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

Department of computer
systems software

INFORMATION RETRIEVAL SYSTEMS AND SERVICES

Syllabus

– Requisites of the Course

Cycle of Higher Education	***Second cycle of higher education (Master's degree)***
Field of Study	*112 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	*1 year, (1 semester)*
Scope of the discipline	*Lectures: 36 hours, computer workshop: 36 hours and 66 hours of self-study.*
Semester control/ control measures	*Exam, modular control work, calendar control*
Course Schedule	*According to the schedule for the autumn semester of the current academic year (rozklad.kpi.ua)*
Language of Instruction	*English*
Course Instructors	Lecturer: PhD, assistant, Volodymyr Pogorelov, volodymyr.pogorelov@gmail.com Teacher of practical work: PhD, assistant, Volodymyr Pogorelov, volodymyr.pogorelov@gmail.com
Access to the course	Google classroom: https://classroom.google.com/u/0/c/MzE5ODIyMjA4MTA3?hl=ua

– Outline of the Course

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

The ***role*** *of discipline "Information Retrieval Systems and Services" is to develop the competencies necessary for solving practical problems of professional activity related to the development of software for information and search systems.*

The ***purpose*** *of studying the discipline "Information Retrieval Systems and Services" is the formation of students' abilities to independently design and develop software that implements methods and algorithms for searching data in information and search systems and services.*

The subject of the discipline "Information Retrieval Systems and Services" are methods, algorithms and models used to develop information and search systems.

Studying the course "Information Retrieval Systems and Services" forms general competences (GC) and professional competences (PC) in students:

PC03 *Ability to design software architecture, model the operation of individual subsystems and modules.*
PC05 *Ability to develop, analyze and apply specifications, standards, rules and guidelines in the field of software engineering.*
PC10 *Ability to plan and perform research in software engineering.*
PC12 *Ability to design complex multimedia and information retrieval systems.*
PC13 *Ability to design and construct, implement and maintain web-based software systems to implement new information retrieval methods.*
PC14 *Ability to implement and maintain information systems.*

Studying of the discipline "Information Retrieval Systems and Services" contributes to the formation in students of the following **program learning outcomes** *(PLO) according to the educational program:*
PLO04 *Identify information needs and classify data for software design.*
PLO18 *Develop mathematical and software for research in software engineering.*
PLO20 *Plan and perform research in the software engineering area, choose methods and tools, analyze the results, justify the conclusions.*
PLO21 *Know the theoretical foundations underlying research methods of information systems and software, research methodologies and computational experiments.*
PLO23 *Know the principles of building software information retrieval systems.*
PLO26 *Know and be able to apply in practice specialized templates for designing information retrieval systems.*
PLO27 *Be able to design and develop multi-agent information retrieval systems.*
PLO28 *Be able to design and develop distributed and centralized 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)

The successful study of the discipline "Information Retrieval Systems and Services" is preceded by the study of the disciplines "Mathematical Analysis", "Linear Algebra and Analytical Geometry", "Probability Theory", "Data Structures and Algorithms", "Programming" and "Software Support of Information and Retrieval Systems" of the curriculum for the preparation of bachelors in the specialty 121 Software engineering.
The theoretical knowledge and practical skills obtained during the mastering of the discipline " Information Retrieval Systems and Services" ensure the successful completion of course projects and master's theses in the specialty 121 Software engineering.

3. Content of the course

The discipline "Information Retrieval Systems and Services" involves the study of the following topics:

Topic 1. Introduction to search engines and services

Topic 2. Inverted indexes

Topic 3. Web graph and link analysis

Topic 4. Infrastructure beyond the index

Topic 5. Users and advertising

Modular control work

Exam

4. Coursebooks and teaching resources

Basic literature:

1. Search algorithms in information systems: method. river / O.L. Sukhyi, V.M. Milenin, V.M. Taradaynik - K.: Institute of the Gifted Child of the National Academy of Sciences of Ukraine, 2015. - 70 p.
2. Rudenko V.D. Databases in information systems K.: Phoenix, 2010, - 235 p.
3. Nesterenko O.V. Intelligent decision support systems: training. manual/ O.V. Nesterenko, O.I. Savenkov, O.O. Falovsky Under the editorship P.I. Bidyuk – Kyiv: National Academy management. - 2016. - 188 p.
4. Introduction to software engineering: teaching. manual / E.V. Levus, N.B. Melnyk - Lviv: Publishing House of Lviv Polytechnic, 2017. - 280 p.

