ניהול נתוני עתק ברשת

Management of Big Web Data

89-542

סוג הקורס: שיעור **Course type:** lecture

היקף שעות: שעתיים בשבוע **Extent:** two hours per week

אתר הקורס באינטרנט: https://lemida.biu.ac.il **Website:**

הקורס יועבר בשפה העברית או האנגלית, לבחירת המרצה.

The course will be taught in Hebrew or English, depending on the lecturer's decision.

א. **מטרות הקורס:** בקורס זה נסקור נושאים עכשוויים ומתקדמים בניהול של מידע המועבר באינטרנט. ניהול מידע אינטרנטי כרוך, בין היתר, באחסון וביצוע יעיל של פעולות על כמויות גדולות של מידע (כפי שנעשה, למשל, במנועי חיפוש), בהתמודדות עם קישוריות של מידע לפי תוכן או בעזרת לינקים, ובחילוץ ידע מתוך נתונים. מטרת הקורס היא לערוך היכרות עם כלים ועקרונות מרכזיים בתחומים הנ"ל.

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 the extraction of structured knowledge from raw, semi-structured or unstructured data. We will introduce prominent tools and principles in these areas.

ב. תוכן הקורס: הקורס יסקור כמה נושאים מרכזיים.

(1) ייצוג של נתונים אינטרנטיים

- HTML, XML, Xpath

(2) חיפוש ודירוג של דפי אינטרנט

- Crawling, PageRank, HITS, top-k algorithms

(3) חישובים על נתוני עתק (big data)

- MapReduce, Spark, query languages, query evaluation

(4) חילוץ נתונים ובניית מאגרי מידע

- 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

ג. חובות הקורס:

דרישות קדם: מערכות מסדי נתונים (89-581) או מסדי נתונים (89-281) או במקביל

חובות / דרישות / מטלות: מבחן מסכם (חובה לעבור אותו כדי לעבור את הקורס!), תרגילים במהלך הסמסטר.

מרכיבי הציון הסופי: 85% בחינה, 15% תרגילים.

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: **ד. ביבליוגרפיה:**

Abiteboul, Serge, Ioana Manolescu, Philippe Rigaux, Marie-Christine Rousset, and Pierre Senellart. **Web data management**. Cambridge University Press, 2011.

Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman. **Mining of massive data sets. 3rd Edition.** Cambridge university press, 2020.

Curé, Olivier, and Guillaume Blin. **RDF database systems: triples storage and SPARQL query processing.** Morgan Kaufmann, 2014.