spark
- Design MapReduce-based algorithms and run them in the Cloud
  - Well-known algorithms: Clustering and data processing
- Challenge yourself in the 4-week Hackathon

Lecture structure

- Short lecture (~30/40 minutes) to introduce a topic and define one or multiple practical problems related to that topic
- Work on the practical problems in team or by yourself
- 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 Monday at 8AM (hard deadline)

Course requirements

- Students have to bring their own laptop to the lecture
- No text 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
  - But if you complete your work in class, you are free for the rest of the week!

Grades

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

Office hours

- Instructor: Monday 3:30PM – 4:30PM or by appointment (sent email to taufer@utk.edu) – room: Min Kao 620
- GRA: Mon 1PM- 2PM, Wed and Thur 10AM: 11:30AM, or by appointment (with 24 hours notice) – room: RA office

Outline of today's lecture

- VIP talk:
  - Genevieve Bell (Intel) on the origin of data analytics
- Establish a collaborative environment
  - Install Jupyter and learn how to use it
- Establish familiarly with Python coding
  - 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