SAF Study Abroad Foundation

COM SCI X 450.1 – Introduction to Data (CSE 450)

Learning Outcomes

- Describe what Data Science is and the skill sets needed to be a data scientist
- Explain in basic terms what Statistical Inference means. Identify probability distributions commonly used as foundations for statistical modeling. Fit a model to data.
- Use R to carry out basic statistical modeling and analysis.
- Identify and explain fundamental mathematical and algorithmic ingredients that constitute a
- Recommendation Engine (dimensionality reduction, singular value decomposition, principal
- component analysis). Build their own recommendation system using existing components.
- Create effective visualization of given data (to communicate or persuade).
- Work effectively (and synergically) in teams on data science projects.

Course Description

This course introduces students to the evolving domain of data science and to the food chain of knowledge domains involved in its application. Students learn a wide range of challenges, questions, and problems that data science helps address in different domains, including social sciences, finance, health and fitness, and entertainment. The course addresses the key knowledge domains in data science, including data development and management, machine learning and natural language processing, statistical analysis, data visualization, and inference. The course also provides an exposure to some of the technologies involved in application of data science, including Hadoop, NoSQL, and Python Programming language. The course includes case studies that require students to work on real-life data science problems.

Sample Modules Taught in Previous Program

- **Module 1:** Introduction: What is Data Science?

    - Big Data and Data Science hype – and getting past the hype
    - Why now? – Datafication
    - Current landscape of perspectives
    - Skill sets needed
- **Module 2:** Statistical Inference
    - Populations and samples
    - Statistical modeling, probability distributions, fitting a model
    - Intro to programming languages – R and Python
- **Module 3:** Exploratory Data Analysis and the Data Science Process
    - Basic tools (plots, graphs and summary statistics) of EDA
    - Philosophy of EDA
    - The Data Science Process
    - Case Study
- **Module 4:** Recommendation Systems: Building a User-Facing Data Product
    - Algorithmic ingredients of a Recommendation Engine
    - Dimensionality Reduction
    - Singular Value Decomposition
    - Principal Component Analysis
    - Exercise: build your own recommendation system
- **Module 5:** Data Visualization
    - Basic principles, ideas and tools for data visualization
    - Examples of inspiring (industry) projects
    - Exercise: create your own visualization of a complex dataset
- At the end of course, students will participate in a competition.