

## 4IZ550 Dolování z webu

Česky	Dolování z webu
Anglicky	Web mining
Německy	Web mining
ECTS kredity	3
Forma výuky	2/0
Garant	Doc. Ing. Vojtěch Svátek, Dr.
Výchozí předměty	4IZ210
Původní předmět (předměty)	

### Anotace

Předmět je zaměřen na technologie dolování užitečných informací a znalostí z velkých objemů webových dat. Důraz je kladen jak na celkový rámcový přehled o oboru, tak na získání znalostí potřebných pro praktické aplikování vybraného okruhu těchto technologií.

### Cíl

Vybavit studenty přehledem o technologiích dolování z webu a schopností některé z nich aplikovat v praxi.

### Osnova

- § Přehled hlavních okruhů metod dolování z webu: dolování z obsahu webu (Web Content Mining), ze struktury webu (Web Structure Mining) a z uživatelského chování na webu (Web Usage Mining)
- § Přehled praktických aplikací založených na dolování z webu
- § Dolování z obsahu webu: indexování a vyhledávání dokumentů (Information Retrieval) ve webovém prostředí – booleovský a vektorový model vyhledávání, indexování latentní sémantiky (LSI); uspořádání nalezených dokumentů; meta-vyhledávání
- § Dolování z obsahu webu: kategorizace a shlukování webových dokumentů
- § Aplikace metod zpracování přirozeného jazyka při dolování z webu: lemmatizace, rozpoznávání slovních druhů, desambiguace, povrchová syntaktická analýza atd.
- § Využívání struktury odkazů: primární procházení webu (crawling, spidering), analýza topologie odkazů, metody PageRank a HITS
- § Globální analýza webu; analýza sociálních sítí na WWW
- § Dolování z uživatelského chování na webu; internetový marketing
- § Extrakce informací jako specifický typ dolování z obsahu webu: wrapperový přístup vs. extrakce aktivovaná příznaky
- § Specifické aplikace: dolování názorů („opinion mining“) vs. dolování faktů („fact mining“); analýza webového spamu; komparativní nakupování; atd.
- § Integrace informací získaných z WWW, využití mapování schémat
- § Vztah dolování z webu a technologií sémantického webu: automatické sémantické anotování, učení ontologií, vyhledávání na sémantickém webu

## Studijní zátěž

### Pro prezenční formu

Účast na přednáškách	13 x 2 hodiny	26 hodin
Příprava na průběžný test 1 (samostatné studium)	24 hodin	24 hodin
Příprava na průběžný test 2 (samostatné studium)	36 hodin	36 hodin
<b>Celková pracovní zátěž</b>		<b>86 hodin</b>

## System ověřování znalostí (požadavky na ukončení předmětu)

### Pro prezenční formu

Průběžný test (pro úspěšné absolvování minimum 15 bodů)	0 až 30 bodů	
Závěrečný test vč. mikroeseje (pro úspěšné absolvování minimum 30 bodů)	0 až 70 bodů	
<b>Celkem – maximum</b>		<b>100 bodů</b>

### Klasifikační stupnice

viz Studijní a zkušební řád VŠE

## Literatura předmětu

Druh lit.	ISBN	Název knihy	Autoři	Rok vydání
Z	3-540-37881-2	Web Data Mining	Liu, B.	2006
Z	80-7248-041-3	Vybrané kapitoly z počítačového zpracování přirozeného jazyka.	Strossa, P.	1999
D	978-0-387-30632-2	Ontology Learning and Population from Text	Cimiano, P.	2006

## Skupina

<b>Studium</b>	<b>Obor, popř. vedlejší specializace</b>	<b>Skupina</b>
Navazující magisterské	Obor znalostní technologie	Volitelný předmět