


**Course Syllabus**

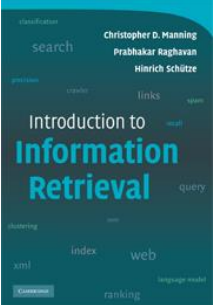
**ISTE-612**

**Knowledge Processing Technologies**

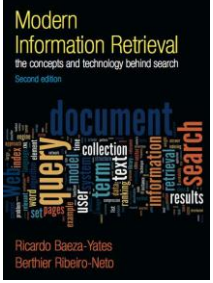
**REMINDER:** The information presented in this syllabus is subject to expansion, change, or modification during the semester.

<p><b>Instructor:</b>          Weishi Shi   email address: ws7586@rit.edu (<b>Please include “612” in the subject line of all email messages that you send.</b>)</p>	<p><b>Office Hours:TBD</b></p> <p><b>Others by appt.</b></p>
---	--

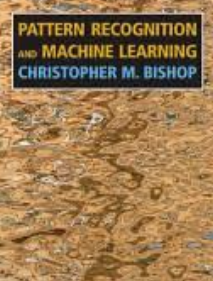
**Course Text and Materials**

	<p>Introduction to Information Retrieval          Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze</p> <p>Cambridge University Press, 2008.          ISBN-10: 0521865719          ISBN-13: 978-0521865715</p> <p>Available Online: <a href="http://nlp.stanford.edu/IR-book/">http://nlp.stanford.edu/IR-book/</a></p>
--	--

**Recommended Reading (Optional):**

	<p>Modern Information Retrieval:          The Concepts and Technology behind Search (2nd Edition)          Ricardo Baeza-Yates and Berthier Ribeiro-Neto</p> <p>ACM Press Books, 2011.          ISBN-10: 0321416910          ISBN-13: 978-0321416919</p>
---	--

**Recommended Reading (Optional):**

	<p>Pattern Recognition and Machine Learning (Information Science and Statistics)          Christopher M. Bishop</p> <p>Springer (October 1, 2007)          ISBN-10: 0387310738          ISBN-13: 978-0387310732</p>
---	---

**Important RIT Deadlines**

**Last day of drop/add: February 2, 2015**

**Last day to withdraw from the course with a “W”: April 24, 2015.**

**NOTE:** IST department policy states that a student has one semester to **challenge** any **grade**. After that, grades cannot be challenged.

### **Course Description**

This is the second course in a 2-course sequence that provides students with exposure to foundational information sciences and technologies. Topics include information processing, modeling, storing, searching, and retrieval technologies, data and text analytics for knowledge extraction, and Internet middleware technologies to access and deliver data and knowledge. Prerequisites: (One year of programming in an object-oriented language, a database theory course, a course in Web development, and a statistics course). Offered: Fall and Spring.

### **Course Goals and Objectives**

The main objectives of this class is to

1. Process, model, store, index, represent, search, and retrieve, modern data collections using fully automatic systems—focusing on information retrieval systems.
2. Analyze data by developing algorithms to extract high-level knowledge from it—focusing on data analytics, especially text analytics, including text classification and clustering.
3. Exploit Internet middleware technologies to access external data sources or deliver information to others—focusing on web service technologies

The major focus will be placed on unstructured data, which may be in the form of text, hypertext, or multimedia (e.g., images and videos). Various unstructured data models will be studied, such as Boolean, probabilistic, and vector space models, which are implemented using inverted files, relational thesauri, special hardware, and other approaches. State-of-the-art technologies and research skills of literature analysis, innovation, evaluation of new ideas, and communication are emphasized via lab assignments and projects. Students will get exposed to a broad overview of the research topics, methodologies, major results, open problems, and potential future research directions. More specifically, this course will help students:

- To identify basic theories and analysis tools as they apply to process modern data collections
- To develop understanding of problems and potentials of current information processing and retrieval systems
- To learn and appreciate different information processing and retrieval algorithms and systems
- To apply various modeling, indexing, matching, organizing, and evaluating methods to data processing, retrieval, and analysis problems

- To become aware of current experimental and theoretical research in the area

**Role of course in curriculum for:**

- Use multiple open-source libraries
- Program effectively within the student's specialty area.
- Develop specialized IT skills in a self-selected specialty area.

**Course Materials**

**Written Exams**

Anticipated test dates are shown on the attached schedule. Requests to take tests at a different time will not be honored except in exceptional circumstances, such as a documented medical excuse, a serious family emergency, or scheduled RIT-approved off-campus event, and must be arranged with the instructor *in advance* if the circumstance can be anticipated.

