498: Information Retrieval and Web Search

Syllabus
Winter 2015

Schedule:
Mon/Wed 12-1:30pm

Instructor
Rada Mihalcea, BBB 3769, mihalcea@umich.edu
Office hours: Wednesday 2:00-3:00pm

GSI
Shibamouli Lahiri, BBB 1695, lahiri@umich.edu
Office hours: T 11:30-1:30pm, Th 11:30-1:30pm, F 12:30-2:30pm

Course description
This course will cover traditional material, as well as recent advances in Information Retrieval (IR), the study of indexing, processing, querying, and classifying data. Basic retrieval models, algorithms, and IR system implementations will be covered. While the course will primarily focus on IR techniques for textual data, it will also address IR for other media, including images/videos, music/audio files, and geospatial information. The course will also address topics in Web search, including Web crawling, link analysis, search engine development, social media, and crowdsourcing. There will be two invited lectures from people working at major companies in the field (e.g., Google, Yahoo, Microsoft, Facebook, Twitter).

Website
Most course materials are made available on the course website, and are considered required reading. A detailed schedule, including lecture topics, assignment due dates and exam dates, will also be available on the course website.

http://web.eecs.umich.edu/~mihalcea/498IR

Forum
We will be using Piazza to host a course forum. You are required to read this regularly; it is the venue we will use for important course announcements and assignment clarifications. In addition, it will be a significant source of help and hints on the assignments. We do not answer technical questions via email. In order to save everyone time, we want all students to have the benefit of seeing each question and its answer, so please use the forum. Please do not post your own solutions, project code, test cases, or output to the forum. Also, please search the forum before posting to avoid questions that have already been asked and answered.
Contact
Please direct technical questions to our Piazza forums. For other questions, you can reach us at mihalcea@umich.edu or lahiri@umich.edu

Prerequisites
EECS 281 Data Structures and Algorithms

List of Topics

<table>
<thead>
<tr>
<th>Course Overview. Introduction to IR models and methods.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text properties. Text processing.</td>
</tr>
<tr>
<td>Web crawling</td>
</tr>
<tr>
<td>Boolean model and extensions</td>
</tr>
<tr>
<td>Vector space model</td>
</tr>
<tr>
<td>Evaluation of IR systems. IR test collections.</td>
</tr>
<tr>
<td>Relevance feedback. Query expansion.</td>
</tr>
<tr>
<td>Text classification. Text clustering.</td>
</tr>
<tr>
<td>Link analysis. PageRank and HITS. Topic sensitive link analysis.</td>
</tr>
<tr>
<td>Question answering. Information Extraction.</td>
</tr>
<tr>
<td>Keyword extraction. Text summarization.</td>
</tr>
<tr>
<td>Image retrieval. Geospatial search.</td>
</tr>
<tr>
<td>Recent trends in information retrieval.</td>
</tr>
</tbody>
</table>

Required/recommended readings

(required) Introduction to Information Retrieval
(online version available) Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze Cambridge University Press, 2008.

(recommended) Readings in Information Retrieval
K.Sparck Jones and P. Willett
Morgan Kaufmann, 1997

(recommended) Modern Information Retrieval
Ricardo Baeza-Yates and Berthier Ribeiro-Neto
Addison-Wesley, 1999

Additional readings will be assigned throughout the semester.
Programming Environment

For this course, you must have a CAEN account. Anyone enrolled in EECS 498 is eligible for one, but it takes a little while to get it set up. You will have to write your programs in Python. We will grade your programs in the CAEN Linux environment and they must run correctly in this environment.

You are free to develop your programs on any platform you like, but they must run on the CAEN platform. To be safe, though, we recommend you develop in the CAEN environment, using the machine pool loginlinux.engin.umich.edu.

Course work and grading

Programming assignments: 30%
Exam I: 20%
Exam II: 20%
Project: 30%

There will be four programming assignments throughout the semester. Each assignment will specify the material to be turned in. Assignments are due by 11:59pm on the due date. Assignments may be turned in up to 3 days late, with a penalty of 10% for each day late. No credit will be given after 3 days. No exceptions.

Accommodations for Students with Disabilities

If you think you need an accommodation for a disability, please let your instructor know at the beginning of the semester or at least 2 weeks in advance. As soon as you make us aware of your needs, we can work with the Services for Students with Disabilities (SSD) office to help us determine appropriate academic accommodations. SSD (734-763-3000; http://ssd.umich.edu) typically recommends accommodations through a Verified Individualized Services and Accommodations (VISA) form. Any information you provide is private and confidential and will be treated as such.