Advanced Methods in Data Science and Big Data Analytics

40 HORAS

OVERVIEW

This course builds on skills developed in the Data Science and Big Data Analytics course. The main focus areas cover Hadoop (including Pig, Hive, and HBase), Natural Language Processing, Social Network Analysis, Simulation, Random Forests, Multinomial Logistic Regression, and Data Visualization. Taking an "Open" or technology-neutral approach, this course utilizes several open-source tools to address big data challenges.

AUDIENCE

This course is intended for aspiring Data Scientists, data analysts that have completed the associate level Data Science and Big Data Analytics course, and computer scientists wanting to learn MapReduce and methods for analyzing unstructured data such as text.

PREREQUISITE KNOWLEDGE/SKILLS

- Completion of the Data Science and Big Data Analytics course
- Proficiency in at least one programming language such as Java or Python

COURSE OBJECTIVES

Upon successful completion of this course, participants should be able to:

- Develop and execute MapReduce functionality
- Gain familiarity with NoSQL databases and Hadoop Ecosystem tools for analyzing large-scale, unstructured data sets
- Develop a working knowledge of Natural Language Processing, Social Network Analysis, and Data Visualization concepts
- Use advanced quantitative methods, and apply one of them in a Hadoop environment
- Apply advanced techniques to real-world datasets in a final lab

COURSE OUTLINE

Module 1: MapReduce and Hadoop

- The MapReduce Framework
- Apache Hadoop
- Hadoop Distributed File System
- YARN

Module 2: Hadoop Ecosystem and NoSQL

- Hadoop Ecosystem
- Pig
- Hive
- NoSQL - Not Only SQL
- HBase
- Spark

Module 3: Natural Language Processing

- Introduction to NLP
- Text Preprocessing
- TFIDF
- Beyond Bag of Words
- Language Modeling
- POS Tagging and HMM
- Sentiment Analysis and Topic Modeling

Module 4: Social Network Analysis

- Introduction to SNA and Graph Theory
- Most Important Nodes
- Communities and Small World
- Network Problems and SNA Tools

Module 5: Data Science Theory and Methods

- Simulation
- Random Forests
- Multinomial Logistic Regression

Module 6: Data Visualization

- Perception and Visualization
- Visualization of Multivariate Data

Teatinos 550, Santiago Centro – Mesa Central +56 22 6576300
www.sondatraining.com - contacto@sondatraining.com