School of Computing and Information - University of Pittsburgh

INFSCI 2140: Information Storage and Retrieval [Current as of: 08/26/2021]

Fall 2021

Class time: **Mondays 12:00pm – 2:50pm**
Location: **IS 405**

Instructor:

Daqing He, PhD, Professor
School of Computing and Information, University of Pittsburgh
Phone: 412-624-2477
E-mail: dah44@pitt.edu
Office: Room 618, Information Science Building
Virtual Office Hours: Thursdays 9pm to 10pm@Zoom

Graduate Student Assistants:
TA: Saeed Javadi
Office: 707 Information Science Building
Email: saz31@pitt.edu
Virtual Office hours: Thursdays 3-5pm

Communication

- Sending email is the best means to talk to the instructor and the TA
  - Expect a 24-hour response time frame
- Attending virtual office hours is also a good way
  - Instructor's office hours: issues related to the course in general
  - TA's office hours: specific assignment related issues
- Submitting Muddiest Points
  - One stone for two-bird approach
  - A quick way to ask help on topics that confuse you
  - An easy way to earn participation points (up to 5% in final score)

Canvas URL: http://canvas.pitt.edu

I. Course Description:

This course offers an examination of problems and techniques related to storing and accessing unstructured information with an emphasis on textual information, an overview of several approaches to information access with a primary focus on search-based information access, an introduction to automated retrieval system design, content analysis, retrieval models, result presentation, and system

evaluation, and applications of retrieval techniques to various issues on the Web, on mobile platforms and other reality settings.

Prerequisites: introduction to logic and statistical analysis, familiarity with JAVA or Python programming language

Course Goals
Upon finishing this course, the students should be able to

- understand the dimensions of the information retrieval "problem";
- master the analysis and design of information retrieval systems;
- consider the factors which optimize the information retrieval process;
- examine current issues in information retrieval

Upon satisfactory completion of this course, students will:

- be able to explain core concepts and terms of information retrieval
- be able to explain different retrieval models and basic algorithms
- be able to evaluate existing information retrieval systems and suggest how the systems can be improved
- be able to apply theories to effectively solve information retrieval problems in real world situations

II. Canvas Information:

The Web-based teaching system for this course is Canvas, whose goal is to facilitate course-related communication as well as distribution of course materials and grades. You can access Canvas at http://canvas.pitt.edu. You must log in with your University Computer Account – this is the one that goes with your 'pitt.edu' e-mail address. Course-related e-mail will be sent to your Pitt e-mail account. If you do not read e-mail on your Pitt account, you are responsible for forwarding any e-mail received on your Pitt account to the e-mail address that you use. See http://accounts.pitt.edu/ for information on managing your Pitt account and forwarding e-mail. If you have trouble logging in to Canvas, you may need to log in to the accounts website above to activate your Pitt e-mail account. Call 412-624-HELP with any problems relating to your account.

III. Recommended books and Readings

There is no required textbook for this class. However, various parts of the following books will be used in the class:

1. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze, "Introduction to Information Retrieval". Cambridge University Press. 2008. Available at http://nlp.stanford.edu/IR-book/. Referred as "IIR" subsequently.
2. Mitra, Bhaskar, and Nick Craswell. *An introduction to neural information retrieval*. Now Foundations and Trends, 2018. https://www.microsoft.com/en-us/research/uploads/prod/2017/06/fntir2018-neuralir-mitra.pdf

3. Stefan Büttcher, Charles L. A. Clarke, Gordon V. Cormack, "Information Retrieval: Implementing and Evaluating Search Engines." MIT Press. 2010. Sample chapters are available at http://www.ir.uwaterloo.ca/book/. Referred as "IES" subsequently.
4. Ricardo Baeza-Yates, Berthier Riberiro-Neto, "Modern Information Retrieval", 2nd Edition. Addison Wesley, 2011. ISBN-10: 9780321416919. http://www.mir2ed.org/. Referred as "MIR" subsequently.

There will be about 3-4 required readings each week. These readings usually are available online whose URLs are shared in the table in Section IV. You are asked to read all these readings before the class each week starts. This reading task should be completed before 11:59pm of the Saturday before the class. As described below, 10 participations are required as part of your final grade, each of which counts for .5 participation point.

