Course Name: Massive Data Analytics

By: Iman Gholampour (For Fall 2020)

Course Description

Data Analytics is the science of analyzing data and converting information into useful knowledge. The past 10 years has seen steeply decreasing costs to gather, store, and process data, creating an even stronger motivation for working on massive datasets and empirical approaches to problem solving. Big data are characterized by several factors that can be broken down into 5 Vs: Volume, Velocity, Variety, Veracity, and Value. Dealing with massive data needs sophisticated mechanisms for Data Management, Storage and Processing. This course addresses a wide range of data analytic techniques to work with big data and massive datasets. The course builds a practical foundation for working with state-of-the-art computational platforms and tools for big data analytics. The main topics of such a course may include:

- Introduction to Data and Data Analytics
- Data Warehousing & Data Mining
- Massive Data Handling and Processing Challenges
- Common Technologies for Data Science
- Cluster Computing Framework
- MapReduce
- Fault Tolerance, Large Files Support, Data/Process Locality
- Distributed File System (DFS)
- Batch Processing and Apache Hadoop and Hadoop DFS (HDFS)
- Real-time and Stream Processing and Apache Spark
- Spark API, R, SQL, Python, Scala, Java …
- Big Data Enhanced Analytics Algorithms
- Graph Analytics
- Data Visualization
- Big Data, Machine Learning and AI Algorithms
- Unsupervised Learning for Big Data
- Data Abstraction Methods, Topic Modeling
- Machine Learning with Gaussian Process Modeling
- Information Security

Syllabus

- Introduction
  - MapReduce and Spark for mining of massive datasets
  - Data Mining
  - Distributed File system (DFS)
  - MapReduce and some algorithms using MapReduce
  - Spark tutorial and basic MapReduce functions
  - Programming Hadoop with Spark
- Finding similar Items in large-scale datasets
  - Locality Sensitive Hashing for documents
  - The theory of Locality Sensitive functions
  - LSH families for different Distance measure
- Frequent Item-set Mining
  - The Market-Basket model
  - Market-Basket A-priori
  - Handling large data set in memory
  - Counting frequent Items
- Dimensionality Reduction
  - Principal Component Analysis
  - Singular Value Decomposition
  - CUR decomposition
- Recommendation Systems
  - Content base
  - Collaborative filtering
  - UV decomposition
- Machine Learning with Big Data
  - Ensemble learning
  - Boosting and Bagging (Bootstrap Aggregation)
  - Stacking
  - Pipeline programming in ML
  - Data Clustering/Classification/Prediction with Gaussian Processes (GP) Modeling
- Data Abstraction Methods
  - Probabilistic Topic modeling (PLSA, LDA, …)
  - Correlated and Dynamic Topic Modeling
  - Non-Probabilistic Topic Modeling (STC, GSTC)
- Mining Data stream
  - Sampling data in streams
  - Filter streams
  - Counting distinct elements in stream
  - Counting ones in window

Prerequisites

- Advanced Programming

Evaluations

- Assignments
- Mid-term and Final Exams
- Final Projects

Practical Activities

- Using Python for Data Science
- Using Scala for Data Science
- Getting familiar with spark
- Getting familiar with programming Hadoop with Spark
- Using pipeline python programming for hyper parameter tuning and best model selection
- …

References

[1] Mining of Massive Data, J. Leskovek et. Al, Stanford University Free Book, 3rd Edition, 2019.
[2] Data Science in Practice, A. Said, V. Torra (Editors), Springer, 2019.
[3] Data Analytics with Spark Using Python, J. Allen, Addison Wesley, 2018.
[4] BIG DATA Algorithms, Analytics, and Applications (Editors), K-C Li, H. Jiang, …, CRC Press, 2015.
[5] Advanced Analytics with Spark, S. Ryza, S. Owen, J. Wils, O'Reilly, 2017.
[6] Scala Programming for Big Data Analytics, I. Elahi, Apress, 2019.
[7] …