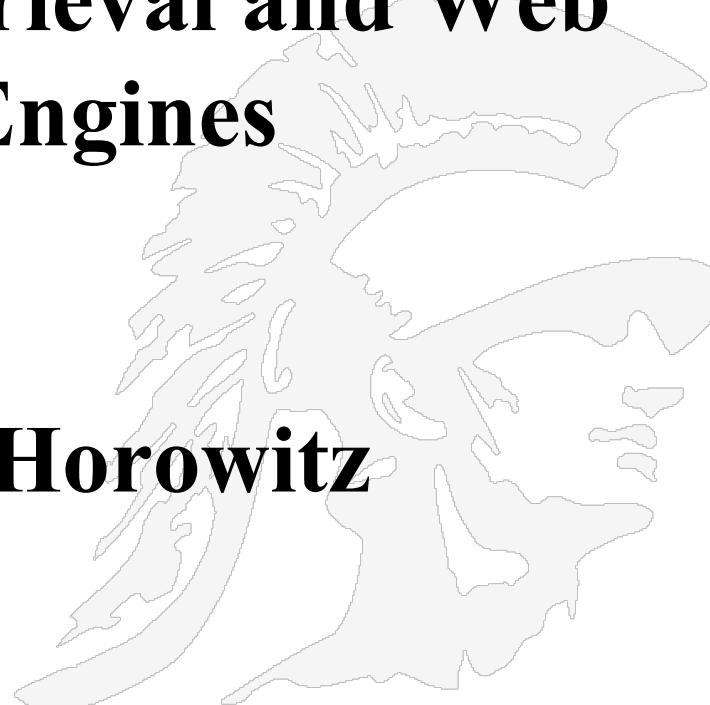


CSCI 572

Information Retrieval and Web

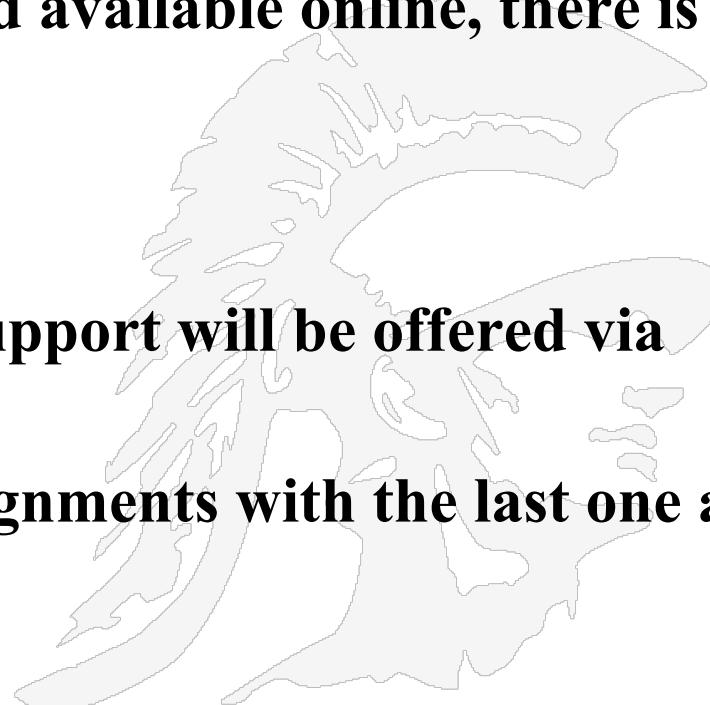
Search Engines

Prof. Ellis Horowitz



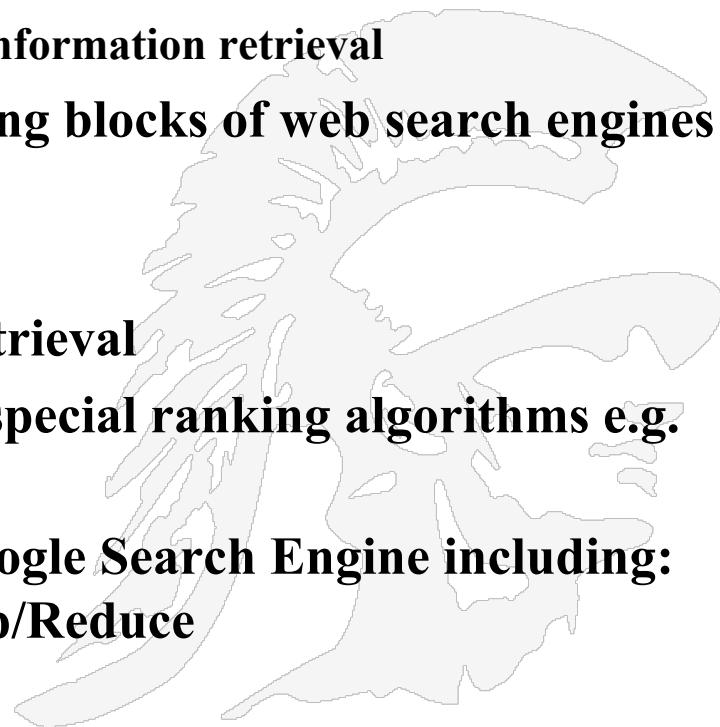
General Requirements

- All lectures will be online and available via DEN
 - <https://courses.uscden.net/>
 - Email me if you wish to attend my recording of the lecture
- Both exams are open book and available online, there is no final exam
 - Exam 1, Sept. 29th
 - Exam 2, Dec. 1st
- TA/grader interactions and support will be offered via Piazza
- There are five homework assignments with the last one a comprehensive project



