UČNI NAČRT PREDMETA / COURSE SYLLABUS	
Predmet:	Iskanje in ekstrakcija podatkov s spleta
Course title:	Web Information Extraction and Retrieval

Študijski program in stopnja **Study programme and level**	**Študijska smer** **Study field**	**Letnik** **Academic year**	**Semester** **Semester**
Interdisciplinarni magistrski študijski program Računalništvo in matematika	ni smeri	1 in 2	prvi ali drugi
Interdisciplinary Masters study programme Computer Science and Mathematics	none	1 in 2	first or second

Vrsta predmeta / Course type	obvezni
Univerzitetna koda predmeta / University course code:	63551

Predavanja **Lectures**	**Seminar** **Seminar**	**Vaje** **Tutorial**	**Klinične vaje** **work**	**Druge oblike študija**	**Samost. delo** **Individ. work**	**ECTS**
45	**10**	**20**			**105**	**6**

Nosilec predmeta / Lecturer:	Marko Bajec

Jeziki / Languages:	**Predavanja / Lectures:**	slovenski/Slovene, angleški/English
	Vaje / Tutorial:	slovenski/Slovene, angleški/English

Pogoji za vključitev v delo oz. za opravljanje študijskih obveznosti:	**Prerequisites:**
Vsebina:	**Content (Syllabus outline):**

Vsebina predavanj: Predmet bo pokrival naslednje vsebine: - Poizvedovanje in iskanje po spletu: Osnovni koncepti poizvedovanja Modeli poizvedovanja Odziv ustreznosti Mere za ocenjevanje točnosti poizvedb Predobdelava besedil in spletnih strani Inverzni index in njegova kompresija Latentno semantično indeksiranje Iskanje po spletu Meta iskanje po sletu: kombiniranje različnih načinov rangiranja; - Spletno pregledovanje in indeksiranje: Osnovni algoritem spletnega pajka Univerzalni spletni pajek Fokusirani spletni pajki Domenski spletni pajki - Ekstrakcija strukturiranih podatkov: Indukcija ovojnice Generiranje ovojnice na osnovi primera Samodejna izdelava ovojnice Ujemanje glede na obliko besede ali drevesne strukture	Content of the course: This course will cover the following topics: - Information Retrieval and Web Search: Basic Concepts of Information Retrieval Information Retrieval Models Relevance Feedback Evaluation Measures Text and Web Page Pre-Processing Inverted Index and Its Compression Latent Semantic Indexing Web Search Meta-Search: Combining Multiple Rankings - Web Crawling: A Basic Crawler Algorithm Implementation Issues Universal Crawlers Focused Crawlers Topical Crawlers - Structured Data Extraction: Wrapper Induction Instance-Based Wrapper Learning Automatic Wrapper Generation String Matching and Tree Matching

Večkratna poravnava	Multiple Alignment
Gradnja DOM dreves	Building DOM Trees
Ekstrakcija glede na stran s seznamom ali več strani	Extraction Based on a Single List Page or Multiple Pages
- Integracija podatkov:	- Information Integration:
Ujemanje glede na podatkovno shemo	Schema-Level Matching
Ujemanje glede na domeno in primere	Domain and Instance-Level Matching
Združevanje podobnosti	Combining Similarities
Ujemanje 1:m	1:m Match
Integracija iskalnikov po spletnih straneh	Integration of Web Query Interfaces
Izgradnja globalnega iskalnika po spletnih straneh	Constructing a Unified Global Query Interface
- Rudarjenje mnenja in analiza sentimenta:	- Opinion Mining and Sentiment Analysis:
Klasifikacija dokumentov po sentimentu	Document Sentiment Classification
Ugotavljanje subjektivnosti v stavkih in klasifikacija sentimenta	Sentence Subjectivity and Sentiment Classification
Slovarji besed in fraz, nosilcev mnenja	Opinion Lexicon Expansion
Aspektno orientirano rudarjenje mnenja	Aspect-Based Opinion Mining
Iskanje in extrakcija mnenja	Opinion Search and Retrieval

Temeljni literatura in viri / Readings:

- Bing Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications, Springer, August 2013
- Ricardo Baeza-Yates , Berthier Ribeiro-Neto: Modern Information Retrieval: The Concepts and Technology behind Search, 2nd Edition, ACM Press Books, 2010

