CSE 4/587
Data Intensive Computing

Dr. Eric Mikida
epmikida@buffalo.edu
208 Capen Hall

Dr. Shamshad Parvin
shamsadp@buffalo.edu
313 Davis Hall

Final Exam Review

Announcements

- Phase 3 due tonight

Final Exam Logistics

- The final is Monday 5/15/23 in **Cooke 121** from 11:45PM to 2:45PM
  - **EVERYONE IS IN COOKE 121**
  - If you have any official conflicts (2 exams at same time, or 3 same day) please let me know ASAP
- Seating is randomized
- All bags/electronics must be placed at the front of the room
- **What to Bring:**
  - Pen/Pencil
  - UB ID Card
  - A non-graphing calculator (if you like, not required)

Final Exam Review

Potential Topics:

1. Data Science Overview
2. Models/Algorithms (Linear Regression, Classifiers, K-Means)
3. HDFS Architecture and Protocol
4. MapReduce Fundamentals
5. MapReduce Applications
6. **Page Rank**
7. **Spark/Spark Streaming/Hive**
8. **Cloud**
9. **Ethics in DIC**

} Most of the depth will be focused on these topics

Data Science Overview [Lec 2-5]

1. Understand the overall goals and challenges of DIC
2. Know the four Vs and what they mean
3. Understand the various skills and components that DIC encompasses
4. Know what data cleaning/EDA is and the difference between then two

Linear Regression [Lec 6]

1. Explain the basic components of a Linear Regression model and what they mean/how to interpret them.
2. Understand and discuss evaluation metrics for determining the effectiveness of a given linear regression model.

K-Means Clustering [Lec 7, 11-12]

1. Understand how K-Means can be used to improve results from other models.
2. Understand and discuss potential issues with K-Means clustering.

Classifiers [8-12]

1. Understand the classification of classifiers
2. Understand the development cycle of a classification problem
3. Understand the basics of the different classifiers we have discussed
4. Understand the pros/cons of the classifiers discussed in class

Hadoop and HDFS Architecture [Lec 14,16]

- Understand and discuss the evolution of Hadoop from 1.0 to 3.0.
- Understand the basics of the HDFS architecture, the different components involved, and their roles and responsibilities.
- Understand and discuss block replication and its importance

MapReduce Basics [Lec 17-22]

- Understand and discuss the roles of the different types of MapReduce tasks that are part of a MapReduce Job.
- Understand the type of data that MapReduce deals with
- Understand the basics of the MapReduce algorithms discussed
  - Word Count
  - Word Co-Occurrence
  - k-mer Counting

Spark [Lec 27-29]

1. Be able to read and understand Spark programs (in Python)
2. Understand what an RDD is, and how it is stored/computed in Spark
    a. Understand the difference between a transformation and an action
    b. Understand the difference between a narrow and wide dependency
    c. Know what a lineage graph is and what it is used for in Spark
    d. Be able to generate DAGs of RDD transformations
    e. Be able to divide DAGs of transformations into stages for execution
3. Understand the fault tolerance mechanisms used by spark
4. Understand the benefits Spark provides

Cloud [Lec 33]

1. Understand the different types of service models in the cloud
2. Understand the challenges that cloud-based applications have to address
    a. Understand the basics of how these issues are addressed by tools we have encountered in class

Ethics in DIC [Lec 34]

1. Understand the different types of bias that may be part of our DIC applications
    a. Be able to explain what the types of bias are
    b. Be able to give examples of what may cause a particular type of bias to appear
    c. Be able to recognize situations that would cause a certain kind of bias to appear
    d. Be able to suggest possible solutions to address the different types of bias
    e. Understand which stages of the DIC pipeline each type of bias may appear in

Misc/Previous Topics [Lec 36]

1. Have a basic understanding of topics covered by the midterms
2. Have a basic understanding of what was covered in Lecture 36 (course recap)