

СОФИЙСКИ УНИВЕРСИТЕТ "СВ. КЛИМЕНТ ОХРИДСКИ"
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА

учебна година: 2009/2010

семестър: летен
(зимен, летен)

наименование на дисциплината: "Извличане на информация"
хорариум: 45+30 вид на дисциплината: избираема
специалност: курс:
лектор: доцент д-р Иван Койчев

1. Кратка анотация на дисциплината

"Извличането на информация" (Information Retrieval) една от най-активно развиващите се области на Информатиката. Това е провокирано най-вече от експлозивното развитие на уеб и желанието ни лесно и бързо намираме полезна информация в него. Към настоящето представена в електронен вид неструктурираната информация (текст) значително надхвърля по обем структурираната информация (данни). Системите за търсене и извличане на полезна информация в океан от неструктурирана (полу-структурирана) информация имат водещ пазарен дял.

Курсът си поставя за цел да запознае студентите с основните методи и технологии на Извличането на Информация (ИзИнф): както класически така и модерни подходи. Ще бъдат разгледани и подходи, които се опитват да отидат по-далече от търсене по ключова дума, като добавят и степен на интелигентност при извличането на информация.

По време на лабораторните занятия студентите ще се запознаят със съответните отворени среди и ще получат задания за самостоятелна работа (проект), който ще представят пред колегите си в края на курса.

2. Предварителни изисквания към студентите (отнася се само за избираемите дисциплини)
3. Форма на проверка на знанията и уменията и начин на формиране на оценката по дисциплината

	% от оценката
Текуща оценка	65%
– курсова работа	55%
– котролна работа	0%
– активно учатие в часовете	5%
– присъствие в час	5%
Изпит	35%
– практически (задачи)	0
– теоретически	35%

4. Тематичен план (конспект) на дисциплината

Въпроси	Глави от [1]
1. Задачи на Извличането на Информация. Булев модел. Обърнат индекс.	1
2. Отделяне на елементите на текста. Изграждане на речници от термини. Стоп думи. Нормализация. Списъци с адреси. Въпроси-фрази.	2
3. Толерантно извличане. Корекция на правописа. Фонетична корекция.	3
4. Конструирание на индекси. Разпределени индекси. Динамични индекси.	4
5. Параметрично и зонално индексирание. Честота и тегла на термините. Теглови функции.	6
6. Векторно-пространствен модел.	7
7. Оценка на системите за ИзИнф. Мерки. Оценка върху корпуси от текстове.	8
8. Обратна връзка и разширяване на въпроси.	9
9. Извличане от XML документи.	10
10. Вероятностни модели в Извличането на Информация	11
11. Модели на езика.	12
12. Класификация на текст – Наивен бейсов подход. Избор на разделящи характеристики.	13
13. Класификация на текст – Векторно-пространствен модел. Rocchio; kNN; линейни и нелинейни класификатори. Повече от един два класа.	14
14. Редуциране на размерността. Неявно семантично индексирание (LSI).	18
15. WWW: уеб характеристики; спам; реклами; търсене; индекси; дублирания.	19
16. WWW: обхождане и индексирание; разпределени индекси	20

5. Литература

ОСНОВЕН УЧЕБНИК:

1. C. Manning P. Raghavan, H. Schütze. *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England 2007
<http://www-csli.stanford.edu/~schuetze/information-retrieval-book.html>

ДРУГИ:

2. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.
<http://www.ischool.berkeley.edu/~hearst/irbook/>
3. van Rijsbergen, C. J. 1979. *Information Retrieval*. London: Butterworths. Second Edition.
<http://www.dcs.gla.ac.uk/Keith/Preface.html>