Cilji in kompetence: **Objectives and competences:**

Cilj predmeta je študente naučiti, kako sprogramirati iskanje po spletu (po indeksiranem in neindeksiranem delu spleta) ter kako razviti programe za ekstrakcijo strukturiranih podatkov s statičnih in dinamičnih spletnih strani. Študentje bodo spoznali osnovne koncepte spletnega iskanja in ekstrakcije podatkov s spleta ter se naučili potrebnih tehnik, ki so za to potrebne. Po uspešno opravljene predmetu bodo sposobni samostojnega razvoja aplikacij, ki avtomatizirajo spletno iskanje in ekstrahirajo podatke s spletnih strani, vključno z ekstrakcijo podatkov iz on-line socialnih medijev.	The main objective of this course is to teach students about how to develop programs for web search (including surface web and deep web search) and for extraction of structural data from both, static and dynamic web pages. Beside basic concepts of the web search and retrieval, students will learn about relevant techniques and approaches. After the course, if successful, students will be able to develop programs for automatic web search and structured data extraction from web pages (including search and extraction from on-line social media).
Predvideni študijski rezultati:	**Intended learning outcomes:**
Po uspešno zaključenem modulu bodo študenti zmožni: - Povzeti najpomembnejše pristope in tehnike s področja iskanja in ekstrakcije podatkov s spleta - presoditi, kateri pristopi s področja iskanja in ekstrakcije podatkov s spleta so najbolj primerni za reševanje posameznih problemov, - razviti aplikacije za zajem in analizo podatkov s spleta, - konstruirati lastne algoritme za ekstrakcijo podatkov s spleta, - pojasniti delovanje in časovno kompleksnost algoritmov iskanja po spletu, - uporabiti in integrirati različne odprto-kodne rešitve s področja iskanja in ekstrakcije podatkov s spleta	After successful completion of the module, students will be able to: - summarize the most important approaches and techniques for searching and extracting data from the web - to select approaches and techniques that are most suitable for individual problems in web information extraction and retrieval. - to develop applications for data acquisition and analysis, - to construct new algorithms for web data search and extraction, - to explain behavior and time complexity of specific web search algorithms, - to integrate and employ different open-source solutions from the field.

Metode poučevanja in učenja:	**Learning and teaching methods:**
Predavanja, računske vaje z ustnimi nastopi, projektni način dela pri domačih nalogah in seminarjih.	Lectures, seminars, homeworks, oral presentations, project work.

Načini ocenjevanja:	Delež (v %) / Weight (in %)	**Assessment:**
Način (pisni izpit, ustno izpraševanje, naloge, projekt): Sprotno preverjanje (domače naloge, kolokviji in projektno delo) Končno preverjanje (pisni in ustni izpit) Ocene: 6-10 pozitivno, 5 negativno (v skladu s Statutom UL).	50% 50%	Type (examination, oral, coursework, project): Continuing (homework, midterm exams, project work) Final (written and oral exam) Grading: 6-10 pass, 5 fail (according to the rules of University of Ljubljana).

Reference nosilca / Lecturer's references:

Marko Bajec:

– ŠUBELJ, Lovro, JELENC, David, ZUPANČIČ, Eva, LAVBIČ, Dejan, TRČEK, Denis, KRISPER, Marjan, BAJEC, Marko. Merging data sources based on semantics, contexts and trust. The IPSI BgD transactions on internet research, ISSN 1820-4503. [Print ed.], 2011, vol. 7, no. 1, str. 18-30, ilustr [COBISS.SI-ID 7850580]

– ŠUBELJ, Lovro, FURLAN, Štefan, BAJEC, Marko. An expert system for detecting automobile insurance fraud using social network analysis. Expert systems with applications, ISSN 0957-4174. [Print ed.], 2011, vol. 38, no. 1, str. 1039-1052, ilustr [COBISS.SI-ID 7870292]

– LAVBIČ, Dejan, BAJEC, Marko. Employing semantic web technologies in financial instruments trading : Dejan Lavbič and Marko Bajec. International journal of new computer architectures and their applications, ISSN 2220-9085. [Online ed.], 2012, vol. 2, no. 1, str. 167-182, ilustr [COBISS.SI-ID 9035348]

– ŽITNIK, Slavko, ŠUBELJ, Lovro, LAVBIČ, Dejan, VASILECAS, Olegas, BAJEC, Marko. General context-aware data matching and merging framework. Informatica, ISSN 0868-4952, 2013, vol. 24, no. 1, str. 119-152, ilustr [COBISS.SI-ID 9735764]

– ŠUBELJ, Lovro, BAJEC, Marko. Group detection in complex networks : an algorithm and comparison of the state of the art. Physica. A, Statistical mechanics and its applications, ISSN

0378-4371. [Print ed.], 1 March 2014, vol. 397, str. 144-156 [COBISS.SI-ID 10333012]