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

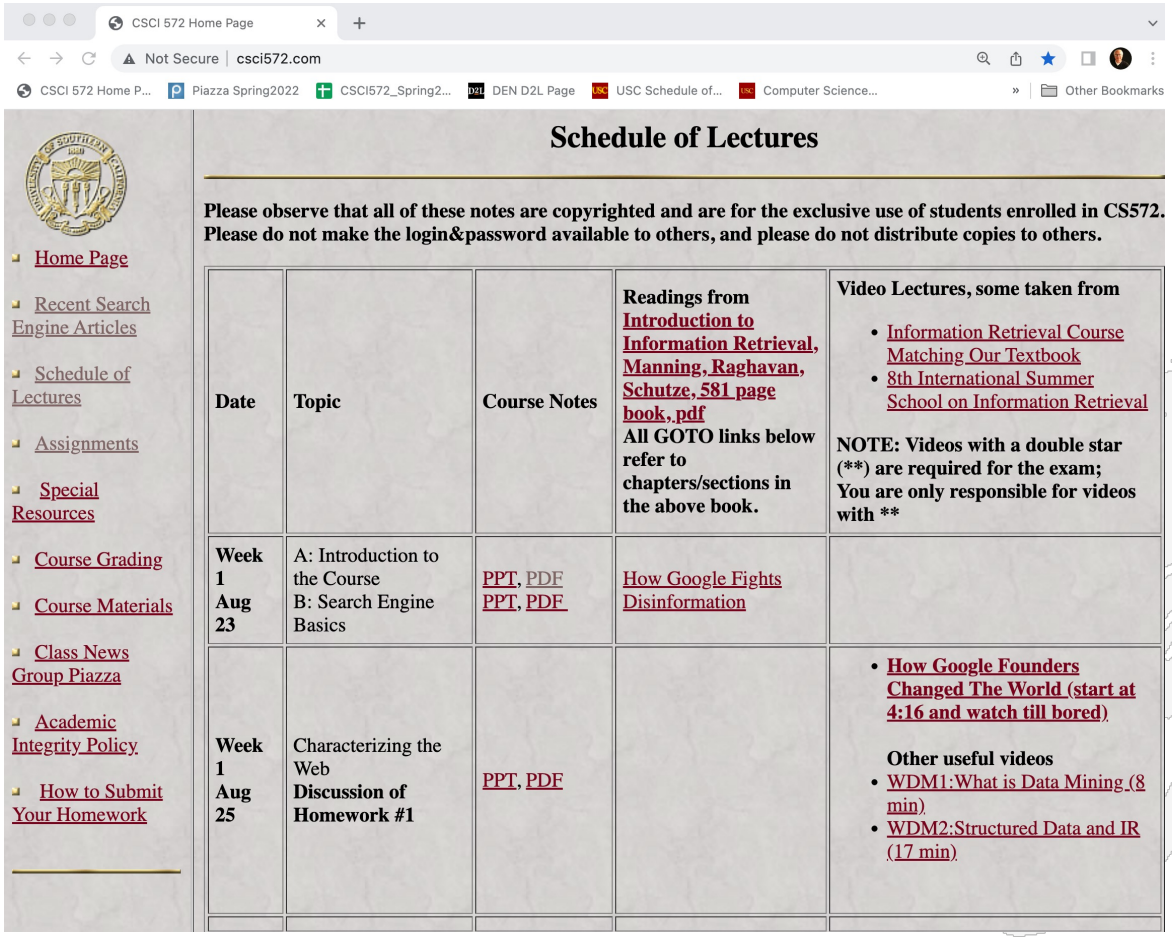
- **The following courses cover similar material and I wish to acknowledge some of the instructors for making their notes available:**
- **Stanford University**
  - **CS 276, Information Retrieval and Web Search, Spring 2019, Manning and Nayak**
  - **http://web.stanford.edu/class/cs276/**
- **Univ. 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 →