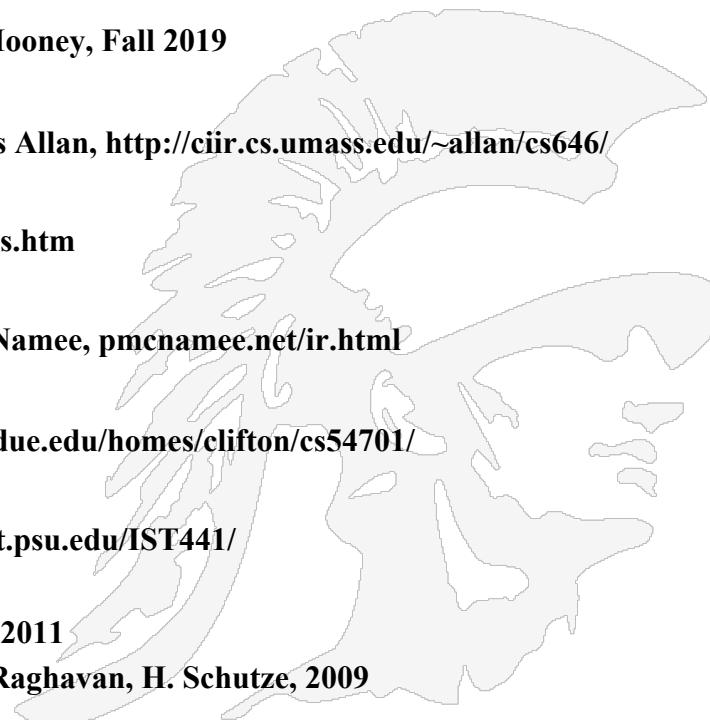
The Class Goals

- This course will give you
 - a broad understanding of how search engines work, and
 - will equip you to participate in designing and developing search engine capabilities
 - while covering the basic elements of information retrieval
- It discusses the fundamental building blocks of web search engines including:
 - crawling the web
 - indexing the data for fast retrieval
 - query processing including special ranking algorithms e.g. PageRank and HITS
 - a detailed analysis of the Google Search Engine including: Google File System and Map/Reduce