Additional literature:

5. S. Abiteboul, M. Preda, and G. Cobena. Adaptive on-line page importance computation. In Proc. 12th International WWW Conference (WWW2003), Budapest, Hungary, pages 280–290, 2003.
6. G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. Knowledge and Data Engineering, IEEE Transactions on, 17(6):734–749, 2005.
7. G. Aggarwal, S. M. Muthukrishnan, D. Pal, and M. Pal. General auction mechanisms for search advertising. In Proc. 18th International World Wide Web Conference (WWW'2009), pages 241–250, April 2009.
8. S. Amer-Yahia, M. D. Choudhury, M. Feldman, N. Golbandi, R. Lempel, and C. Yu. Automatic construction of travel itineraries using social breadcrumbs. In Proc. 21st ACM Conference on Hypertext and Hypermedia (Hypertext'2010), pages 35–44, June 2010.
9. C. Anderson. The Long Tail - Why the Future of Business is Selling Less of More. Hyperion Books, New York NY, 2006.
10. V. Anh and A. Moffat. Index compression using fixed binary codewords. In Proc. of the 15th Int. Australasian Database Conference, pages 61–67, 2004. 4
11. A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the web. ACM Transactions on Internet Technology, 1(1):2–43, 2001.
12. R. Baeza-Yates. Graphs from search engine queries. In Proc. 33rd conference on Current Trends in Theory and Practice of Computer Science, SOFSEM'07, pages 1–8, 2007.
13. R. Baeza-Yates, A. Gionis, F. Junqueira, V. Murdock, V. Plachouras, and F. Silvestri. The Impact of Caching on Search Engines. In SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, 2007. ACM Press.
14. R. Baeza-Yates, C. Hurtado, and M. Mendoza. Improving search engines by query clustering. J. Am. Soc. Inf. Sci. Technol., 58:1793–1804, October 2007.
15. R. Baeza-Yates, F. Junqueira, V. Plachouras, and H. F. Witschel. Admission Policies for Caches of Search Engine Results. In SPIRE, 2007.

Use to master the practical skills of the discipline. The materials are freely available on the Internet.

– Educational content

5. Methods of mastering an educational discipline (educational component)

№	*Type of training session*	*Description of the training session*
Topic 1. Introduction to search systems and services		
1	*Lecture 1. Course content, introduction to search engines, technical overview of search engine components*	*Overview of course content, introduction to search engines. Introduction to information retrieval: Boolean*

		model, vector model, TF/IDF indicator. Technical overview of search engine components *Tasks on self-study: p. 6 No. 1.*
2	*Lecture 2. Probabilistic search for information, lemma Neumann-Pearson*	*Probabilistic information search. Neumann-Pearson Lemma. Linguistic models in information retrieval* *Tasks on self-study: p. 6 No. 2.*
Topic 2. Inverted indexes		
3	*Lecture 3. Basics: concept of inverted index; efficient construction of the index; the operations it supports; additional payload, which is usually stored in search engines; accompanying lexicon*	*Basics: the concept of an inverted index. Methods of effective construction of the index. Index supported operations. An additional payload that is usually stored by search engines. Companion lexicon B-Tree, Min-Heap* *Tasks on self-study: p.6 №3.*
4	*Lecture 4. Request evaluation schemes: term by term (term-at-a-time) and document by document (doc-at-a-time),* *heaps of results, early termination/simplification, WAND*	*Request evaluation schemes: term by term (term-at-a-time) and document by document (doc-at-a-time). Identification of a near-duplicate page. Heaps of results, early termination/simplification, WAND* *Tasks on self-study: p. 6 No. 4.*
5	*Lecture 5. Compressing the index and changing the order of documents*	*Compressing the index and changing the order of documents. Apache Lucene Search Library (prerequisite for homework)* *Tasks on self-study: p. 6 No. 5.*
6	*Lecture 6. Distributed Index Architectures: Global/Local Schemas, Combinatorial Data Distribution Problems, Google Cluster Architecture*	*Architectures of distributed indexes. Global/local schemas, combinatorial data partitioning problems, Google's cluster architecture.* *Tasks on self-study: p. 6 No. 6.*
7	*Computer workshop 1. Construction of inverted indexes*	*Task: Analyze software tools and build an index in the information search system.* *Tasks on self-study: p. 6 No. 7.*
Topic 3. Web graph and link analysis		
8	*Lecture 7. Web graph structure: power laws, Bow-tie structure, self-similarity*	*Web graph structure: power laws, Bow-tie structure, self-similarity* *Tasks on self-study: p. 6 No. 8.*
9	*Lecture 8. Fundamentals of link analysis: Google PageRank, Kleinberg HITS, a brief overview of Perron-Frobenius theory and ergodicity*	*Basics of link analysis: Google PageRank (topic sensitive), Kleinberg HITS, a brief overview of Perron-Frobenius theory and ergodicity* *Tasks on self-study: p. 6 No. 9.*

