Національний технічний університет України
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Computer Systems Software Department

Information Retrieval Systems Software. Part 2. Methods of Information Retrieval Organization

Syllabus

Requisites of the Course	
Cycle of Higher Education	***First cycle of higher education (Bachelor's degree)***
Field of Study	*12 Information Technologies*
Speciality	*121 Software engineering*
Education Program	*Software Engineering of Multimedia and Information Retrieval Systems*
Type of Course	*Normative*
Mode of Studies	*full-time*
Year of studies, semester	*4 year (8 semester)*
ECTS workload	*4,5 credits (ECTS). Time allocation: 36 hours for lectures, 18 hours for lab classes, 81 hours for self-study*
Testing and assessment	*Exam, midterm test*
Course Schedule	*According to rozklad.kpi.ua*
Language of Instruction	*English*
Course Instructors	*Lecturer: teaching assistant, Yakiv Yusyn, yusyn@pzks.fpm.kpi.ua* *Teacher of laboratory work: teaching assistant, Yakiv Yusyn, yusyn@pzks.fpm.kpi.ua*
Access to the course	*Google Classroom. Link to be provided to registered students*

Outline of the Course

1. Course description, goals, objectives, and learning outcomes

The study of the discipline "Information Retrieval Systems Software 2. Methods of Organization of Information Retrieval" allows students to form the competencies necessary to solve practical problems of professional activities related to developing and using of information retrieval methods in software.

*The **purpose** of studying the discipline "Information Retrieval Systems Software 2. Methods of Organization of Information Retrieval" is to form in students the ability to develop software for information retrieval, as well as use third-party software and libraries for information retrieval.*

*The **subject** of the discipline "Information Retrieval Systems Software 2. Methods of Organization of Information Retrieval" is mathematical and algorithmic methods of presentation and search of unstructured documental information in information retrieval systems.*

*Study of the discipline "Information Retrieval Systems Software 2. Methods of Organization of Information Retrieval" contributes to the formation of **professional competencies (FC)** in students, necessary for solving practical tasks of professional activity, related to the development and use of information retrieval methods in software.*

***PC07** Knowledge of information data models, the ability to create software for data storage, retrieval and processing.*

***PC13** Ability to reasonably select and master software development and maintenance tools.*

***PC14** Ability to algorithmic and logical thinking.*

PC15 *Ability to apply fundamental and interdisciplinary knowledge to build advanced retrieval algorithms.*

PC17 *Ability to develop software for information retrieval systems.*

PC20 *Ability to apply the acquired fundamental mathematical knowledge to develop calculation methods in the multimedia and information retrieval systems creation.*

Study of the discipline "Information Retrieval Systems Software 2. Methods of Organization of Information Retrieval" contributes to the formation in students of the following ***program learning outcomes (PLO)*** *according to the educational program:*

PLO03 *To know the software life cycle basic processes, phases and iterations.*

PLO05 *To know and apply relevant mathematical concepts, domain methods, system and object-oriented analysis and mathematical modeling for software development.*

PLO06 *Ability to select and use the appropriate task of software development methodology.*

PLO07 *To know and to apply in practice the fundamental concepts, paradigms and basic principles of the functioning of language, instrumental and computational tools of software engineering.*

PLO08 *To know and to be able to develop a human-machine interface.*

PLO10 *To conduct a pre-project survey of the subject area, system analysis of the design object.*

PLO11 *To select initial data for design, guided by formal methods of describing requirements and modeling.*

PLO12 *To apply effective approaches to software design in practice.*

PLO13 *To know and apply methods of developing algorithms, designing software and data and knowledge structures.*

PLO15 *To choose programming languages and development technologies to solve the problems of creating and maintaining software.*

PLO16 *To have the software development, design approval and all types of software documentation release skills.*

PLO17 *To be able to apply methods of component software development.*

PLO18 *To know and be able to apply information technology of processing, storage and transmission of data.*