Course Lectures



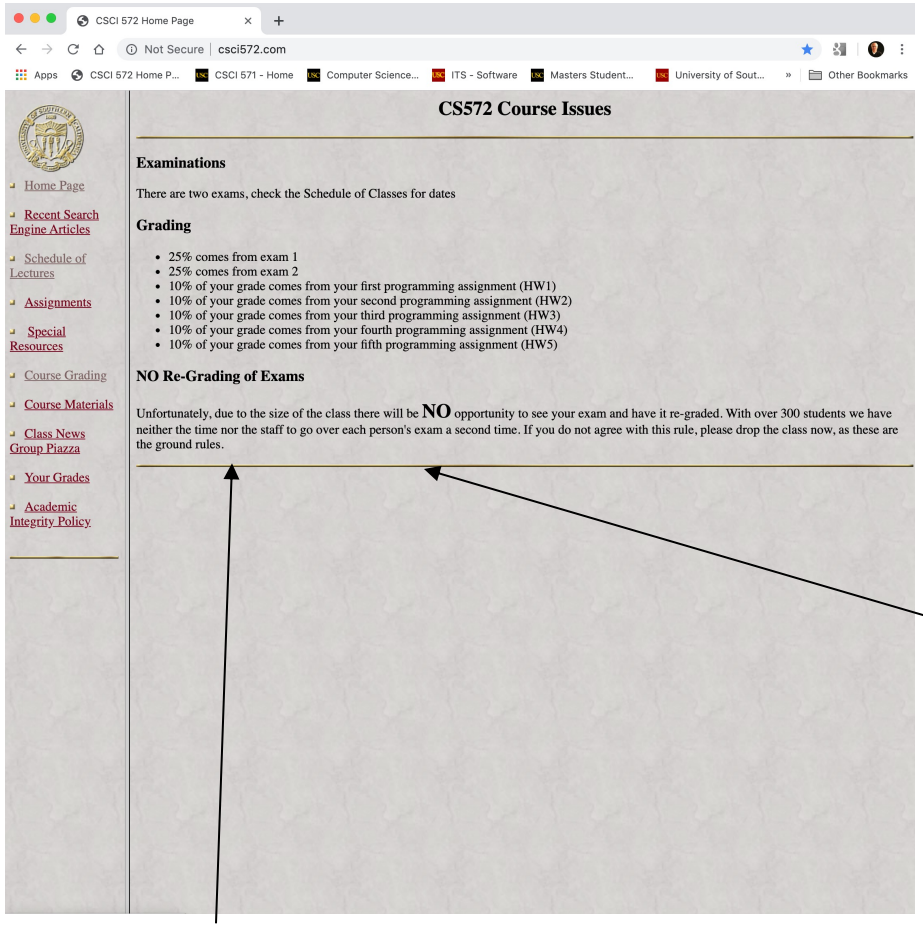
Schedule of Lectures

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 • 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		• How Google Founders Changed The World (start at 4:16 and watch till bored) Other useful videos • 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**

Course Grading



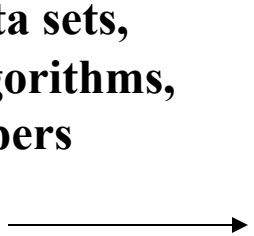
- **50%, 5 class programming assignments each worth 10%;**
- **50% two exams;**
- **Re-grading opportunities only for those scoring *5 or more points below the average***

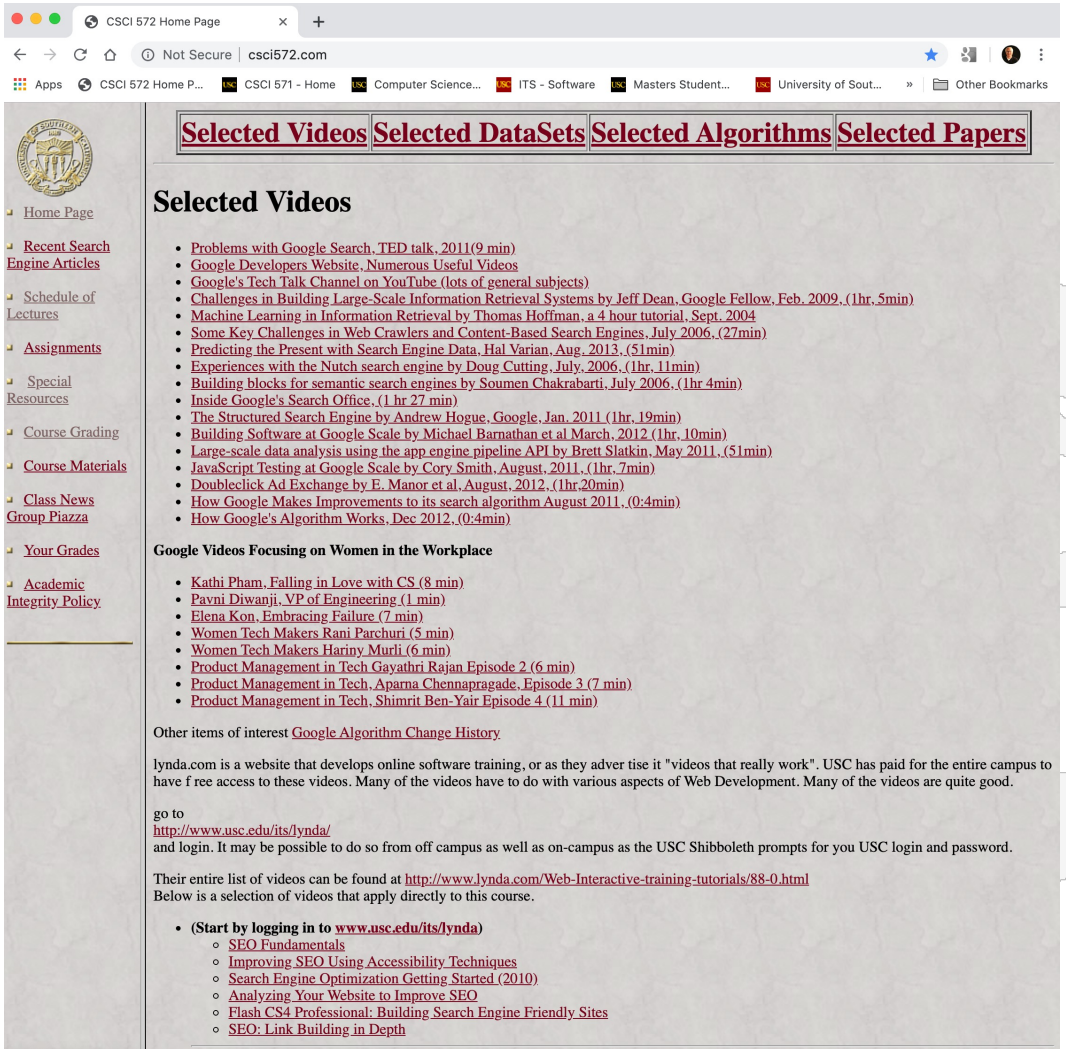
LIMITED Re-Grading Opportunities of Exams