Special Acknowledgments

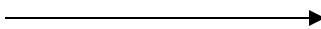
- The following courses cover similar material and I wish to acknowledge some of the instructors for making their notes available:
- Stanford University
 - CS 276, Information Retrieval and Web Search, Spring 2019, Manning and Nayak
 - <http://web.stanford.edu/class/cs276/>
- Univ. of Texas, Austin
 - CS 371R, Information Retrieval and Web Search, Ray Mooney, Fall 2019
- Univ. of Mass.
 - CMPSCI 646, Information Retrieval, Spring 2014, James Allan, <http://ciir.cs.umass.edu/~allan/cs646/>
- Cornell University
 - <https://www.cs.cornell.edu/courses/cs4300/2013fa/lectures.htm>
- Johns Hopkins
 - 605.766 Information Retrieval, Fall 2019, Prof. Paul McNamee, pmcnamee.net/ir.html
- Purdue University
 - CS 54701, Information Retrieval Fall, 2019 <http://cs.purdue.edu/homes/clifton/cs54701/>
- Penn State
 - IST 441, Prof. C. Lee Giles, Spring, 2019, <http://clgiles.ist.psu.edu/IST441/>
- Also, the following textbooks provided a variety of material
 - *Mining Massive Data Sets*, A. Rajaraman and J. Ullman, 2011
 - *Introduction to Information Retrieval* by C. Manning, P. Raghavan, H. Schutze, 2009
 - *Mining the Web*, S. Chakrabarti, 2003



Course Website

<http://csci572.com/>

Home Page



Screenshot of a web browser displaying the course website at <http://csci572.com/>. The page features the University of Southern California logo at the top right. The main content area is titled "Course 572: Information Retrieval and Web Search Engines Fall, 2022". It includes a video player showing a thumbnail of Professor Horowitz, with options to "Watch later" and "Share". Below the video, the USC Viterbi School of Engineering logo is displayed. A table provides contact information for the instructor:

Instructor: Professor Horowitz	
Class Meeting Time:	I do not take attendance All lectures will be available on DEN All exams can be taken remotely
Instr. Phone:	(213) 740-8056
E mail:	horowitz (at) usc.edu
Class Location:	SGM 124
Instr. Office:	SAL 320
Office Hours:	email Prof. Horowitz

Course Lectures

CSCI 572 Home Page Not Secure | csci572.com

CSCI 572 Home P... Piazza Spring2022 CSCI572_Spring2... DEN D2L Page USC Schedule of... Computer Science... Other Bookmarks

 Home Page Recent Search Engine Articles Schedule of Lectures Assignments Special Resources Course Grading Course Materials Class News Group Piazza Academic Integrity Policy How to Submit Your Homework

Schedule of Lectures

Please observe that all of these notes are copyrighted and are for the exclusive use of students enrolled in CS572. Please do not make the login&password available to others, and please do not distribute copies to others.

Date	Topic	Course Notes	Readings from Introduction to Information Retrieval, Manning, RagHAVAN, Schutze, 581 page book, pdf All GOTO links below refer to chapters/sections in the above book.	Video Lectures, some taken from <ul style="list-style-type: none"> Information Retrieval Course Matching Our Textbook 8th International Summer School on Information Retrieval NOTE: Videos with a double star (**) are required for the exam; You are only responsible for videos with **
Week 1 Aug 23	A: Introduction to the Course B: Search Engine Basics	PPT, PDF PPT, PDF	How Google Fights Disinformation	
Week 1 Aug 25	Characterizing the Web Discussion of Homework #1	PPT, PDF		<ul style="list-style-type: none"> How Google Founders Changed The World (start at 4:16 and watch till bored) Other useful videos <ul style="list-style-type: none"> WDM1:What is Data Mining.(8 min) WDM2:Structured Data and IR (17 min)

- I will normally lecture for 1 hour;

- We will then watch one or more relevant videos

- All slides are available in PPT and PDF
- Login/password for the class notes are:
csci572 notes

CSCI 572 Home Page Not Secure | csci572.com

Apps CSCI 572 Home P... CSCI 571 - Home Computer Science... ITS - Software Masters Student... University of Sout... Other Bookmarks

 Home Page Recent Search Engine Articles Schedule of Lectures Assignments Special Resources Course Grading Course Materials Class News Group Piazza Your Grades Academic Integrity Policy

CS572 Course Issues

Examinations

There are two exams, check the Schedule of Classes for dates

Grading

- 25% comes from exam 1
- 25% comes from exam 2
- 10% of your grade comes from your first programming assignment (HW1)
- 10% of your grade comes from your second programming assignment (HW2)
- 10% of your grade comes from your third programming assignment (HW3)
- 10% of your grade comes from your fourth programming assignment (HW4)
- 10% of your grade comes from your fifth programming assignment (HW5)

NO Re-Grading of Exams

Unfortunately, due to the size of the class there will be **NO** opportunity to see your exam and have it re-graded. With over 300 students we have neither the time nor the staff to go over each person's exam a second time. If you do not agree with this rule, please drop the class now, as these are the ground rules.