The second exam will be given during the exam week of the semester. Again, requests to take this exam at a different time will not be honored except in the exceptional circumstances discussed above.

Students are responsible for all material covered in lectures. Examinations will heavily emphasize conceptual understanding of the material. All exams will be analyzed after they have been administered; the instructor will look for -- and eliminate -- invalid or poorly framed questions.

All examinations are closed book. You may prepare a *hand-written* two-sided 8 1/2 by 11 inch "crib sheet" (even if it is empty) and bring it to your examinations. This crib sheet must have your name in the upper left hand corner (even if it is otherwise empty).

**Labs**

Each student will be given a "lab check-off" sheet during the first lab period. As you complete a lab, you demonstrate it to your lab instructor and answer questions about the lab. If the lab is correct and the answers are satisfactory, the lab instructor will initial your sheet. At the end of the course, whatever portion of the lab percentage you have earned will be awarded to you.

This must be done as you go along! The idea behind this is educational: it is to give you incentive to keep up throughout the course. Nothing will be awarded to someone who -- at the last minute -- digs up all of his or her labs and asks for them to be evaluated; this defeats the purpose of the award. **In keeping with this, Labs 1 and 2 must be completed and initialed before the midterm exam.**

Obviously, do not lose the sheet: we will not evaluate them until the end of the semester. If it is lost, so is your chance to achieve *any* credit for your labs.

### **Project**

For the project, you will work in teams of **either two or three students** on a problem of your choosing that is interesting, significant, and relevant to building an information storage, processing, and retrieval system and/or developing data analytics algorithms for knowledge extraction from non-trivial data that your team collects. You will have great latitude in what you choose to work on, so take advantage of this opportunity to make a big impact! Please see the project description document for details about the project.

### **Class Attendance**

Your actions in the classroom and the lab should reflect the standards of behavior set in the commercial environment: you should be respectful of your classmates, the professor, the teaching assistant (TA) and the course support personnel (the note takers and interpreters) and you should willingly participate when asked to do so.

### **myCourses**

MyCourses will serve as the primary communication mechanism for this course. You should log in on mCourses and check for changes in labs or schedule on a regular basis. All course materials will be posted to the content section of myCourses.

## Topics

- **Introduction to Knowledge and Data Processing:** Motivation, Definition, and Conceptual Model
- **Modeling of Unstructured Data:** Boolean, Vector Space, and Probabilistic Models
- **Evaluation:** Assumptions, Measures of retrieval performance, Test Collections, Evaluation Methodology
- **Search Strategies:** Query Languages, Query Reformulation: query expansion, term reweighting, Relevance Feedback, Information Visualization
- **Data and File Structures:** Inverted files
- **Automatic Indexing:** Lexical Analysis: stoplists, stemming, Segmentation strategies for long text, Thesaurus Construction <sup>[L]</sup><sub>SEP</sub> manually derived (WordNet), automatic (Latent Semantic Indexing)
- **Knowledge extraction:** Text Classification, Clustering
- **Internet Middleware Technologies:** Web services, SOAP, WSDL, and REST

## Grading

The grading scale used along with the grading criteria is as follows:

Component	Weight
Lab	20
Midterm exam	25
Final exam	30
Project	25

Range	Grade
$\geq 90.0\%$	A
$\geq 80.0\% \ \& \ < 90\%$	B
$\geq 70.0\% \ \& \ < 80.0\%$	C
$\geq 60.0\% \ \& \ < 70.0\%$	D
$< 60.0\%$	F

Project	Weight
<b>Checkpoint 1:</b>	10%
• Project proposal	
<b>Checkpoint 2:</b>	10%
• Data collection, description, and processing	
<b>Checkpoint 3:</b>	10%
• Core algorithm	
<b>Checkpoint 4:</b>	10%
• System integration and demo feedback	
Final demo and project presentation	40%
Peer evaluation	20%
Total	25

### Course Schedule

The estimated course schedule is below. All dates, lecture topics, and assignments are subject to reasonable change at the discretion of your instructor. Any changes will be announced. In the event that you feel a change will be detrimental to you, please make your concerns known when the change is announced.