Readings will generally be available online or via Canvas (if available in electronic format). Additional readings may be added as needed.

IV. Course Schedule Summary

Date	Unit and Readings	Assignment and Others
Aug. 30	1: introduction and course overview Readings 1. Mitra, B, Craswell, N. *An Introduction to Neural Information Retrieval*, Foundations and Trends in Information Retrieval. 2018 Section 2.1 https://www.microsoft.com/en-us/research/uploads/prod/2017/06/INR-061-Mitra-neuralir-intro.pdf 2. IES section 1.1 and 1.2 (avaiable at http://www.ir.uwaterloo.ca/book/01-introduction.pdf) 3. MIR sections 1.1-1.4 (available at http://www.mir2ed.org/, the content section at the left side. Chapter 1)	Assignment 1 Out
Sep. 6	Labor Day (University Closed)	
Sep. 13	2: document and query processing Using BERT for query intent? https://blog.google/products/search/search-language-understanding-bert/ Readings *1*. IIR sections 1.2, chapters 2 and 3.	*Team Project Introduction*

Sep. 20	3: index construction and compression Readings: 1. IIR chapters 4 and 5.	Assignment 2 Out
Sep. 27	4: matching models 1: Boolean and vector space 1. IIR sections 1.3 and 1.4, chapter 6.	*Team Formation Deadline* Assignment 1 Due
Oct. 4	5: matching models 2: statistical language model Readings: 1. IIR chapter 12. 2. Djoerd Hiemstra and Arjen de Vries. (2000) Relating the New Language Models of Information Retrieval to the Traditional Retrieval Models. Technical Report, TR-CTIT-00-09, Centre for Telematics and Information Technology. http://citeseer.ist.psu.edu/299514.html	
Oct. 11	6: evaluation Readings: 1. IIR chapter 8. 2. Karen Sparck Jones, (2006). What's the value of TREC: is there a gap to jump or a chasm to bridge? ACM SIGIR Forum, Volume 40 Issue 1 June 2006 http://doi.acm.org/10.1145/1147197.1147198 3. Kalervo Järvelin, Jaana Kekäläinen. (2002) Cumulated gain-based evaluation of IR techniques ACM Transactions on Information Systems (TOIS) Volume 20 , Issue 4 (October 2002) Pages: 422 – 446 http://doi.acm.org/10.1145/582415.582418	Assignment 3 Out
Oct. 18	7: relevance feedback and query expansion Readings: 1. IIR chapter 9. 2. Xu, J. and Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.* 18, 1 (Jan. 2000), 79-112. (DOI= http://doi.acm.org/10.1145/333135.333138) 3. Wang, X., Fang, H., and Zhai, C. (2008). A study of methods for negative relevance feedback. In *Proceedings of the 31st Annual*	Assignment 4 Out Assignment 2 Due

	international ACM SIGIR Conference on Research and Development in information Retrieval (Singapore, Singapore, July 20 - 24, 2008). SIGIR '08. ACM, New York, NY, 219-226. (DOI= http://doi.acm.org/10.1145/1390334.1390374) 4. Donna Harman, (1992). Relevance feedback revisited. Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval. Pages: 1 - 10. Copenhagen, Denmark. 1992. (http://doi.acm.org/10.1145/133160.133167)	
Oct. 25	8: matching models 3: learning to rank and latent semantics models 1. Liu, T. Y. (2009). Learning to Rank for Information Retrieval. Foundations and Trends® in Information Retrieval, 3(3), 225-331. http://didawikinf.di.unipi.it/lib/exe/fetch.php/magistraleinformatica/ir/ir13/1_-_learning_to_rank.pdf (sections 1.2, 2.1. 2.2, 3, 4.1.3) 2. IIR chapters 18 3. Hofmann, T., *Probabilistic latent semantic indexing*, in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 1999, ACM: Berkeley, California, USA. p. 50-57.[17] http://dl.acm.org/citation.cfm?id=312649	
Nov. 1	9: matching models 4: word embedding and neural models 1. Mitra, B, Craswell, N. *An Introduction to Neural Information Retrieval*, Foundations and Trends in Information Retrieval. 2018 https://www.microsoft.com/en-us/research/uploads/prod/2017/06/INR-061-Mitra-neuralir-intro.pdf 2. Guo, Jiafeng, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W. Bruce Croft, and Xueqi Cheng. "A deep look into neural ranking models for information retrieval." *Information Processing & Management* 57, no. 6 (2020): 102067. 3. Lin, Jimmy, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. "Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations." In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*. 2021. 4. Lin, Jimmy, Rodrigo Nogueira, and Andrew Yates. "Pretrained transformers for text ranking: Bert and beyond." *arXiv preprint arXiv:2010.06467* (2020).	*Assignment 3 Due*
Nov. 8	10: user interaction and interactive information retrieval Readings: 1. MIR chapter 10 (available at http://people.ischool.berkeley.edu/~hearst/irbook/) 2. Marti A. Hearst. Ch. 1: The Design Of Search User Interfaces. Search User Interfaces.	*Project Initial Presentation (Online)*

