1. Credit: This is a 5-credit course.

2. Prerequisites: This course is open to all students in the Masters in Computer Science and Systems program. There are no specific prerequisites for this course.

3. Catalog Description: Covers the basic principles and techniques used in information retrieval (IR) and web search, including keyword based search, content analysis (vector space model, probabilistic language models), link analysis (PageRank), indexing, text classification, document clustering, aggregated search, user-system interaction in IR, and evaluation of IR systems.

4. Course Details: The field of information retrieval (IR) is concerned with the analysis, organization, storage, and retrieval of unstructured and semi-structured data. This course will be a practice-oriented course that focuses on: IR techniques, retrieval models, indexing, text classification, document clustering, the concept of relevance, evaluation of IR systems and search engine technologies. Given a collection of information (e.g., web) and a retrieval system (e.g., search engine), we will examine three major information retrieval steps:
   1) How to interpret the user's information need, and retrieve relevant information (queries, retrieval models, etc).
   2) How to organize and present the retrieved information so that identifying the topic of interest is easy for the user (ranking, clustering, text classification, etc).
   3) Once the retrieval, organization and presentation of information is devised, it is then required to monitor how effective or beneficial these methods are in information access (user-system interaction and evaluation).
This course will primarily focus on “text” retrieval.

5. Learning Goals: Upon successful completion of this course, the student will be able to:
   1. Explain the main techniques that major web search engines rely on to rank results for a keyword based query, including content analysis (vector space model) and link analysis (PageRank)
   2. Explain back-end as well as front-end technologies related to retrieval systems.
   3. Perform research (identify state-of the art, compare different methods, list limitations) in the area of information retrieval.
   4. Understand the concept of dictionaries and inverted index used in retrieval systems.
   7. Measure different types of system-user interaction using experimental and observational methods.
   8. Understand the importance of linear algebra and probability in IR.
6. **Textbook:**

- **Required Textbook:**
  Introduction to Information Retrieval  
  C. Manning, P. Raghavan and H. Schutze  
  Cambridge University Press, 2008  
  Available online: http://nlp.stanford.edu/IR-book/

- **Additional Resources:**
  1. Search Engines – Information Retrieval in Practice  
     W. B. Croft, D. Metzler, and T. Strohman  
     Cambridge University Press, 2009  
  2. Foundations of Statistical Natural Language Processing  
     C. Manning and H Schutze  
     MIT Press, 1999

- **Other Readings:** Selected conferences papers and chapters from other books will be assigned for reading. These will be available online.

7. **Grading:**

- **Homeworks (20%)** –
  There will be two assignments (10% each), involving problems related to concepts and techniques discussed in class, as well as experiments with various tools or systems. Unless otherwise specified, these assignments must be done individually.

- **Term Project (30%)** –
  Each student will be expected to complete a term project, as a form of problem-based learning in which the student acquires skills and understand their importance in a real context. The project will involve the design and implementation of an information retrieval system (or a specific part of a system). The project work starts with gaining an in-depth understanding (review) of a particular area of Information Retrieval (IR) that is of interest to the student. The review can take the form of a research paper examining the use of a specific IR technique or IR model, or it can be a detailed survey of a particular IR related topic. The review should contain a summary and a technical evaluation of the state-of-the-art related to the topic being studied. Broadly, the review should contain four parts: (1) Problem Description, (2) Existing approaches and solutions (3) Relevant evaluation methods (4) Limitations, Discussion and Conclusion. In addition to this review paper, the oral and written deliverables for the project include a project proposal, regular updates, documented source code, a term paper and a final presentation.

- **Midterm (20%) and Final exam (30%)**
  Students need to have a passing grade on the assignments and project (combined) as well as a passing grade on the tests (combined) to pass the class.
## 8. Schedule:

<table>
<thead>
<tr>
<th>Week</th>
<th>Tentative Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td>Introduction to Information Retrieval</td>
</tr>
<tr>
<td>Week 2</td>
<td>Indexing and Query Processing</td>
</tr>
<tr>
<td>Week 3</td>
<td>Retrieval Models (Boolean, Vector Space, Query-Likelihood)</td>
</tr>
<tr>
<td>Week 4</td>
<td>Concept of Relevance (Relevance Feedback)</td>
</tr>
<tr>
<td>Week 5</td>
<td>Evaluation (Metrics, Collection based evaluation, Log analyses)</td>
</tr>
<tr>
<td>Week 6</td>
<td>Human Computer Interaction in IR (Controlled Studies)</td>
</tr>
<tr>
<td>Week 7</td>
<td>Text Classification</td>
</tr>
<tr>
<td>Week 8</td>
<td>Document Clustering</td>
</tr>
<tr>
<td>Week 9</td>
<td>IR systems and WWW (PageRank, Search engines)</td>
</tr>
<tr>
<td>Week 10</td>
<td>Aggregated Search (Introduction, Source-Orientation, User behavior)</td>
</tr>
</tbody>
</table>