Week	Lectures	Reading	Due	Lab
1	Course Introduction Boolean Model	Chap 1		
2	Boolean Model – Inverted Index	Chap 1	Team Selection	Lab#1
3	Text Processing Technologies	Chap 2	Project Checkpoint 1	
4	Data Representation I— Postings Lists and Variations	Chap 2		
5	Data Representation II— Term Dictionary	Chap 3		Lab#2
6	Binary Tree Examples		Project Checkpoint 2	
7	Vector Space Model	Chap 6		Lab#3
8	Model and Retrieval Evaluation	Chap 8		
9	Review & Midterm			
10	Text Classification I	Chap 14	Project Checkpoint 3	Lab#4
11	Text Classification II			
12	Text Clustering	Chap 16		Lab#5
13	Internet Middleware Technologies		Project Checkpoint 4	
14	Project Work & Demo		Project Report	
15	Project Demo			

**Cheating Policy:** Please review the departmental policy on cheating as described at <http://www.it.rit.edu/dishonesty.php>. (Also attached to the end of this syllabus) I will strictly follow the department policy on dealing with academic dishonesty.

Note that if you get accused of cheating, the evidence has already been checked by other faculty members to verify it will withstand an appeal.

***Additional Policies:***

**Cell Phones/Pagers:**

Shut them off, I don't want to hear them going off during class.

**Late Policy:**

If you are having problems with an assignment or an emergency that may make you late in submitting your work, **contact me before the due date**. Excuses made after the fact will not be honored.

**Extra Credit:**

My policy on extra credit is simple: I do not offer extra credit assignments for any reason. Please do not ask for one.

**Notices of Accommodation:**

If you have a "Notice of Accommodation", you must provide me with a copy within a week of starting this course. If you provide me with the notice later in the course, it will not be retroactive. (In other words, an NOA is not a license to retake an exam or practical that you have done poorly on.)

**Final Exam Date:**

The final exam date for this course is set of the Registrar's Office about the middle of the quarter. Please do not make travel plans without checking the exam schedule, since I will not give an early exam to accommodate your plans.

**Contact Information:**

Any updates to assignments and any emails that I need to send to individual students will be done through MyCourses. **What this means is that you should check your email and the MyCourses conference for this course periodically.**

I generally have email running whenever I am logged in, so you should get a reply to any email you send me within a day.

**Finally...**

Any or all of the previous information is subject to change or modification during the semester.

**ACADEMIC DISHONESTY POLICY  
DEPARTMENT OF INFORMATION TECHNOLOGY**

The following statement is the Policy on Academic Dishonesty for the Department of Information Technology:

The Department of Information Technology does not condone any form of academic dishonesty. Any act of improperly representing another person's work as one's own (or allowing someone else to represent your work as their own) is construed as an act of academic dishonesty. These acts include, but are not limited to, plagiarism in any form or use of information and materials not authorized by the instructor during an examination or for any assignment.

If a faculty member judges a student to be guilty of any form of academic dishonesty, the student will receive a **FAILING GRADE FOR THE COURSE**. Academic dishonesty involving the abuse of RIT computing facilities may result in the pursuit of more severe action.

If the student believes the action by the instructor to be incorrect or the penalty too severe, the faculty member will arrange to meet jointly with the student and with the faculty member's immediate supervisor. If the matter cannot be resolved at this level, an appeal may be made to the GCCIS Academic Conduct Committee.

If the faculty member or the faculty member's immediate supervisor feels that the alleged misconduct warrants more severe action than failure in the course, the case may be referred to the GCCIS Academic Conduct Committee. The Academic Conduct Committee can recommend further action to the dean of the student's college including academic suspension or dismissal from the Institute.

The following definitions will be used to clarify and explain unacceptable conduct. This is not intended to be an exhaustive list of specific actions but a reasonable description to guide one's actions.

**CHEATING** includes knowingly using, buying, stealing, transporting or soliciting in whole or part the contents of an administered/unadministered test, test key, homework solution, paper, project, software project or computer program, or any other assignment. It also includes using, accessing, altering, or gaining entry to information held in a computer account or disk owned by another.

**COLLUSION** means the unauthorized collaboration with another person in preparing written work or computer work (including electronic media) offered for credit. Final work submitted by a student must be substantially the work of that student. Collaboration on an assignment is expressly forbidden unless it is explicitly designated as a group project. When there is any doubt, a student should consult the



instructor (NOT ANOTHER STUDENT) as to whether some action is considered collusion.

Whenever there is any question as to whether a particular action is considered academic dishonesty, the instructor should be consulted.

The penalty for academic dishonesty in a course is an automatic "F" in that course.