LIMITED Re-Grading Opportunities of Exams

Unfortunately, due to the size of the class there will be **NO** opportunity for all to see your exam and have it re-graded.
Please do not take this class if you do not agree with this policy

Course Grading

- **50%, 5 class programming assignments**

each worth 10%;

- **50% two exams;**
- **Re-grading opportunities only for those scoring 5 or more points below the average**

Special Resources

Videos, Data sets, Algorithms, Papers



CSCI 572 Home Page Not Secure | csci572.com

Selected Videos Selected DataSets Selected Algorithms Selected Papers

Selected Videos

- Problems with Google Search. TED talk, 2011(9 min)
- Google Developers Website, Numerous Useful Videos
- Google's Tech Talk Channel on YouTube (lots of general subjects)
- Challenges in Building Large-Scale Information Retrieval Systems by Jeff Dean, Google Fellow, Feb. 2009. (1hr. 5min)
- Machine Learning in Information Retrieval by Thomas Hoffmann, a 4 hour tutorial, Sept. 2004
- Some Key Challenges in Web Crawlers and Content-Based Search Engines, July 2006. (27min)
- Predicting the Present with Search Engine Data, Hal Varian, Aug. 2013. (51min)
- Experiences with the Nutch search engine by Doug Cutting, July, 2006. (1hr. 11min)
- Building blocks for semantic search engines by Soumen Chakrabarti, July 2006. (1hr 4min)
- Inside Google's Search Office, (1 hr 27 min)
- The Structured Search Engine by Andrew Hogue, Google, Jan, 2011 (1hr. 19min)
- Building Software at Google Scale by Michael Barnathan et al March, 2012 (1hr. 10min)
- Large-scale data analysis using the app engine pipeline API by Brett Slatkin, May 2011. (51min)
- JavaScript Testing at Google Scale by Cory Smith, August, 2011. (1hr. 7min)
- Doubleclick Ad Exchange by E. Manor et al, August, 2012. (1hr.20min)
- How Google Makes Improvements to its search algorithm August 2011.,(0:4min)
- How Google's Algorithm Works, Dec 2012. (0:4min)

Google Videos Focusing on Women in the Workplace

- Kathi Pham, Falling in Love with CS (8 min)
- Pavni Diwanji, VP of Engineering (1 min)
- Elena Kon, Embracing Failure (7 min)
- Women Tech Makers Rani Parchuri (5 min)
- Women Tech Makers Haringi Muri (6 min)
- Product Management in Tech Gayathri Rajan Episode 2 (6 min)
- Product Management in Tech, Aparna Chennapragade, Episode 3 (7 min)
- Product Management in Tech, Shimrit Ben-Yair Episode 4 (11 min)

Other items of interest [Google Algorithm Change History](#)

lynda.com is a website that develops online software training, or as they advertise it "videos that really work". USC has paid for the entire campus to have free access to these videos. Many of the videos have to do with various aspects of Web Development. Many of the videos are quite good.

go to <http://www.usc.edu/its/lynda/>
and login. It may be possible to do so from off campus as well as on-campus as the USC Shibboleth prompts for you USC login and password.

Their entire list of videos can be found at <http://www.lynda.com/Web-Interactive-training-tutorials/88-0.html>
Below is a selection of videos that apply directly to this course.

- (Start by logging in to www.usc.edu/its/lynda/)
 - SEO Fundamentals
 - Improving SEO Using Accessibility Techniques
 - Search Engine Optimization Getting Started (2010)
 - Analyzing Your Website to Improve SEO
 - Flash CS4 Professional: Building Search Engine Friendly Sites
 - SEO: Link Building in Depth



Course Assignments

CSCI 572 Home Page Not Secure | csci572.com

CSCI 572 Home P... Piazza Fall 2022 Fall2022Schedule... CSCI572_Fall202... DEN D2L Page Computer Science... University of Sout...