10	Computer workshop 2. Link analysis	Task: Using software to implement a link analysis module for the information system. Tasks on self-study: p. 6 No. 10.
11	Lecture 9. Stability and similarity of schemes based on connections, TKC effect	Stability and similarity of connection-based schemes, TKC effect. Evolutionary models of the web graph Tasks on self-study: p. 6 No. 11.
Topic 4. Infrastructure beyond the index		
12	Lecture 10. Crawlers. Purpose and architecture, optimization of scan order, calculation of importance metrics during scanning	Crawlers. Purpose and architecture, optimization of scan order, calculation of importance metrics during scanning. Bloom filters Tasks on self-study: p. 6 No. 12.
13	Lecture 11. Effective caching and prefetching of query results	Efficient caching and prefetching of query results Tasks on self-study: p. 6 No. 13.
Topic 5. Users and advertising		
14	Lecture 12. Computer advertising: models and definitions. CPM, CPC, CPA; sponsored search (adwords), content matching (adsense), media advertising	Computer advertising: models and definitions. CPM, CPC, CPA. Sponsored search (adwords), content matching (adsense), media advertising Tasks on self-study: p. 6 No. 14.
15	Lecture 13. Computer advertising: auction mechanisms	Computer Advertising (TBD): Models and Definitions. CPM, CPC, CPA; sponsored search (adwords), content matching (adsense), media advertising Tasks on self-study: p. 6 No. 15.
16	Lecture 14. Extraction and interception of implicit content created by users	Mining and intercepting implicit user-generated content. Query log analysis Tasks on self-study: p. 6 No. 16.
17	Lecture 15. Task execution and search assistance - from spelling correction and simple shortcuts to multimedia, mashups, query completion and aspects	Task completion and search assistance – from spelling corrections and simple shortcuts to multimedia, mashups, query completion, and aspects. Tasks on self-study: p. 6 No. 17.
18	Lecture 15. Long Tail, recommendation systems and joint filtering	Long Tail, Recommender Systems and Collaborative Filtering. Context-sensitive search and user modeling Tasks on self-study: p. 6 No. 18.
19	Computer workshop 3. Building a context-sensitive search system	Task: to build a context-sensitive search system. Tasks on self-study: p. 6 No. 19.
Modular control work		

6. Self-study

The discipline "Information Retrieval Systems and Services" is based on independent preparations for classroom classes on theoretical and practical topics.

№	The name of the topic submitted for self-study	Number of hours	Literature
1	Preparation for the lecture 1	1	1; 2; 3; 8
3	Preparation for the lecture 2	2	1; 4, 7
4	Preparation for the lecture 3	2	1; 3; 5; 9; 10
5	Preparation for the lecture 4	2	11-13
6	Preparation for the lecture 5	1	3; 6; 10
7	Preparation for the lecture 6	1	3; 8; 10
8	Preparation for a computer workshop 1	5	11-13
9	Preparation for the lecture 6	2	2; 5
10	Preparation for the lecture 7	2	2; 9
11	Preparation for the lecture 8	1	11-13
12	Preparation for a computer workshop 2	5	2; 6
13	Preparation for the lecture 9	1	1; 2; 3
14	Preparation for the lecture10	2	11-13
15	Preparation for the lecture 11	2	4; 5
16	Preparation for the lecture 12	2	4; 5
17	Preparation for the lecture 13	2	9-10
18	Preparation for the lecture 14	2	4, 12
19	Preparation for the lecture 15	1	4, 10
20	Preparation for a computer workshop 3	5	3, 11-13
27	Preparation for modular control work	10	1-15
28	Preparation for the exam	16	1-15

– Policy and control

7. Policy of academic discipline (educational component)

Attending lectures is mandatory.
Attending computer workshop classes may be occasional and as needed for consultation/protection of computer workshop works.
Rules of behavior in classes: activity, respect for those present, turning off phones.
Adherence to the policy of academic integrity.
Rules for protecting the work of the computer workshop: the work must be done in accordance with the tasks and according to the option.
The rules for assigning incentive and penalty points are as follows. Incentive points are awarded for:
- accurate and complete answers in surveys based on lecture materials (maximum number of points for a blitz survey - 3 points).

