NUS SOC Summer Workshop 2026
Media, Analytics, Security & AI
Web Mining

Course Information

Pre-requisites

🐯 Which year of study is appropriate for your topic?

Any level.

🐯 What background and programming languages are required for your topic?

They should at least have basic programming knowledge. We will be using Python but if there are others who prefer R, it should be fine also (I can do a bridging or adapt the materials accordingly).

🐯 What do you think is attractive/unique about your topic to students?

This topic aims at addressing the practical aspects of data mining (not just on the modeling), but also how one can scrap data from any online sources. Most of the time, we assume that datasets will be provided or can be downloaded off the Internet. However, very often, we do not have this luxury and there is a need to mine the data ourselves. We will cover the different strategies and provide a systematic approach how to go about doing this.

This time round, we aim to launch an online coding system that allows student to practice the concepts learned. It will complement the in-class teaching and also cover additional skills.

Learning content and Teaching

🐯 What will be covered during "trial" lectures?

Lecture 1: Introduction to Analytics and Web Mining (3 hours):

- Introduction to Analytics and Web Mining
- Types of Decision Problem

Web Mining

- Statistical Learning
- Overview of techniques for performing Web Content Mining
- Some hands-on web scraping

Course Information

🐞 What will be covered during the "advanced" seminars?

Lecture 2: Regression

- Simple Linear Regression
- Multi Linear Regression
- Coding Scheme for Categorical Variables -Problems with Linear Regression

Lecture 3: Classification and Clustering

- Introduction to Classification
- Various Classification Techniques
- Assessing Model Accuracy
- Resampling Methods

Lecture 4: Mining Web Content I (3 hours):

- Web Basics
- Techniques for performing Web Content Mining
- Extracting Content from HTML Source Lecture 5: Mining Web Content II
- Document Object Model (DOM)
- XPath
- CSS Selectors
- Extracting Content using HTML Parser -GUI Web Scrapers Lecture 6: Mining Web

Content III

- Web Application Design

Web Mining

- Mining Web Data using APIs
- Scraping using an actual browser/headless browser Lecture 7: Recommender Systems
- Introduction to Recommender Systems
- Content-based Recommender Systems
- Collaborative Filtering Recommender Systems

Course Information

What will be the nature of the project work? How do you intend to split students into project groups, each consisting of 3 or 4 students?

Students will self-propose projects in the area of web mining. It would be a good idea to split the groups by university.

Do you have any recommendations for references (books) students can study to prepare for your topic before coming to NUS?

We will train the students from ground up so they do not need to prepare beforehand.

Besides their own personal laptops, what other equipment or software would students need for your topic?

They would need Python installed in their machine.