COURSE DETAIL

DATA ANALYSIS AND RETRIEVAL

Country
Netherlands

Host Institution
Utrecht University

Program(s)
Utrecht University

UCEAP Course Level
Upper Division

UCEAP Subject Area(s)
Computer Science

UCEAP Course Number
140

UCEAP Course Suffix

UCEAP Official Title
DATA ANALYSIS AND RETRIEVAL

UCEAP Transcript Title
DATA ANALYSIS & R

UCEAP Quarter Units
6.00

UCEAP Semester Units
4.00

Course Description

In this course, we build on foundation from database systems, focusing on two important issues. The first issue concerns how to deal with large volumes of data that do not have the precise record structure found in databases. The amount of unstructured data (primarily text) in the world far exceeds the amount of structured data. Searching through text requires a very different approach, especially because the number of results can be extremely large, making ranking based on relevance essential. This field is known as Information Retrieval (IR). Although this discipline has existed for quite some time, its relevance has increased in recent years due to the demand for web search engines. Become familiar with basic IR concepts such as precision, recall, Boolean search, indexing and posting lists, term weighting, the vector space model, and relevance feedback. Also take a detailed look at Google’s PageRank algorithm. This part includes a practical assignment in which IR techniques are applied to processing queries on relational databases, addressing the problem that the number of results can be either too large or too small. the second issue is how to extract interesting patterns and models from data. This is the domain of data mining and machine learning. Here too, the emphasis is on the analysis of unstructured data (again, primarily text), such as using data mining for document classification and clustering, as well as for ranking documents based on their relevance to a given query. The term “document” should be interpreted broadly: it may refer to web pages, email messages (spam or not spam?), posts to a newsgroup, or even tweets. The techniques covered include, among others, Naive Bayes classification, nearest neighbor, support vector machines, hierarchical clustering, and partitioning methods such as k-means clustering. This part also includes a practical assignment in which the data analysis techniques discussed in the lectures be applied to problems as described above. For this, we use the data analysis system R. Assumed previous knowledge in Databases (INFODB), Graphics (INFOGR), and Research Methods in Computer Science or Game Technology. If you have not passed these courses (or other courses in which you acquired comparable prior knowledge), we advise you not to choose this course.

Language(s) of Instruction

English

Host Institution Course Number

INFOB3DA

Host Institution Course Title

DATA ANALYSIS AND RETRIEVAL

Host Institution Course Details

https://osiris-student.uu.nl/onderwijscatalogus/extern/cursus

Host Institution Campus

Science

Host Institution Faculty

Host Institution Degree

Host Institution Department

Information and Computing Sciences

Course Last Reviewed

2025-2026

Print