Information Retrieval and Web Agents
601.466/666 - Spring 2024

Instructor: Prof. David Yarowsky
Hackerman 324G
410-516-5372
`yarowsky@jhu.edu`

TA: Niyati Bafna
`nbafna1@jh.edu`

CA: Shreayan Chaudhary
`schaud31@jhu.edu`

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

Classroom: Hackerman B-17

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

Office Hours: Instructor - Thursday 2-3, 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, 2008. [**Primary Text**]
http://nlp.stanford.edu/IR-book/information-retrieval-book.html
- C. Wong. *Web Client Programming.* O'Reilly & Associates, 1997. (available online at `http://oreilly.com/openbook/webclient/`)
- Selected papers distributed in class.

Supplemental Readings:

- A Python language reference book of your choice (suggestions given on class website)
- R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison Wesley Longman, 1999.
- 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 601.226 (Data Structures) or its equivalent. Knowledge of Python (or a willingness to learn the language on your own relatively quickly) is also important.

The material covered will be complementary to that in 601.465/665 (Natural Language Processing) and 520.666 (Information Extraction). Similarities and differences will be discussed in the first class. No background in NLP is assumed, and although 601.465/665 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:	30%
Final Project:	32%
Class Participation:	6%

Assignments:

1. Machine Learning for preliminary Text Analysis and Corpus Processing
2. A Vector-model Information Retrieval System
3. Vector-based and Bayesian Text Classification and Information Extraction
   (a) Named Entity Classification and Word Sense Disambiguation
   (b) Email/News Routing and Filtering - Supervised IR
   (c) Related Text Classification Problems (Gender detection, Authorship ID, Language ID, Sentiment Classification)
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.

Lateness Policy:

Although students have 2-3 weeks to complete most assignments, recognizing that last-minute illness or unplanned events may occur, homework assignments may optionally be handed in late *up to a total of 5 days combined across all homeworks* without penalty and without the need for permission or excuse. Each 24 hour period after the due date and time counts as 1 late day, and are counted in granularities of whole days (no partial days).

Preliminary Class Schedule

Tu 1/23	Course Overview. Discussion of problems and issues in Information Retrieval
Th 1/25	Introduction to IR models and methods (Boolean/vector/probabilistic)
Tu 1/30	Preliminary stages of text analysis and document processing. Boolean IR models.
Th 2/01	Inverted files, indexing, signature files, PAT trees, suffix arrays
Tu 2/06	Inverted files, indexing, signature files, PAT trees, suffix arrays (cont.)
Th 2/08	Vector-based IR models
Tu 2/13	Vector-based IR models (cont.) - including term weighting, similarity measures
Th 2/15	Query expansion, thesaurus creation, clustering algorithms, SVD/LSI
Tu 2/20	Evaluation metrics, test collections and issues.
Th 2/22	Relevance Feedback and Probabilistic IR models
Tu 2/27	(cont.) - including user modeling, automatic feedback acquisition
Th 2/29	Document routing/filtering/topic-classification; Spam detection
Tu 3/05	(Large) Language-model-based IR and Industry Standard IR tools
Th 3/07	Information extraction, Text Classification and Question Answering
Tu 3/12	IE (cont.) - named entity recognition/tagging, semantic frame analysis
Th 3/14	IE (cont.) - sense tagging and general semantic disambiguation
Tu 3/19	Spring Break (no class)
Th 3/21	Spring Break (no class)
Tu 3/26	IE (cont.) - Sentiment classification, authorship ID, language ID, gender detection
Th 3/28	Information visualization - Dotplot, Texttiling, graphical queries
Tu 4/02	Web robots, spiders, crawlers, ants, HTTP, robot exclusion
Th 4/04	IR on the WWW cont. - Harvest, collection fusion, Metacrawler
Tu 4/09	IR on the World Wide Web - new technologies and protocols
Th 4/11	Music Information retrieval and Image Search
Tu 4/16	Collaborative filtering. Web Agents.
Th 4/18	Web agents - webshopper, bargainfinder, case studies
Tu 4/23	Web agents - case studies, economic, ethical, legal and political issues
Th 4/25	Future directions, overview and conclusion
We 5/08	Final Examination (9AM-12PM, official T/Th 3pm exam slot)

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 final examination date and time will be held in the official slot for our course meeting time. Anyone unable to take an exam during that slot for religious reasons or other reasons known in advance should inform the instructor prior to April 25 so alternate arrangements can be discussed.

Additional Sources for Readings (major conference proceedings):

Information Retrieval:

SIGIR	(ACM Conference on R&D in Information Retrieval)
TREC	(Text Retrieval Conference)
WSDM	(Web Search and Data Mining)

Natural Language Processing / Information Extraction:

ACL/NAACL	(Association for Computational Linguistics)
EMNLP	(Empirical Methods in Natural Language Processing)
COLING	(International Conference on Computational Linguistics)
DUC/TAC	(Document Understanding Conference/Text Analysis Conference)

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 601.466/666 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 601.466/666 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).