COURSE DESCRIPTION CARD - SYLLABUS

Course name
Information retrieval and search engines [N1Inf1>ZI]

Course

Field of study Computing	Year/Semester 4/7
Area of study (specialization) –	Profile of study general academic
Level of study first-cycle	Course offered in Polish
Form of study part-time	Requirements elective

Number of hours

Lecture 12	Laboratory classes 12	Other 0
Tutorials 0	Projects/seminars 0	

Number of credit points

2,00

Coordinators

dr inż. Irmina Masłowska
irmina.maslowska@put.poznan.pl

dr hab. inż. Miłosz Kadziński prof. PP
milosz.kadzinski@put.poznan.pl

Lecturers

Prerequisites

Students taking this course should have basic knowledge of object programming, algorithms and data structures, statistic and data analysis, linear algebra, and artificial intelligence. They should be capable of formulating and solving fundamental mathematical programming problems, constructing an object model of a simple system, programming in at least one object programming language, and collecting information from the indicated sources. Within the social competencies, they should understand that knowledge and skills in computer science quickly become outdated. They must present honesty, responsibility, perseverance, cognitive curiosity, creativity, personal culture, and respect for others.

Course objective

1. Conveying knowledge of primary methods of information retrieval, indexing, and search models. 2. Explaining the basics of natural language processing. 3. Presenting methods for ranking web resources concerning their adequacy for the user query and web structure and evaluating the quality of search results. 4. Explaining the fundamental laws of describing the structure of links between web resources. 5. Presenting the application of data mining and machine learning techniques for pattern mining in analyzing web resources and user behavior. 6. Explaining selected threats related to the use of the Internet. 7. Extending the capabilities of using data analysis, linear algebra, artificial intelligence, and machine learning for analyzing the content of information resources, the structure of links between these resources, and usage patterns. 8. Revealing how to interpret the results of using the methods mentioned above in the context of analysis of content, structure, and usage of information resources.

Course-related learning outcomes

Knowledge:
1. has extended, in-depth knowledge of mathematics on matrix process, probability theory, and graph theory - [K1st_W1]
2. has a detailed, well-grounded knowledge of information retrieval and search methods, algorithms and their complexity, artificial intelligence, data analysis, natural language processing, and machine learning - [K1st_W4]
3. has the knowledge needed for analyzing and processing information resources (in particular, collecting, processing, and ordering semi-structured data) and selecting a suitable method for realizing these tasks - [K1st_W4]
4. has a basic knowledge of key directions and the most important successes of computer sciences and related disciplines within the scope of information retrieval and search - [K1st_W5]
5. knows and understands the basic techniques, methods, algorithms, and tools used for solving computer problems related to information retrieval and search as well as natural language processing - [K1st_W7]
6. knows selected threats faced when using the Internet - [-]
Skills:
1. can formulate and solve complex problems within the scope of information retrieval and search by applying appropriately selected methods (including analytical, simulation, or experimental approaches) of data analysis, linear algebra, artificial intelligence, natural language processing, and machine learning for analyzing the context, structure, and usage of web resources - [K1st_U4]
2. can interpret the results of applying the above-mentioned methods in the context of web content, structure, and usage mining; can collect information about information retrieval and search from various sources, including the relevant literature and databases (in Polish and English), integrate them appropriately, interpret and conduct a critical analysis - [K1st_U1]
3. can employ information and communication tools at different stages of carrying out IT projects related to information retrieval and search - [K1st_U2]
4. can design - following a pre-defined specification - and create an information retrieval and search system by first selecting and then using the available methods, techniques, and computer tools (including programming languages) - [K1st_U10]
5. can formulate and implement the algorithms using at least popular tools serving for information retrieval and search - [K1st_U11]
6. can plan and carry out life-long learning related to information retrieval and search - [K1st_U19]
Social competences:
1. understands that knowledge and skills related to information retrieval and search quickly become outdated [K1st_K1]
2. knows the examples of poorly functioning information retrieval and search systems, which led to economic, social, or environmental losses - [K1st_K2]
3. can think and act in an enterprising way, finding the commercial application for the created IR systems, having in mind the economic benefits as well as legal and social issues - [K1st_K3]

Methods for verifying learning outcomes and assessment criteria

Learning outcomes presented above are verified as follows:

Lecture: Verification takes place during a written exam. Students solve several computational tasks concerning the subjects presented during all lectures and answer some short questions. Alltogether there are at most 20 tasks. Each task is evaluated individually, being allocated a certain number of points. In

order to receive a passing grade, the student must score more than 50% of the total points.
Laboratory classes: After each class, students solve practical programming assignments and report their solutions to the instructors leading the laboratory classes. Each assignment is evaluated. The number of points can be increased by discussing additional aspects of the task, efficiently using the comprehended knowledge for problem-solving, and indicating perception problems that might improve the didactic process.

Programme content

Methods of information retrieval, indexing, and search models for unstructured/semi-structured data (like texts), basics of natural language processing, methods for ranking web resources concerning their adequacy for the user query and web structure and evaluating the quality of search results, fundamental laws of describing the structure of links between web resources, application of data mining and machine learning techniques for pattern mining in analyzing web resources and user behavior.

Course topics

Lectures: Classification of web resources and access methods. Review of methods and applications of Web mining: analysis of content, structure, and usage. Characteristics of the description level of natural language and the respective sub-fields of linguistics. The phases of pre-processing natural language to enable the search: lexical analysis, stopwords, stemming/lemmatization, selection of index terms, and construction of categorical structures. Spelling correction (Levenshtein distance, edit distance). POS-tagging. Identification of meaning and relations between terms. Representation of texts, TF-IDF representation.
Similarity measures for text documents. Classical and non-classical search models (Boolean, probabilistic, VSM, fuzzy, neural networks, LSI incorporating SVD decomposition).
Search systems - history, architecture, working rules, methods of organization, and presentation of results. Ranking of web documents in terms of their adequacy to query; HITS, PageRank, and their extensions. Aspects considered by the web searchers. Evaluation of the quality of the search results, including precision and recall. Spamming techniques; preventing spamming.
Indexing of text documents, basic kinds of indices, and their applications. Inverse index, trees, and suffix arrays - time and memory complexity. Algorithm for constructing inverted indices for large text collections. Distributed indexing, MapReduce model. Analysis of web structure: Bowtie model, power law, Zipf law. Characteristics of log files, methods for discovering and analyzing patterns - statistic methods, data mining, and machine learning. Classification and clustering of documents, users, or behavior patterns. Analysis of usage patterns for personalization of content and web services, application in e-commerce, collaborative filtering, and recommender systems. Opinion mining and
sentiment analysis - an exploration of opinions available on the Internet: identification, classification, summarization, search. Spamming of web opinions and recommender systems, spam hiding methods, and spam identification methods.
Laboratory classes: organized in the form of 2-hour exercises, conducted in the laboratory. Individual issues discussed during the lecture are illustrated with tasks during laboratory classes. Text processing: tokenization, stop words, normalization, stemming, lemmatization. Practical use of the vector space model (TF-IDF and cosine similarity) for ranking text documents according to their suitability for a given query. Using OpenNLP package for natural language processing, POS tagging, parsing, and sentiment analysis. Practical use of HITS and PageRank for ranking websites based on link structure. Log server files processing and basics of exploratory data analysis. Investigating the success rate of ad campaigns. Algorithms for constructing indices, suffix trees, and arrays. Using Lucene for indexing, parsing, and ranking documents. Using Tika for analyzing and parsing documents of various types. Developing a web crawler.

Teaching methods

Lecture: slide show presentations on different sub-fields of information retrieval and search, illustrated with examples and practical assignments.
Laboratory classes: solving illustrative examples on board and coding problem solutions in Python, conducting computational experiments, discussing the chosen methods, teamwork, demonstration of selected information retrieval systems, and multi-media show.

Bibliography

1. Eksploracja zasobów internetowych, Z.Markov, D.T.Larose, PWN, 2009
2. Introduction to Information Retrieval, Ch.D.Manning, P.Raghavan, H.Schütze, Cambridge University Press, 2008 (available online: https://nlp.stanford.edu/IR-book/)
3. Mining of Massive Datasets, Anand Rajaraman, Jeffrey David Ullman, Cambridge University Press, 2011 (available online: http://infolab.stanford.edu/~ullman/mmds/book.pdf)
4. Modern Information Retrieval, Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Addison-Wesley, 1999
5. Data intensive text-processing with MapReduce, Jimmy Lin, Chris Dyer, University of Maryland, Morgan & Claypool Synthesis, 2010 (available on-line: https://lintool.github.io/MapReduceAlgorithms/MapReduce-book-final.pdf)

Additional:
1. Speech and Language Processing (3rd ed. draft), D. Jurafsky and J.H. Martin (2024 version available online: https://web.stanford.edu/~jurafsky/slp3)
2. Foundations of Statistical Natural Language Processing, Ch.D.Manning, H. Schütze, MIT Press, Cambridge Massachusetts, MIT Press Cambridge Mass, 1999
3. Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. B. Liu, Springer, 2009
4. Mining the Web: Discovering Knowledge from Hypertext Data. S. Chakrabarti, Morgan Kaufmann, 2002
5. The Text Mining Handbook. R. Feldman, J. Sanger, Cambridge University Press, 2006
6. Surveys published at https://searchenginewatch.com, https://searchengineland.com/

Breakdown of average student's workload

	Hours	ECTS
Total workload	50	2,00
Classes requiring direct contact with the teacher	24	1,00
Student's own work (literature studies, preparation for laboratory classes/ tutorials, preparation for tests/exam, project preparation)	26	1,00