|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Module Name | **Text Mining and Search** | | | | | | |
| Module Responsibility | *Prof. Dr. Englmeier* | | | | | | |
| Qualification Targets | ***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.*  ***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.*  ***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.*  ***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.*  *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.* | | | | | | |
| **Content** | Knowing | Perceivin g | Applying | Analyzing | Evaluatin g | Synthesiz ing |
| Basics | X | X |  |  |  |  |
| User  interaction |  | X | X | X |  | X |
| Retrieval  models & evaluation | X | X | X | X | X | X |
| Apache Lucene |  | X | X | X | X | X |
| Content  extraction | X | X | X | X | X | X |
| Indexing |  | X | X | X | X | X |
| Query  matching |  | X | X | X | X | X |

|  |  |
| --- | --- |
| Module Contents | *1. Fundamentals in Information Retrieval (IR)*  •*Basic IR concepts*   •*Regular Expressions*   •*XML*  *2. User Interaction*   •*User story structure & validation*  •*Feature charts*   •*User support*  *3. Retrieval models & evaluation*  *4. Apache Lucene*   •*Modules*   •*Integration (Java)*  *5. Indexing*   •*Tokenization*   •*Stopwords*   •*Stemming*   •*Synonyms*  *6. Query matching*   •*Query vectors*  •*Matching models* |
| Teaching Modalities | *Lectures, workshops, team cooperation* |
| Requirements for Participation | *Solid practical programming skills* |
| Literature / Multimedia-based Teaching Material | *Baeza-Yates, R.; Ribeiro-Neto, B.: “Modern Information Retrieval”, ACM Press, New York, 1999.*  *McCandless, M. et al: “Lucene in Action”, Second Edition, Manning, Stamford, 2010*  *Application examples from search engines in practice* |
| Applicability | *Master Applied Computer Science* |
| Effort/  Total Workload | *Total 90 hours. Attendance: 30 hours, Self-Study: 30 hours, Practical work: 30 hours* |
| ECTS/ Emphasis of the Grade for the final Grade | *3 CP (Emphasis of the Grade for the final Grade 3/120)* |
| Performance Record | *Project work* |
| Semester | *2nd semester* |
| Frequency of the course | *Once during the academic year (summer semester)* |
| Duration | *One semester* |
| Type of Course | *Compulsory elective course* |