Home Page Recent Search Engine Articles Schedule of Lectures Assignments Special Resources Course Grading Course Materials Class News Group Piazza Academic Integrity Policy How to Submit Your Homework

CS572 Course Assignments
Last Modified: July 12, 2022

Homework 1: Comparing Search Engine Results

- [Search Engine Comparison Exercise](#)
- [100QueriesSet1 Google Result1](#)
- [100QueriesSet2 Google Result2](#)
- [100QueriesSet3 Google Result3](#)
- [100QueriesSet4 Google Result4](#)
- [Grading Guidelines](#)
- [Homework #1 Due Sep 06](#)

Homework 2: Web Crawling

Homework 3: Creating an Inverted Index Using a Hadoop Cluster

Homework 4: Indexing the Web Using Solr

Homework 5: Enhancing Your Search Engine

Late Assignment Policy
Homework submitted for grading before or on the "Homework Due Date", as listed in the Schedule of Lectures, will be eligible for 100% of the grade points for the assignment. Homework submitted late will be accepted for up to 7 calendar days after the due date, and will receive an automatic 10% penalty. Homework submitted more than 7 days after the due date will not be accepted.

- There will be five programming assignments

Deliverables

- a short write-up of your solution to the assignment, and
- any source code you have written

- Due dates are listed on the *Assignments* page and the *Schedule of Lectures* page

Course Materials

A set of class notes will be made available.

Required Text

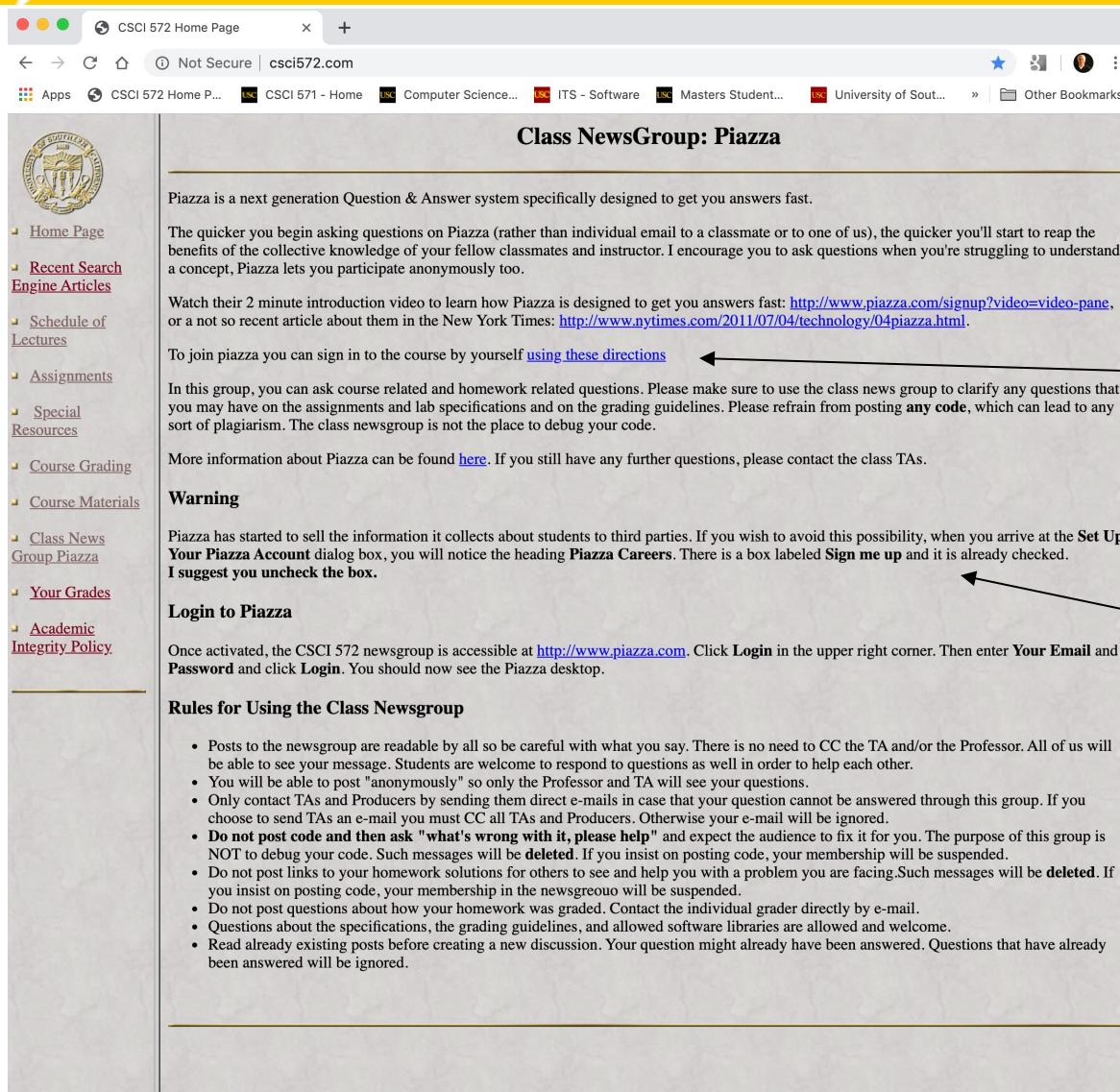
- Introduction to Information Retrieval*, Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Cambridge University Press, 2008.
[HTML Edition](#)
[PDF for online viewing](#)
[PDF of the book for printing](#)
[Slides are located here](#)
[If all else fails try this link to download the book](#)