PLO19 *To know and be able to apply software verification and validation methods.*

PLO20 *To know approaches to evaluation and quality assurance of software.*

PLO25 *To know and to be able to use fundamental mathematical tools in the algorithms construction and modern software development.*

PLO31 *To be able to identify, analyze and document software requirements for multimedia and information retrieval systems.*

PLO38 *To be able to apply programming technologies for multimedia and information retrieval systems software development.*

PLO39 *To know the types of search engines, the principles of their construction, the methods and algorithms for performing different kinds of information retrieval in them.*

PLO40 *To know and be able to apply in practice the methods and criteria for estimating the effectiveness of information retrieval.*

PLO42 *To know the basic presentation models of textual and multimedia information and methods of its pre-processing for use in the design of information retrieval systems.*

PLO43 *To know and be able to use in practice the existing software resources and libraries for processing of textual information and multimedia data in information retrieval systems.*

PLO44 *To know the most common query languages used in the development of information retrieval systems.*

2. Prerequisites and post-requisites of the course (the place of the course in the scheme of studies in accordance with curriculum)

Successful study of the discipline "Information Retrieval Systems Software 2. Methods of Organization of Information Retrieval" is preceded by the study of discipline "Information Retrieval Systems Software 1. NoSQL Databases" of the curriculum for Bachelors in 121 Software Engineering.

The theoretical knowledge and practical skills obtained during the mastering of the discipline "Information Retrieval Systems Software 2. Methods of Organization of Information Retrieval" ensure the successful implementation of course and diploma projects in the specialty 121 Software engineering. Also, the acquired knowledge and skills are a prerequisite for studying the disciplines "Information Retrieval Systems and Services" and "Artificial Intelligence Technologies for Information Retrieval Systems" of the Master's degree curriculum in 121 Software Engineering.

3. Content of the course

Discipline "Information Retrieval Systems Software 2. Methods of Organization of Information Retrieval" involves the study of the following topics:

Topic 1. Introduction to information retrieval

Topic 2. Presentation models of documents for information retrieval

Topic 3. Full-text search engine "Elasticsearch"

Topic 4. Overview of advanced tasks of information retrieval

Topic 5. Validation and verification of information retrieval results

Midterm test

Exam

4. Coursebooks and teaching resources

Main resources:

1. Manning C. D. An Introduction to Information Retrieval / C. D. Manning, P. Raghavan, H. Schütze. – Cambridge: Cambridge University Press, 2008. – p. 544.
Get acquainted with the sections related to the following topics of the discipline: introduction to information retrieval, presentation models of documents for information retrieval, overview of advanced tasks of information retrieval, validation and verification of information retrieval results. Open-Access Resource.

2. Elasticsearch Guide [Online]. – 2022. – Available: https://www.elastic.co/guide/en/elasticsearch/reference/master/index.html.
Get acquainted with the sections related to the following topics of the discipline: full-text search engine "Elasticsearch". Open-Access Resource.

3. Kibana Guide [Online]. – 2022. – Available: https://www.elastic.co/guide/en/kibana/current/index.html.
Get acquainted with the sections related to the following topics of the discipline: full-text search engine "Elasticsearch". Open-Access Resource.