	(http://searchuserinterfaces.com/book/sui_ch1_design.html) 3. Marti A. Hearst. Ch. 11: Information Visualization For Text Analysis. Search User Interfaces. http://searchuserinterfaces.com/book/sui_ch11_text_analysis_visualization.html	
Nov. 15	11. Exam	
Nov. 22	Thanksgiving Break	
Nov. 29	12. Web information retrieval Readings: 1. IIR chapters 19 and 21. *OR MIR chapter 13* 2. J. Kleinberg. (1998) Authoritative sources in a hyperlinked environment. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998. www.cs.cornell.edu/home/kleinber/auth.pdf 3. S. Brin, L. Page: (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. WWW7 / Computer Networks 30(1-7): 107-117 (1998) http://dbpubs.stanford.edu:8090/pub/1998-8	Assignment 4 Due
Dec. 6	13. intelligent information retrieval Readings: 1. Susan Gauch, Mirco Speretta, Aravind Chandramouli and Alessandro Micarelli. User Profiles for Personalized Information Access. Chapter 2 in Brusilovsky, P., Kobsa, A., Neidl, W. (eds.) (2007) The Adaptive Web: Methods and Strategies of Web Personalization. Lecture Notes in Computer Science, Vol. 4321. Springer-Verlag, Berlin Heidelberg New York. 2. Shen, Xuehua, Bin Tan, and ChengXiang Zhai. "Context-sensitive information retrieval using implicit feedback." *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2005. http://dl.acm.org/citation.cfm?id=1076045 3. Wang, Xuanhui, and ChengXiang Zhai. "Learn from web search logs to organize search results." In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 87-94. ACM, 2007. http://dl.acm.org/citation.cfm?id=1277759 4. Ahn, J., Brusilovsky, P., He, D., Grady, J., and Li, Q. (2008). Personalized web exploration with task models. In Proceeding of the 17th international Conference on World Wide Web (Beijing, China, April 21 - 25, 2008). WWW '08. ACM, New York, NY, 1-10. DOI= http://doi.acm.org/10.1145/1367497.1367499.	
Dec. 13	14: *Team Project Final Presentation*	*Term Project*

		Poster (final) Due

V. Assessment

Participation 10%

Class attendance is required for success in this course, as material will be covered in class that is not included in the readings. Participation is based on active participation to each week's "readings" before the class and "my muddiest points" after the class. The "readings" is planned to complete on an online reading system, which will be stated in detail in the class. Your muddiest points should be posted into the discussion section dedicated to the muddiest points in each unit. The deadline for posting muddiest point is marked in Canvas, generally should be the Saturday 11:59pm after the week class. Just list any questions regarding the issues covered during the class. Again, 10 responses to the muddiest points are required as part of your final grade, each of which counts .5 participation point.

If you must miss a class, please notify the instructor, and make arrangement to obtain course notes and handouts. Makeup exams will not be offered except under extreme circumstances.

Assignment 36%

There are total four assignments, each of which will count 9% in the final course score. The deadline of submitting each assignment is before 11:59pm of the due date. Each 24-hours delay will have 40% deduction of the maximal score. No submission later than 2 days will be accepted except in the case of emergencies and personal disasters.