Recommended Texts

- Mining Massive Data Sets*, A. Rajaraman and J. Ullman, Pre-print version
- Mining Massive Data Sets, 2nd Edition*, J. Leskovec, A. Rajaraman, J. Ullman
- Data-Intensive Text Processing with MapReduce*, Jimmy Lin and Chris Dyer, Morgan & Claypool
- Search Engines: Information Retrieval in Practice*, Bruce Croft, Donald Metzler, Trevor Strohman, Addison Wesley, 2010
[Slides are located here](#)
- Modern Information Retrieval*, Ricardo Baeza-Yates and Berthier Ribeiro-Neto, Addison-Wesley, 1999.

- *Introduction to Information Retrieval* by Manning et al is our main textbook;
- The book is available for free via PDF download
- It presents all of the classic Information Retrieval concepts

Class NewsGroup Piazza

A screenshot of a web browser window titled "CSCI 572 Home Page". The address bar shows "csci572.com". The page content is titled "Class NewsGroup: Piazza". It contains several sections of text and links. On the left sidebar, there is a list of course links including "Home Page", "Recent Search Engine Articles", "Schedule of Lectures", "Assignments", "Special Resources", "Course Grading", "Course Materials", "Class News Group Piazza", "Your Grades", and "Academic Integrity Policy". A large arrow points from the right side of the page towards the sidebar.

Piazza is a next generation Question & Answer system specifically designed to get you answers fast.

The quicker you begin asking questions on Piazza (rather than individual email to a classmate or to one of us), the quicker you'll start to reap the benefits of the collective knowledge of your fellow classmates and instructor. I encourage you to ask questions when you're struggling to understand a concept, Piazza lets you participate anonymously too.

Watch their 2 minute introduction video to learn how Piazza is designed to get you answers fast: <http://www.piazza.com/signup?video=video-pane>, or a not so recent article about them in the New York Times: <http://www.nytimes.com/2011/07/04/technology/04piazza.html>.

To join piazza you can sign in to the course by yourself [using these directions](#)

In this group, you can ask course related and homework related questions. Please make sure to use the class news group to clarify any questions that you may have on the assignments and lab specifications and on the grading guidelines. Please refrain from posting **any code**, which can lead to any sort of plagiarism. The class newsgroup is not the place to debug your code.

More information about Piazza can be found [here](#). If you still have any further questions, please contact the class TAs.

Warning

Piazza has started to sell the information it collects about students to third parties. If you wish to avoid this possibility, when you arrive at the **Set Up Your Piazza Account** dialog box, you will notice the heading **Piazza Careers**. There is a box labeled **Sign me up** and it is already checked. I suggest you uncheck the box.