8. Types of control and rating system for evaluating learning outcomes (PCO)

During the semester, students perform 3 computer practicals. The maximum number of points for each computer workshop: 15 points.

Points are awarded for:

- quality of the computer workshop: 0-10 points;

- answer during the defense of the computer workshop: 0-6 points;

- timely submission of work for defense: 0-4 points.

Performance evaluation criteria:

10 points – the work is done qualitatively, in full;

8-9 points - the work is done qualitatively, in full, but has shortcomings;

6-7 points – the work is completed in full, but contains minor errors;

2-5 points – the work is completed in full, but contains significant errors;

0 points - the work is not completed in full.

Answer evaluation criteria:

3 points – the answer is complete, well-argued;

2 points – the answer is correct, but has flaws or minor errors;

1 points – there are significant errors in the answer;

0 points - there is no answer or the answer is incorrect.

Criteria for evaluating the timeliness of work submission for defense:

2 points – the work is presented for defense no later than the specified deadline;

0 points – the work is submitted for defense later than the specified deadline.

The maximum number of points for performing and defending computer practicals:

15 points × 3 comp. practice = 45 points.

During the semester, lectures take place on the topic of the current lesson. Maximum points for all surveys: 3 points. The number of surveys on the topic of the current lesson for one student is unlimited.

The assignment for the modular test consists of 3 theoretical and 2 practical questions. The answer to each question is evaluated by 3 points.

Evaluation criteria for each test question:

3 points – the answer is correct, complete, well-argued;

2 points - in general, the answer is correct, but has flaws;

1 points – there are significant errors in the answer;

0 points - there is no answer or the answer is incorrect.

The maximum number of points for a modular control work:

3 points × 5 questions = 15 points.

The rating scale for the discipline is equal to:
$R = R_C$ = 45 points + 15 points + 40 points = 100 points.
According to the description: $R = R_{comp.practice} + R_{МКР} + R_{exam}$ = 45+15+40 points = 100 points

Calendar control: conducted twice per semester as a monitoring of the current status of meeting the syllabus requirements.
At the first certification (8th week), the student receives "credited" if his current rating is at least 10 points (50% of the maximum number of points a student can receive before the first certification).
At the second certification (14th week), the student receives "passed" if his current rating is at least 20 points (50% of the maximum number of points a student can receive before the second certification).

Semester control: exam
Conditions for admission to semester control:

With a semester rating (rC) of not less than 24 points and the enrollment of all the work of the computer workshop, the student is admitted to the exam. After passing the exam, a grade is assigned according to the table (Table of correspondence of rating points to grades on the university scale).
Completion and defense of a computer workshop is a necessary condition for admission to the exam.

Table of correspondence of rating points to grades on the university scale:

Points	*Grade*
100-95	Excellent
94-85	Very good
84-75	Good
74-65	Satisfactory
64-60	Sufficient
Below 60	Fail
Course requirements are not met	Not Graded

9. Additional information about the course

The list of questions submitted for semester control is given in Appendix 1.

Working program of the academic discipline (syllabus):

Is designed by teacher PhD, assistant, Volodymyr Pogorelov;

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)

Appendix 1. List of questions submitted for semester control

1. Probabilistic search for information, Neumann-Pearson lemma
2. Concept of inverted index; efficient index construction.
3. Additional payload, which is usually stored in search engines; accompanying lexicon.
4. Request evaluation schemes: term by term (term-at-a-time) and document by document (doc-at-a-time).
5. Search engine components.
6. Efficient caching and prefetching of query results.
7. Google's cluster architecture.
8. Distributed index architectures: global/local schemes, combinatorial problems related to data distribution.
9. Web graph structure: power laws, Bow-tie structure, self-similarity.
10. Basics of link analysis: Google PageRank, Kleinberg HITS, a brief overview of Perron-Frobenius theory and ergodicity.
11. Stability and similarity of schemes based on connections, TKC effect.
12. Crawlers. Purpose and architecture, optimization of scan order, calculation of importance metrics during scanning.
13. Computer advertising: models and definitions. CPM, CPC, CPA; sponsored search (adwords), content matching (adsense), media advertising.
14. Mining and Interception of Implicit User Generated Content.
15. Compressing the index and changing the order of documents.
16. Long Tail, Recommender Systems and Collaborative Filtering.
17. Computer advertising: auction mechanisms.
18. Purpose and architecture, optimization of scan order, calculation of importance metrics during scanning.
19. Mining and Interception of Implicit User Generated Content.
20. Search tasks, means and technologies of information search.