[*Information Retrieval for Data Science*]

Course Administrative Details

Course Title	*Information Retrieval for Data Science*		
Instructor(s)	*Stanislav Protasov*	**Instructor's e-mail**	*s.protasov@innopolis.ru*
Course #	*XXX*	**Course Type**	*Core*
Faculty	*Computer Science and Engineering*	**Major**	*Data Science*
Academic year	*2019-2020*	**Semester Offered**	*Spring*
No. of Credits	*6 ECTS*	**Total workload on average**	*3 hours in auditory per week* *3 hours homework per week* *20 hours project*
Lecture Hours	*2 per week*	**Lab Hours**	*2 per week*
Language	*English*	**Frequency**	*Weekly*
Target Audience	*Masters*	**Anticipated Enrollment**	*30 students*
Studying year	*1 (5)*		
Grading Mode	*A, B, C, Fail*	**Keywords**	*Information retrieval, information search, search engines, data mining*

Course outline

The course is designed to prepare students to understand and learn contemporary tools of information retrieval systems. The course will focus on applying mathematical and programming tools to building such systems. Throughout the course, students will be involved in discussions, readings and assignments to experience real world systems. The technologies and algorithms covered in class includes advanced algorithms, linear algebra, machine learning, data mining, natural language processing and so on.

Course Delivery

The course will be given weekly from January to April 2020. Every week, there will be a 2-hour lecture followed by a 2-hour lab session. Lab sessions are followed by homework. In the end of the semester students at the exam present their projects – information retrieval services.

Prerequisite courses

Machine learning, Python, algorithms and data structures

Required background knowledge

Solid knowledge of imperative and object-oriented programming concepts and good programming skills in Python are assumed. Good understanding of machine learning together with a skill to run models required. Basic probability theory and statistics understanding is also required.

Course structure

[IA – Individual Assignment]

Week# / Date	***Topic***	***Assignments***
Week 1	*Introduction to information retrieval*	*Non-graded introduction test*
Week 2	*Building inverted index. Language, tokenization, stemming, searching, scoring.*	*Inverted index service with scoring*
Week 3	*Distributive semantics. Vector model. Dimension reduction.*	*Vector representation*

Week 4	*ML approaches to vector modelling.*	*ML for document representation*
Week 5	*Indexing for vector model. Kd-trees, quad trees, Annoy, FAISS, HNSW*	*Fast vector index. Recommender system*
Week 6	*Web basics. Internet crawling. XML and HTML processing. Dynamic documents processing.*	*Dynamic site parsing and robot ethics*
Week 7	*Spellchecking and query correction*	*Spellchecker*
Week 8	*Query expansion and suggest*	*Query expansion and suggest lab*
Week 9	*Language model. Topic model. Clustering and classification*	*Topic modeling*
Week 10	*On newsfeeds*	*Newsfeed*
Week 11	*Image and video processing. Understanding and enhancing*	*Image/Video to text*
Week 12	*Audio processing. Speech to text. Acoustic fingerprinting*	*Shazam*
Week 13	*Quality assessment. A/B testing, SBS, pFound, DGC, nDGC.*	*A/B test framework*
Week 14	*Web search specific topics. PageRank. Duplicates. CTR.*	*PageRank*
Week 15	*[Extension slot: selected topic]*	--

Textbook(s)

An Introduction to Information Retrieval by Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Cambridge University Press (any edition)

Reference Materials

TBD

Computer Resources

Students should have laptops with Anaconda installer.

Laboratory Exercises

Tutorial exercises will be set on weekly basis

Laboratory Resources

No laboratory resources are required for this course.

Grading criteria

Homework = 50%, Final Exam (project defense) = 50%

Late Submission Policy

No late submission allowed.

Cooperation Policy and Quotations

We encourage vigorous discussion and cooperation in this class. You should feel free to discuss any aspects of the class with any classmates. However, we insist that all assignments should be done by you alone. We will run automatic code comparison software on assignment submissions to compare solutions. Violations of this policy will be investigated and will result in zero scores on any assignments brought under suspicion.