Login to Piazza

Once activated, the CSCI 572 newsgroup is accessible at <http://www.piazza.com>. Click **Login** in the upper right corner. Then enter **Your Email** and **Password** and click **Login**. You should now see the Piazza desktop.

Rules for Using the Class Newsgroup

- Posts to the newsgroup are readable by all so be careful with what you say. There is no need to CC the TA and/or the Professor. All of us will be able to see your message. Students are welcome to respond to questions as well in order to help each other.
- You will be able to post "anonymously" so only the Professor and TA will see your questions.
- Only contact TAs and Producers by sending them direct e-mails in case that your question cannot be answered through this group. If you choose to send TAs an e-mail you must CC all TAs and Producers. Otherwise your e-mail will be ignored.
- **Do not post code and then ask "what's wrong with it, please help"** and expect the audience to fix it for you. The purpose of this group is NOT to debug your code. Such messages will be **deleted**. If you insist on posting code, your membership will be suspended.
- Do not post links to your homework solutions for others to see and help you with a problem you are facing. Such messages will be **deleted**. If you insist on posting code, your membership in the newsgroup will be suspended.
- Do not post questions about how your homework was graded. Contact the individual grader directly by e-mail.
- Questions about the specifications, the grading guidelines, and allowed software libraries are allowed and welcome.
- Read already existing posts before creating a new discussion. Your question might already have been answered. Questions that have already been answered will be ignored.

If you are not
already signed up
you can sign up
yourself

BEWARE

Review the rules

Academic Integrity Policy

CSCI 572 Home Page Not Secure | csci572.com

Apps CSCI 572 Home P... USC CSCI 571 - Home USC Computer Science... USC ITS - Software USC Masters Student... USC University of Sout... Other Bookmarks

 [Home Page](#)
[Recent Search Engine Articles](#)
[Schedule of Lectures](#)
[Assignments](#)
[Special Resources](#)
[Course Grading](#)
[Course Materials](#)
[Class News Group Piazza](#)
[Your Grades](#)
[Academic Integrity Policy](#)

Course Academic Integrity Policy

Statement on Academic Conduct and Support Systems

Academic Conduct Plagiarism - presenting someone else's ideas as your own, either verbatim or recast in your own words - is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in SCampus in Section 11, Behavior Violating University Standards <https://scampus.usc.edu/1100-behavior-violating-university-standards-and-appropriate-sanctions/>. Other forms of academic dishonesty are equally unacceptable. See additional information in SCampus and university policies on scientific misconduct, <http://policy.usc.edu/scientific-misconduct/>. Discrimination, sexual assault, and harassment are not tolerated by the university. You are encouraged to report any incidents to the Office of Equity and Diversity <http://equity.usc.edu/> or to the Department of Public Safety <http://capsnet.usc.edu/department/department-public-safety/online-forms/contact-us>. This is important for the safety whole USC community. Another member of the university community - such as a friend, classmate, advisor, or faculty member - can help initiate the report, or can initiate the report on behalf of another person. The Center for Women and Men <http://www.usc.edu/student-affairs/cwm> provides 24/7 confidential support, and the sexual assault resource center webpage [sarc@usc.edu](http://sarc.usc.edu) describes reporting options and other resources.

Support Systems

A number of USC's schools provide support for students who need help with scholarly writing. Check with your advisor or program staff to find out more. Students whose primary language is not English should check with the American Language Institute <http://dornsife.usc.edu/ali> which sponsors courses and workshops specifically for international graduate students. The Office of Disability Services and Programs http://sait.usc.edu/academic-support/center-programs/dsp/home_index.html provides certification for students with disabilities and helps arrange the relevant accommodations. If an officially declared emergency makes travel to campus infeasible, USC Emergency Information <http://emergency.usc.edu/> will provide safety and other updates, including ways in which instruction will be continued by means of blackboard, teleconferencing, and other technology.

Statement on Diversity

The diversity of the participants in this course is a valuable source of ideas, problem solving strategies, and engineering creativity. We encourage and support the efforts of all of our students to contribute freely and enthusiastically. We are members of an academic community where it is our shared responsibility to cultivate a climate where all students and individuals are valued and where both they and their ideas are treated with respect, regardless of their differences, visible or invisible.

