

Information Retrieval and Web Agents

600.466 - Spring '08

Instructor: Prof. David Yarowsky **TA:** TBA
NEB 324
(410) 516-5372
yarowsky@cs.jhu.edu

Meeting Time: Tu,Th: 3:00-4:15 PM

Classroom: Shaffer 3

Web Page: <http://www.cs.jhu.edu/~yarowsky/cs466.html>

Office Hours: Instructor - Wed 3-4, Tuesday/Thursday after class and by appointment.
TAs - TBA, special review sections, and by appointment.

Primary Readings:

- C. Manning, P. Raghavan and H. Schuetze, *Introduction to Information Retrieval*, Cambridge University Press. June 2008.
<http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html>
- R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley Longman, 1999.
- L. Wall, T. Christiansen and R. Schwartz, *Programming Perl* O'Reilly & Associates. **or another Perl reference of your choice**
- C. Wong. *Web Client Programming*. O'Reilly & Associates, 1997.
- Selected papers distributed in class.

Recommended Readings:

- I. Witten, A. Moffat and T. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*, 2nd Edition. Morgan Kaufmann, 1999.
- W. Frakes and R. Baeza-Yates. *Information Retrieval: Data Structures and Algorithms*. Englewood Cliffs, N.J. : Prentice Hall, 1992.
- D.A. Grossman, O. Frieder. *Information Retrieval: Algorithms and Heuristics*. Springer, 2004.
- G. Salton and M. McGill, *An Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

Prerequisites:

Students should have a solid programming background and have taken 600.226 (Data Structures) or its equivalent. Knowledge of Perl (or a willingness to learn the language on your own relatively quickly) is also very important.

The material covered will be complementary to that in 600.465 (Natural Language Processing) and 600.666 (Information Extraction). Similarities and differences will be discussed in the first class. No background in NLP is assumed, and although 600.465 is helpful, it is not necessary as a prerequisite.

Course Requirements: Final grades will be based on the following (subject to change):

Assignments (4):	32%
Comprehensive Exam:	31%
Final Project:	31%
Class Participation:	6%

Assignments:

1. Preliminary Text Analysis and Corpus Processing
2. A Vector-model Information Retrieval System
3. (a) Email/News Routing and Filtering - Supervised IR
(b) Named Entity Classification and Information Extraction
4. Build (and unleash) a Web Agent

Considerable infrastructure will be provided in support of each assignment. These will include partial code, supporting routines and training data.

The first 3 assignments will be empirically evaluated on held-out (previously unseen) test sets. A portion of the grade will be based on this objective measure of performance. Code for self-evaluation on a secondary test set will also be provided so students may receive feedback during assignment development and debugging.

Final Project:

The final project for the course will be on a topic of your own choosing. Several options will be suggested. A brief (1-2 page) proposal will be required.

Lateness Policy:

One homework assignment may be handed in up to 5 days late without penalty, and without the need for permission or excuse. No other late homeworks will be accepted.

Preliminary Class Schedule

Tu 1/29	Course Overview. Discussion of problems and issues in Information Retrieval
Th 1/31	Introduction to IR models and methods (Boolean/vector/probabilistic)
Tu 2/05	The Perl Language
Th 2/07	Preliminary stages of text analysis and document processing.
Tu 2/12	Text analysis (cont). Boolean IR models.
Th 2/14	Inverted files, indexing, signature files, PAT trees, suffix arrays
Tu 2/19	Vector-based IR models
Th 2/21	Vector-based IR models (cont.) - including term weighting, similarity measures
Tu 2/26	Evaluation metrics, test collections and issues.
Th 2/28	Query expansion, thesaurus creation, clustering algorithms, SVD/LSI
Tu 3/04	Relevance Feedback and Probabilistic IR models
Th 3/06	(cont.) - including user modelling, automatic feedback acquisition
Tu 3/11	Document routing/filtering/topic-classification; Spam detection
Th 3/13	Information extraction and "text understanding"
Tu 3/18	NO CLASS - Spring Break
Th 3/20	NO CLASS - Spring Break
Tu 3/25	Information Extraction (cont.) - named entity recognition/tagging
Th 3/27	IE (cont.) - sense tagging, co-reference resolution, MUC template filling
Tu 4/01	IE (cont.) Text summarization, event detection and tracking
Th 4/03	Information visualization - Dotplot, Texttiling, graphical queries
Tu 4/08	IR on the World Wide Web - new technologies and protocols
Th 4/10	Web robots, spiders, crawlers, ants, HTTP, robot exclusion
Tu 4/15	WWW search engines - case studies and methods
Th 4/17	IR on the WWW cont. - Harvest, collection fusion, Metacrawler
Tu 4/22	Collaborative filtering. Web Agents.
Th 4/24	Web agents - webshopper, bargainfinder, case studies
Tu 4/29	Web agents - case studies, economic, ethical, legal and political issues
Th 5/01	Future directions, overview and conclusion

Note: Because the time devoted to individual topics depends on the length of class discussion and other factors, the schedule above is tentative and subject to change.

The date for the final examination will be discussed in class and resolved early in the semester.

Additional Sources for Readings (major conference proceedings):

Information Retrieval:

- SIGIR (ACM Conference on R&D in Information Retrieval)
- TREC (Text Retrieval Conference)
- SDAIR (Symposium on Document Analysis and Information Retrieval)
- TDT Topic Detection and Tracking Workshops

Natural Language Processing:

- ACL/NAACL (Association for Computational Linguistics)
- COLING (International Conference on Computational Linguistics)
- EMNLP (Empirical Methods in Natural Language Processing)
- MUC (Message Understanding Conference)
- HLT (Human Language Technology Conferences)

Computer Science Department Academic Integrity Code

The strength of the university depends on academic and personal integrity. In your studies, you must be honest and truthful. Ethical violations include cheating on exams, plagiarism, reuse of assignments, improper use of the Internet and electronic devices, unauthorized collaboration, alteration of graded assignments, forgery and falsification, lying, facilitating academic dishonesty, and unfair competition.

Academic honesty is required in all work you submit to be graded. Except where the instructor specifies group work, you must solve all homework and programming assignments without the help of others. For example, you must not look at any other solutions (including program code) to your homework problems or similar problems. However, you may discuss assignment specifications with others to be sure you understand what is required by the assignment.

If your instructor permits using fragments of source code from outside sources, such as your textbook or on-line resources, you must properly cite the source. Not citing it constitutes plagiarism. Similarly, your group projects must list everyone who participated.

*Students in 600.466 are allowed free use of all partial example solutions and other source code made available in the course directories or in class (without the need for citation). Students may also use small code fragments for general problems found in reference books (with clear appropriate citation if exceeding 3-4 lines), but students in 600.466 are **not** allowed to use or examine any other solutions to problems that are the same or reasonably similar to those covered on the 4 course homeworks or chosen course project.*

Falsifying program output or results is prohibited.

Your instructor is free to override parts of this policy for particular assignments. To protect yourself: (1) Ask the instructor if you are not sure what is permissible. (2) Seek help from the instructor or TA, as you are always encouraged to do, rather than from other students. (3) Cite any questionable sources of help you may have received.

Students who cheat will suffer a serious course grade penalty in addition to being reported to university officials. You must abide by JHU's Ethics Code: Report any violations you witness to the instructor. You may consult the associate dean of students and/or the chairman of the Ethics Board beforehand. For more information, see the guide on Academic Ethics for Undergraduates (<http://www.advising.jhu.edu/ethics.html>) and the Ethics Board web site (<http://ethics.jhu.edu>).