

Title of course	Text Mining and Search
Responsible instructor	<i>Prof. Dr. Englmeier</i>
Learning objectives	<p>Knowing/Perceiving: Students learn essentials in content extraction and information retrieval as the basis of content analysis in texts, which, in turn, provide the theoretical basis for the successful design of advanced content analysis.</p> <p>Applying: The students implement the methods they learn while using well-established tools for data analysis (for example, Apache Lucene), which are valuable for the design of search engines.</p> <p>Analyzing/Evaluating: In the teamwork of the project, the students apply their theoretical design knowledge in the development of specialty search engines. They embrace thus the design versatility in the development of features for text analysis and retrieval. In their practical work they can reflect the effectiveness and potentials of their design approaches.</p> <p>Synthesizing: The result of the course is manifested in a course-wide project that involves the development of a search engine with special search features. Application development is thereby broken down into smaller work packages. Each team (two or three students) assumes a work package, organizes its individual tasks, and contributes to the management of the overall project.</p> <p>The self-empowered organization of the project work also includes explorative learning. Students are so encouraged to learn new methodologies or tools on their own (with support from the professor), provided their individual part of the project work requires that.</p>
Course contents	<ol style="list-style-type: none"> 1. <i>Fundamentals in Information Retrieval (IR)</i> <ul style="list-style-type: none"> • Basic IR concepts • Regular Expressions • XML 2. <i>User Interaction</i> <ul style="list-style-type: none"> • User story structure & validation • Feature charts • User support 3. <i>Retrieval models & evaluation</i> 4. <i>Apache Lucene</i> <ul style="list-style-type: none"> • Modules • Integration (Java) 5. <i>Indexing</i> <ul style="list-style-type: none"> • Tokenization • Stop words • Stemming • Synonyms

	6. <i>Query matching</i> <ul style="list-style-type: none"> • <i>Query vectors</i> • <i>Matching models</i>
Teaching methods	<i>Lectures, workshops, team cooperation</i>
Prerequisites	<i>Solid practical programming skills</i>
Suggested reading	<i>Baeza-Yates, R.; Ribeiro-Neto, B.: "Modern Information Retrieval", ACM Press, New York, 1999.</i> <i>McCandless, M. et al: "Lucene in Action", Second Edition, Manning, Stamford, 2010</i> <i>Application examples from search engines in practice</i>
Applicability	<i>Master Applied Computer Science</i>
Workload	<i>Total 90 hours. Attendance: 30 hours, Self-Study: 30 hours, Practical work: 30 hours</i>
ECTS credit points and weighting factor	<i>3 CP (Emphasis of the Grade for the final Grade 3/120)</i>
Basis of student evaluation	<i>Project work</i>
Time	<i>2nd semester</i>
Frequency	<i>Once during the academic year (summer semester)</i>
Duration	<i>one semester</i>
Course type	<i>Compulsory elective course</i>
Remarks	<i>Teaching language is English.</i>