Honor Code

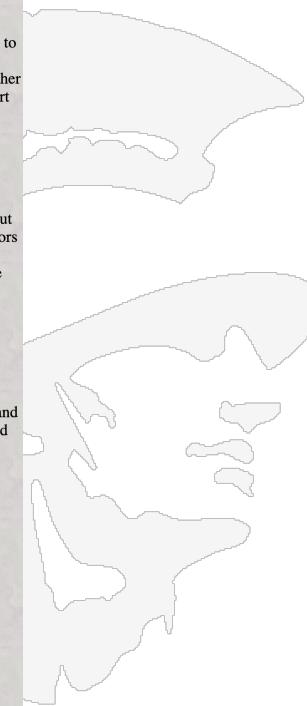
Engineering enables and empowers our ambitions and is integral to our identities. In the Viterbi community, accountability is reflected in all our endeavors.

Engineering+ Integrity.
Engineering+ Responsibility.
Engineering+ Community.
Think good. Do better. Be great.

These are the pillars we stand upon as we address the challenges of society and enrich lives.

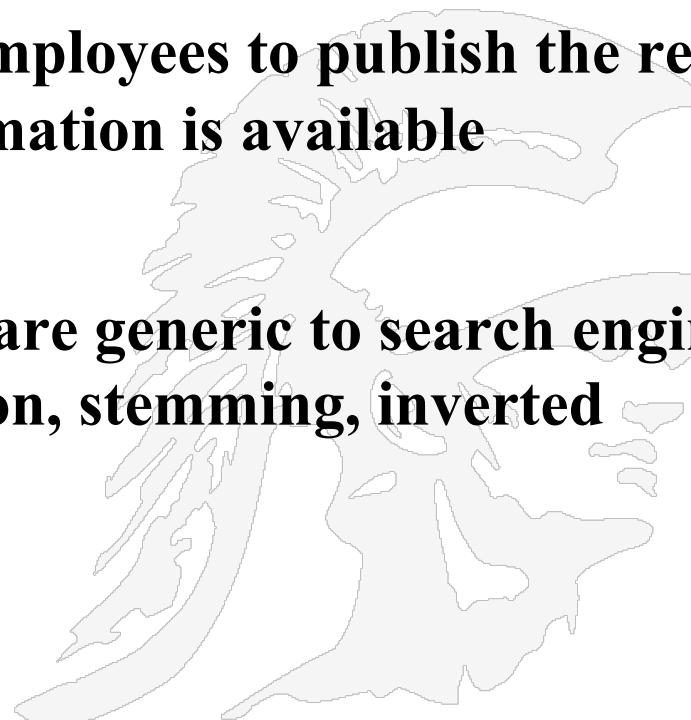
Public Cloud Source Code Repositories

Some students have been posting the code for CSCI 572 exercises on Github's public repository. Repositories like GitHub and Bitbucket are crawled by search engines. Thus they are easily found and copied. Anyone putting their source code in a public repository is playing with fire and putting their grade at risk. **Public repositories like GitHub and Bit bucket should not be used to store CSCI 572 homework source code.** Private repositories from the same services can be used.



Is this Course Just About Google?

- Yes
 - Google has been innovating in search engines so it makes sense to study what they have done
 - Google encourages their employees to publish the results of their work, so the information is available
- No
 - Many topics in the course are generic to search engines, e.g. crawling, de-duplication, stemming, inverted indexes, etc



So...today

- Quick lecture on search engine history and usage



Once Again

- All video lectures are available to everyone at <https://courses.uscden.net/>
- Email me if you wish to attend my recording of the lecture
- The two exams will be open book and online and are scheduled for:
 - Exam 1, Sept 29th
 - Exam 2, Dec 1st
 - **POTENTIAL PROBLEM:** csci571 has exams the SAME DAY at 6:00-6:40pm
- Email me if you wish to attend my taping session
- All slide lectures are available to everyone at the website:
Login/password for the slides are: csci572/notes
- **REMINDER: Re-Grading of Exams will only be permitted for those people scoring 5 or more points below the average**

Please do not take this class if you do not agree with this policy
Copyright Ellis Horowitz, 2011 - 2022