Educational content

5. Methodology

No	Type of a class	A class description

Topic 1. Introduction to information retrieval		
1	*Lecture 1. Basic concepts of information retrieval*	*Introduction. The purpose and tasks of the discipline of information retrieval. The main conceptual apparatus of the discipline.* *Materials for self-studying: p. 6, No. 1.*
2	*Lecture 2. Information retrieval*	*The task of information retrieval. Problems of information retrieval. Main types of information retrieval. Applications of information retrieval. Requirements for information retrieval results.* *Materials for self-studying: p. 6, No. 2.*
Topic 2. Presentation models of documents for information retrieval		
3	*Lecture 3. Set-theoretic models*	*Standard Boolean model. Extended Boolean model. Fuzzy retrieval.* *Materials for self-studying: p. 6, No. 3.*
4	*Laboratory class 1. Implementation of presentation models (part 1)*	*Task: to implement a set-theoretic model.* *Materials for self-studying: p. 6, No. 4.*
5	*Lecture 4. Algebraic models. Part 1*	*Vector space model. TF-IDF.* *Materials for self-studying: p. 6, No. 5.*
6	*Lecture 5. Algebraic models. Part 2*	*Generalized vector space model. Topic-based vector space model. Latent semantic indexing. Latent semantic analysis.* *Materials for self-studying: p. 6, No. 6.*
7	*Laboratory class 2. Implementation of presentation models (part 2)*	*Task: to implement an algebraic model.* *Materials for self-studying: p. 6, No. 7.*
8	*Lecture 6. Probabilistic models. Part 1*	*Binary Independence Model. Probabilistic relevance model. Uncertain inference.* *Materials for self-studying: p. 6, No. 8.*
9	*Lecture 7. Probabilistic models. Part 2*	*Language models. Latent Dirichlet allocation.* *Materials for self-studying: p. 6, No. 9.*
10	*Laboratory class 3. Implementation of presentation models (part 3)*	*Task: to implement an algebraic model.* *Materials for self-studying: p. 6, No. 10.*
Topic 3. Full-text search engine "Elasticsearch"		
11	*Lecture 8. Full-text search engine "Elasticsearch": overview, main concepts.*	*Elastic stack. Main properties of Elasticsearch engine. Glossary.* *Materials for self-studying: p. 6, No. 11.*
12	*Lecture 9. Indexing. Index creation. Filters and simple*	*Index creation in Elasticsearch, document storing in it. Difference between filters and queries. Filters*

	queries using Elastic Query DSL.	*execution. Simple queries.* *Materials for self-studying: p. 6, No. 12.*
13	*Laboratory class 4. Metadata search using Elasticsearch (part 1)*	*Task: to implement software for documents storing in Elasticsearch.* *Materials for self-studying: p. 6, No. 13.*
14	*Lecture 10. Complex queries using Elastic Query DSL*	*Complex queries: fuzzy, wildcard, regex, range.* *Materials for self-studying: p. 6, No. 14.*
15	*Lecture 11. Elasticsearch analyzers: stop-word removal, stemming*	*Analyzers in Elasticsearch. Stop-word definition. Stemming process. Standard Elasticsearch analyzer. Analyzer for English.* *Materials for self-studying: p. 6, No. 15.*
16	*Laboratory class 5. Metadata search using Elasticsearch (part 2)*	*Task: to add the possibility of metadata search using Elasticsearch filters.* *Materials for self-studying: p. 6, No. 16.*
17	*Lecture 12. Elasticsearch integration into existing software systems*	*Elasticsearch cluster. Best practices of Elasticsearch integration.* *Materials for self-studying: p. 6, No. 17.*
18	*Lecture 13. Data visualization tool Kibana. Query language KQL*	*Kibana as a tool for data visualization from Elasticsearch. Main concepts. Simple queries using Kibana Query Language.* *Materials for self-studying: p. 6, No. 18.*
19	*Laboratory class 6. Metadata search using Elasticsearch (part 3)*	*Task: to add the possibility of metadata search using Elasticsearch filters.* *Materials for self-studying: p. 6, No. 19.*
Topic 4. Overview of advanced tasks of information retrieval		
20	*Lecture 14. Documents classification*	*The task of documents classification. Methods of documents classification.* *Materials for self-studying: p. 6, No. 20.*
21	*Laboratory class 7. Full-text search using Elasticsearch (part 1)*	*Task: to implement software for full-text search using Elasticsearch.* *Materials for self-studying: p. 6, No. 21.*
22	*Lecture 15. Documents clustering*	*The task of documents clustering, the difference from classification. Methods of document clustering.* *Materials for self-studying: p. 6, No. 22.*
23	*Laboratory class 8. Full-text search using Elasticsearch (part 2)*	*Task: to add Elasticsearch analyzer.* *Materials for self-studying: p. 6, No. 23.*
Topic 5. Validation and verification of information retrieval results		
24	*Lecture 16. Criteria for evaluating the effectiveness of information retrieval*	*Criteria for evaluating the effectiveness: precision, recall, F-measure.* *Materials for self-studying: p. 6, No. 24.*
25	*Laboratory class 9. Full-text search using Elasticsearch (part 3)*	*Task: to add Elasticsearch analyzer.* *Materials for self-studying: p. 6, No. 25.*

