Discipline	Internet Information Retrieval code: 43 7 semester – winter	
Specialty	SOFTWARE AND INTERNET TECHNOLOGIES	
ECTS credits: **6**	Form of assessment: Continuous assessment	
Lecturer	Professor, Ph.D Hristo Valchanov Room 207-4E Phone: +359 052 383 439 E-mail: hristo@tu-varna.bg	
Department	Computer Science and Engineering	
Faculty	Faculty of Computing and Automation	

Learning objectives:

The discipline aims to acquaint students with the principles of information retrieval on the Internet. The basic concepts and methods of extracting data from documents are examined, with an emphasis on modern approaches and algorithms for searching for information in the Web space. Issues related to information indexing, information retrieval models, ranking, and querying are addressed. The principles of building search engines on the Web, as well as the features of modern systems for extracting information on the Internet, are considered.
In laboratory exercises, students must develop a small search engine from scratch. The search engine includes crawler, inverse index, DNS and web clients. The implementation is under Visual Studio on C or C++ languages.

CONTENTS:

Training Area	Hours lectures	Hours seminar exercises

Basic principles of information extraction.	2	
Architecture of a search engine. Basic components. Functioning	2	
Retrieve web pages. Web Crawling. Storage of retrieved documents.	2	
Word processing. Evaluation of the resulting set	2	
Parsing a document. Link analysis.	2	
Ranking and Indexing. Building indexes. Inverted indexes. Compression	2	
Queries. Transformations and refinement of queries. Display results.	2	
Models of information retrieval.	2	
Evaluation of search engines.	2	
Classification and clustering. Spam detection.	2	
Social Search. Tags. Document filtering.	2	
Retrieval of XML documents. Peculiarities.	2	
Information Retrieval Systems. LEXIS/NEXIS, SMART, Dialog, Dow Jones News/Retrieval, INQUERY.	2	
Google Search Engine Architecture.	2	
Digital Libraries	2	
String searching		2
Implementation of web crawler		4
Text parsing.		4
Parsing of HTML documents		4
Implementation of HTTP client		4
Implementation of DNS client		4
Information indexing		4
Control of the process of following links. Robots.txt		2
Query processing		2
TOTAL: **60** h	**30**	**30**