Course type: lecture
Extent: two hours per week
Website: MOODLE

Outline: in this course we will review advanced topics in Web Data Management. The management of online data involves, among others, storage and effective processing of large volumes of data (as done, e.g., in search engines); accounting for data linkage by explicit hyperlinks or semantic connections; and in the extraction of structured knowledge from raw, semi-structured or unstructured data. We will introduce about prominent tools and principles in these areas.

1. Representation of online data
   - HTML, XML, Xpath
2. Crawling, PageRank, HITS, top-k algorithms
3. MapReduce, Spark, query languages, query evaluation
4. Information extraction, ontologies, RDF, SPARQL
Contents: the course will introduce the following main topics.

1. Representation of Web data
   - HTML, XML, Xpath

2. Search and ranking of Web pages
   - Crawling, PageRank, HITS, top-k algorithms

3. Big data frameworks
   - MapReduce, Spark, query languages, query evaluation

4. Information extraction and semantic Web
   - information extraction, ontologies, RDF, SPARQL

Requirements:

Prerequisites: first year courses, database systems (89-851) or databases (89-281) or in parallel

Assignments and other duties: final exam (passing grade in the exam is required to pass the course), programming assignments and exercises

Grade: 85% final exam, 15% assignments

Bibliography:
