COURSE DESCRIPTION CARD - SYLLABUS

Course name
Information retrieval [S1SI1E>WINF]

Course

Field of study
Artificial Intelligence

Year/Semester
3/5

Area of study (specialization)
–

Profile of study
general academic

Level of study
first-cycle

Course offered in
English

Form of study
full-time

Requirements
compulsory

Number of hours

Lecture
15

Laboratory classes
15

Other
0

Tutorials
0

Projects/seminars
0

Number of credit points

3,00

Coordinators

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

Lecturers

Prerequisites

Basic mathematical knowledge from mathematical analysis and linear algebra. Programming skills in Python. Knowledge acquired during the courses on Introduction to AI, Data Mining, and Machine learning.

Course objective

The course aims at conveying the classical material in information retrieval. Each of the lectures will have its main hero - Boolean retrieval, web usage mining, the vector space model, famous PageRank and HITS algorithms, query expansion, collaborative filtering, index construction, and the Map-Reduce framework. The course will make it easier for you to understand some aspects that will be covered during the sixth semester in the courses on Natural Language Processing or parallel systems.

Course-related learning outcomes

Knowledge:
K1st_W3: has a well-grounded knowledge of fundamental computer science problems within the scope of information retrieval, inculding web content, structure, and usage mining
K1st_W4: knows and understands the basic techniques, methods, algorithms, and tools used for solving computer problems as well as problems in information retrieval (e.g., indexes and retrieval systems)

K1st_W5: has a basic knowledge of key directions and the most important successes of information retrieval understood as an essential sub-domain of artificial intelligence, making use of the achievements of other scientific disciplines and providing solutions with a high practical impact

Skills:
K1st_U1: understands that knowledge and skills quickly become outdated in computer science and, in particular, AI, and perceives the need for constant additional training and raising one"s qualifications.
K1st_U3: can formulate and solve complex decision problems within the scope of information retrieval (e.g., finding relevant information or evaluting the documents" relevance to the query), by applying appropriately selected methods
K1st_U4: can efficiently plan and carry out experiments, including computer measurements and simulations, interpret the obtained results and draw conclusions based on the experimental outcomes in the context of information retrieval tasks (e.g., assessing the quality of the results returned by the IR systems)
K1st_U9: can adapt the existing algorithms as well as formulate and implement the novel algorithms in Python, including the algorithms typical for different IR tasks
K1st_U10: can retrieve, analyze and transform different types of data (while focussing on unstructured data), and carry out data synthesis to knowledge and conclusions useful for solving a variety of information retrieval tasks
K1st_U11: can adapt and make use of the models of intelligent behavior

Social competences:
K1st_K1: understands that knowledge and skills quickly become outdated in information retrieval, and perceives the need for constant additional training and raising one"s qualifications.
K1st_K2: is aware of the importance of scientific knowledge and research related to AI in solving practical problems which are essential for the functioning of individuals, firms, organizations as well as the entire society
K1st_K5: can think and act in an enterprising way, finding the commercial application for the created AI-based systems, having in mind the economic benefits as well as legal and social issues

Methods for verifying learning outcomes and assessment criteria

Learning outcomes presented above are verified as follows:

Learning outcomes presented above are verified as follows:
Lecture: Assessment test is conducted at the last lecture. The students need to solve several computational task concerning the subjects presented during all lectures. Each task is evaluated individually, being allocated a certain number of points. The points are summed up and a standard scale is used to derive the final marks: <50% - 2.0, [50% , 60%) - 3.0, [60% , 70%) - 3.5, [70% , 80%) - 4.0, [80% , 90%) - 4.5, and [90% , 100%] - 5.0.
Laboratory classes: After each class, students solve practical, programming assignments and report their solutions to the instructors leading the laboratory classes within two weeks. Each assignment is evaluated on a scale from 2.0 to 5.0. The final grade is computed as an average from the individual marks.

Programme content

The course introduces you to the world of information retrieval. It is closely related to search, recommendation, and web mining. The covered topics offer an overview of the classical but diverse subjects in information retrieval. The most essential ones include text processing, vector space model, PageRank and HITS, query expansion, collaborative filtering, index construction and compression, and Map-Reduce.

Course topics

Introduction to information retrieval - Boolean retrieval, text processing, and mining navigational patterns: information retrieval, types of data, web mining, unstructured data search, indexing - incidence matrix, Boolean retrieval model, query optimization, major steps in inverted index construction (tokenization, stop words, normalization, stemming, lemmatization), web usage analysis, log files, user identification, sessionization, path completion, navigational patterns, Markov chains. Vector space model and latent semantic indexing: complete search system, binary representation, bag

of words, term frequency, inverse document frequency, vector space model, cosine similarity, term-document matrix, principal component analysis, latent semantic indexing.
Evaluation in information retrieval and PageRank: evaluation of IR systems, measures: precision, recall, accuracy, measure F, precision and recall at k, mean average precision, R-precision, web structure, PageRank algorithm, link farm, TrustRank algorithm, Google Penguin and Panda.
HITS, relevance feedback, and spelling correction: HITS - Hubs and Authorities, query refinement methods, relevance feedback, Rocchio algorithm, pseudo-relevance feedback, thesaurus, spelling correction, Levenshtein distance, Soundex.
Recommender systems and Adwords: famous recommender systems, content-based recommendation, user-based collaborative filtering, item-based collaborative filtering, slope one predictor, Adwords problem, BALANCE algorithm.
Index construction and compression: inverted index, positional inverted index, K-gram index, suffix tree, naive construction algorithm, Ukkonen algorithm, suffix array, qsufsort algorithm, data compression, Heaps" law, Zipf"s law, binary and unary coding, gamma coding, delta coding.
Introduction to MapReduce: big ideas behind processing big data, Map-Fold, what is MapReduce, processing (key, value) pairs, Mapper, Reducer, Combiner, Partitioner, word count, an average of evaluations, word co-occurrence matrix, inverted indexing, retrieval, PageRank, when is MapReduce (less) useful?

Teaching methods

Lecture: slide show presentations on information retrieval methods, illustrated with examples and practical assignments that serve as a summary of the lectures and preparation for the assessment test.
Laboratory classes: solving illustrative examples on board and coding problem solutions in Python, conducting computational experiments, discussion on the chosen methods, teamwork.

Bibliography

Basic
C. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008, http://nlp.stanford.edu/IR-book/
A. Rajaraman, J. Ullman, Mining of Massive Datasets, Cambridge University Press, 2011 http://infolab.stanford.edu/~ullman/mmds/book.pdf)
R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley, 1999
J. Lin, C. Dyer, 5. Data intensive text-processing with MapReduce, J. Lin, C. Dyer, University of Maryland, Morgan & Claypool Synthesis, 2010, https://lintool.github.io/MapReduceAlgorithms/MapReduce-book-final.pdf
Additional
D. Jurafsky, J.H. Martin, Speech and Language Processing, https://web.stanford.edu/~jurafsky/slp3
C. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge Massachusetts, MIT Press Cambridge Mass, 1999
B. Liu, Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. Springer, 2009
S. Chakrabarti, Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann, 2002
R. Feldman, J. Sanger, The Text Mining Handbook. Cambridge University Press, 2006

Breakdown of average student's workload

	Hours	ECTS
Total workload	75	3,00
Classes requiring direct contact with the teacher	30	1,50
Student's own work (literature studies, preparation for laboratory classes/ tutorials, preparation for tests/exam, project preparation)	45	1,50