26	*Lecture 17. Automated testing of information retrieval systems*	*Automated testing of information retrieval systems. Oracle problem. Metamorphic testing. Materials for self-studying: p. 6, No. 26.*
Midterm test		

6. Self-study

The discipline "Information Retrieval Systems Software 2. Methods of Organization of Information Retrieval" is based on independent preparation for classroom classes on theoretical and practical topics.

No	*Topic name*	*Number of hours*	*Resources*
1	*Preparation for the lecture 1*	*1.5*	*1, preface*
2	*Preparation for the lecture 2*	*1.5*	*1, preface*
3	*Preparation for the lecture 3*	*1.5*	*1, pp. 1-16*
4	*Preparation for the laboratory class 1*	*2*	*All resources for lecture(s) No 3*
5	*Preparation for the lecture 4*	*1.5*	*1, pp. 109-132*
6	*Preparation for the lecture 5*	*1.5*	*1, pp. 109-132*
7	*Preparation for the laboratory class 2*	*2*	*All resources for lecture(s) No 4-5*
8	*Preparation for the lecture 6*	*1.5*	*1, pp. 219-251*
9	*Preparation for the lecture 7*	*1.5*	*1, pp. 219-251*
10	*Preparation for the laboratory class 3*	*2*	*All resources for lecture(s) No 4-5*
11	*Preparation for the lecture 8*	*1.5*	*2, sections “What is Elastic?”, “Glossary”*
12	*Preparation for the lecture 9*	*1.5*	*2, section «Query DSL»*
13	*Preparation for the laboratory class 4*	*2*	*All resources for lecture(s) No 8-9*
14	*Preparation for the lecture 10*	*1.5*	*2, section “Query DSL”*
15	*Preparation for the lecture 11*	*1.5*	*2, section “Text analysis”*
16	*Preparation for the laboratory class 5*	*2*	*All resources for lecture(s) No 8-10*
17	*Preparation for the lecture 12*	*1.5*	*All resources for lecture(s) No 8-9*
18	*Preparation for the lecture 13*	*1.5*	*3, sections “What is Kibana?”, “Quick Start”, “Kibana Query Language”*

19	*Preparation for the laboratory class 6*	*2*	*All resources for lecture(s) No 8-10, 12*
20	*Preparation for the lecture 14*	*1.5*	*1, pp. 289-307*
21	*Preparation for the laboratory class 7*	*2*	*All resources for lecture(s) No 8-12*
22	*Preparation for the lecture 15*	*1.5*	*1, pp. 349-395*
23	*Preparation for the laboratory class 8*	*2*	*All resources for lecture(s) No 8-12*
24	*Preparation for the lecture 16*	*1.5*	*1, pp. 151-172*
25	*Preparation for the laboratory class 9*	*2*	*All resources for lecture(s) No 8-12*
26	*Preparation for the lecture 17*	*1.5*	*All resources for lecture(s) No 12*
27	*Preparation for the midterm test*	*7.5*	*All resources for lecture(s) No 1-9, 14-16*
28	*Preparation for the exam*	*30*	*All resources for the semester*