Exam 24%

Due to the hybrid format of the course, the exam will be conducted online. The exact arrangement of the exam will be provided near the exam time. Common exam questions include short calculation, short discussions, and long discussion questions.

Term Project 30%

Please see section VI for detail description of term project.

Course Grading Scale:

The final grade depends on the percentage of points you have earned, and the definition of letter grades is:

- 90 <= A- < 93, 93 <= A < 98, 98 <= A+ <= 100
- 80 <= B- < 83, 83 <= B < 88, 88 < = B+ < 90
- 70 <= C- < 73, 73 <= C < 78, 78 <= C+ < 80
- 60 <= D < 70,

- F < 60

VI. Term Projects

Introduction:
The term project is designed for students to integrate and extend knowledge acquired throughout the course and to apply that knowledge to solve a problem of substantial scope. Students are required to work in groups of 3 people. Experience suggests that successful teams require expertise in design, implementation, and project management.

Your task is to design and develop a prototype retrieval system, using online APIs, Open Source software (e.g., Lucene, Lemur/Indri, etc) or Amazon Web Services. Each team propose their project, and the instructor sometimes may provide project ideas too.

If a collection is needed to compose for the project, to realistically demonstrate the usefulness of the retrieval systems, the collection should contain at least 500-1000 documents.

Milestones for the project:

Introduction of term project:	Unit 2
Team formation deadline:	Unit 4
Project Initial Presentation	Unit 10
Final project presentation:	Unit 14
Project Demo video:	Last week

VII. Course Policies

Ground rules for class discussion

On-class interaction and discussion will be an important means of learning in this course, therefore, it is important that we work together to create a constructive environment by observing these rules:

- You should participate in the discussion of ideas.
- You should respect diverse points of view.
- You should aware the diverse backgrounds of peers.
- You may not belittle or personally criticize another individual for holding a point of view different than your own
- Your use of language should be respectful of other individuals or groups

Plagiarism

It is expected that the work you submit in this course will be your own. While collaboration is allowed for the course project, it should be approved in advance and the nature of each contribution should be specified in the project proposal and the final submission.

The following statement is taken from *The Teaching Assistant Experience: A Handbook for Teaching Assistants and Teaching Fellows at the University of Pittsburgh* (A.P. Haley and J.M. Nicoll, eds.)]

> Plagiarism means submitting work as your own that is someone else's. For example, copying material from a book or other source without acknowledging that the works or ideas are someone else's and not your own is plagiarism. If you copy an author's words exactly, treat the passage as a direct quotation and supply the appropriate citation. If you use someone else's ideas, even if you paraphrase the wording, appropriate credit should be given. You have committed plagiarism if you purchase a term paper or submit a paper as your own that you did not write[1].
>
> Plagiarism is a violation of the University of Pittsburgh's standards on academic honesty, and violations of this policy are taken seriously. **From the *Guidelines on Academic Integrity: Student and Faculty Obligations and Hearing Procedures* (effective September, 1995):**

A student has an obligation to exhibit honesty, and to respect the ethical standards of the historical profession in carrying out his or her academic assignments. Without limiting the application of this principle, a student may be found to have violated this obligation if he or she:

- Presents as one's own, for academic evaluation, the ideas, representations, or words of another person or persons without customary and proper acknowledgment of sources.
- Submits the work of another person in a manner which represents the work to be one's own. [Quotation ellipsed.] [2]

Special Needs

Students with disabilities who require special accommodations or other classroom modifications should notify the instructor and the University's Office of Disability Resources & Services (DRS) no later than the 2nd week of the term. Students may be asked to provide documentation of their disability to determine the appropriateness of the request. DRS is located in 216 William Pitt Union and can be contacted at 648-7890 (Voice), 624-3346(Fax), and 383-7355(TTY). Students who must miss an exam or class due to religious observances must notify the instructor ahead of time and make alternative arrangements.

---

[1] B. G. Davis, *Tools for Teaching* (San Francisco: Jossey-Bass, 1993), 300.

[2] University of Pittsburgh, *Guidelines on Academic Integrity: Student and Faculty Obligations and Hearing Procedures* (Pittsburgh: University of Pittsburgh, 1995), 7-8.