Policy and Assessment

7. Course policy

- *Attendance at lectures is mandatory.*
- *Attendance at laboratory classes can be sporadic and if you need to defend a laboratory work.*
- *Rules of conduct in the classroom: activity, respect for those present.*
- *Adherence to the policy of academic integrity.*
- *Rules for protecting the laboratory work: the work should be done according to the option of the student, which is determined by his number in the group list.*
- *The rules for assigning penalty points are as follows.*

Penalty points are awarded for:

- *Plagiarism (completed task does not correspond to the variant of the task) in the laboratory work: - 5 points for each attempt.*

8. Monitoring and grading policy

During the semester, students complete 3 laboratory works. Maximum number of points for each work: 8 points

Points are awarded for:
- quality of the work: 0-3 points;
- answer during the work defense of: 0-3 points;
- timely submission of work to the defense: 0-2 points.

Criteria for evaluating the quality of work:
3 points – the work is done and performed qualitatively;
2 points – the work is done qualitatively, but has shortcomings;
1 point – the work is done qualitatively, but has some issues;
0 points – the work is done qualitatively, but has serious issues.

Answer evaluation criteria:
3 points – complete answer, well-reasoned;
2 points – complete answer, with minor issues;
1 point – complete answer, with major issues;
0 points – not complete or incorrect answer.

Timely submission of work to the defense:
2 points – the work is submitted for defense no later than the specified deadline;
0 points – the work is submitted for defense later than the specified deadline.

Maximum number of points for performing and defending laboratory works:
8 points × 3 laboratory works = 24 points.

The task for the midterm test consists of 18 test questions – 10 questions with one correct answer and 8 questions with few correct answers. Maximum number of points for each question with one correct answer is 1 point, maximum number of points for each question with few correct answers is 2 points.

Answer evaluation criteria for a question with one correct answer:
1 point – correct answer;
0 points – no answer or incorrect answer.

Answer evaluation criteria for a question with few correct answers:
2 points – all correct answers and no incorrect answers are selected;
1 point – at least 50% of all correct answers are selected;
0 points – no answer or all answers are wrong.

Maximum number of points for the midterm test:
1 point × 10 questions with one correct answer + 2 points × 8 questions with few correct answers = 26 points.

Semester component of the rating scale R_C = 50 points, it is defined as the sum of points received for the performance and defense of laboratory works and points received for the midterm test.
R_C = 24 points + 26 points = 50 points.

Composition and evaluation criteria of the exam answer:
The exam consists of 3 questions – 2 theoretical and 1 practical.
Exam component of the rating scale R_E = 50 points.

Maximum number of points for the first theoretical question is 20 points, for the second theoretical question is 10 points, and for the practical question is 20 points.

Evaluation criteria for the first theoretical question:
18-20 points – complete answer, well-reasoned;
10-17 points – complete answer, with minor issues;
5-9 points – complete answer, with major issues;
0-4 points – not complete or incorrect answer.

Evaluation criteria for the second theoretical question:
9-10 points – complete answer, well-reasoned;
7-8 points – complete answer, with minor issues;
4-6 points – complete answer, with major issues;
0-3 points – not complete or incorrect answer.

Evaluation criteria for the practical question:
18-20 points – complete answer, well-reasoned;
10-17 points – complete answer, with minor issues;
5-9 points – complete answer, with major issues;
0-4 points – not complete or incorrect answer.

Maximum number of points for the exam:

R_E =20+10+20= 50 points.

The rating scale for the discipline is: $R = R_C + R_E$= 50 points + 50 points = 100 points

Semester assessment: Exam.
Prerequisite for admission to the exam is the performance and defense of all laboratory works.

The table of compliance between overall points and the final grade:

Points	*Grade*
100-95	Excellent
94-85	Very good
84-75	Good
74-65	Satisfactory
64-60	Fair
Less than 60	Unsatisfactory
Course requirements are not met	Not graded

Syllabus of the course

Is designed by teaching assistant, Yakiv Yusyn

Adopted by Computer Systems Software Department (protocol № 12 from 26.04.23)

Approved by the Faculty Board of Methodology (protocol № 10 from 26.05.23)