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

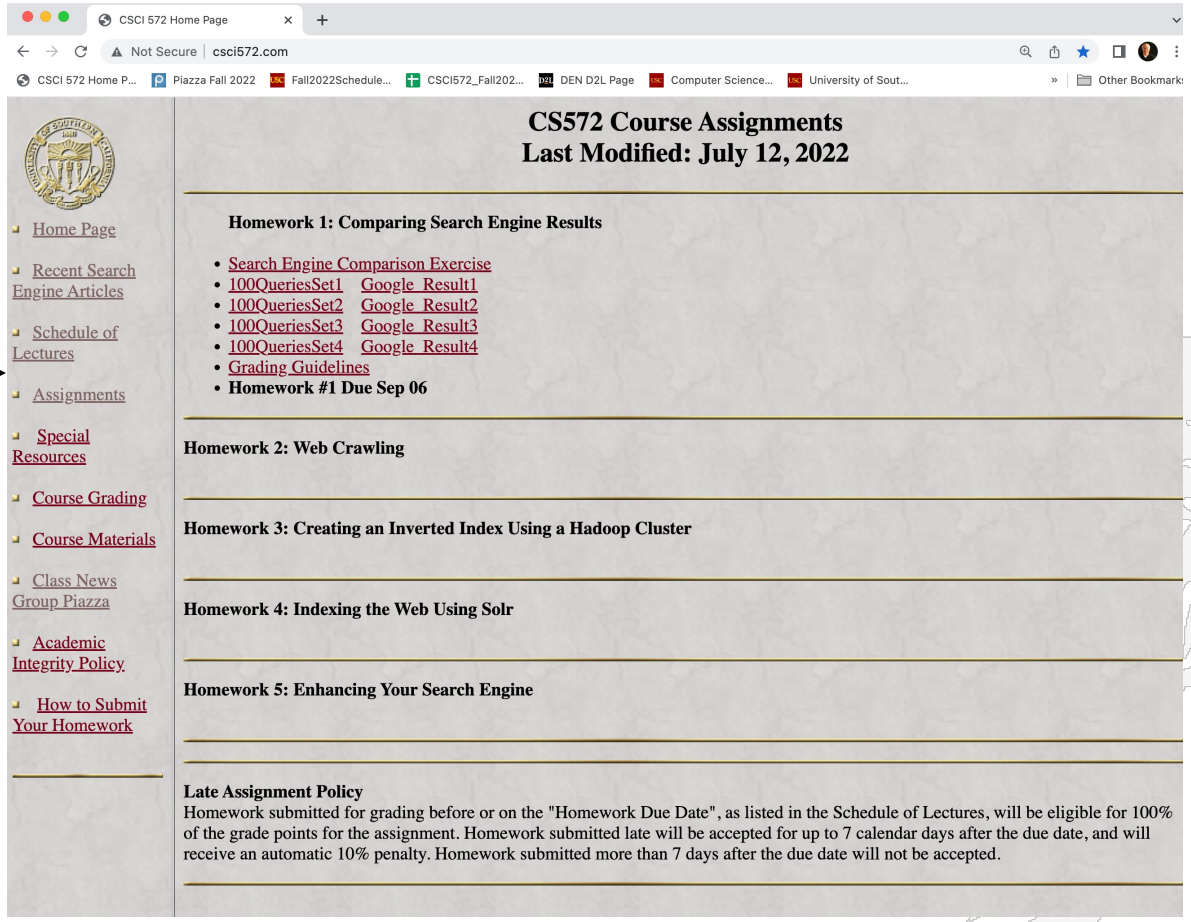
Special Resources

Videos,
Data sets,
Algorithms,
Papers



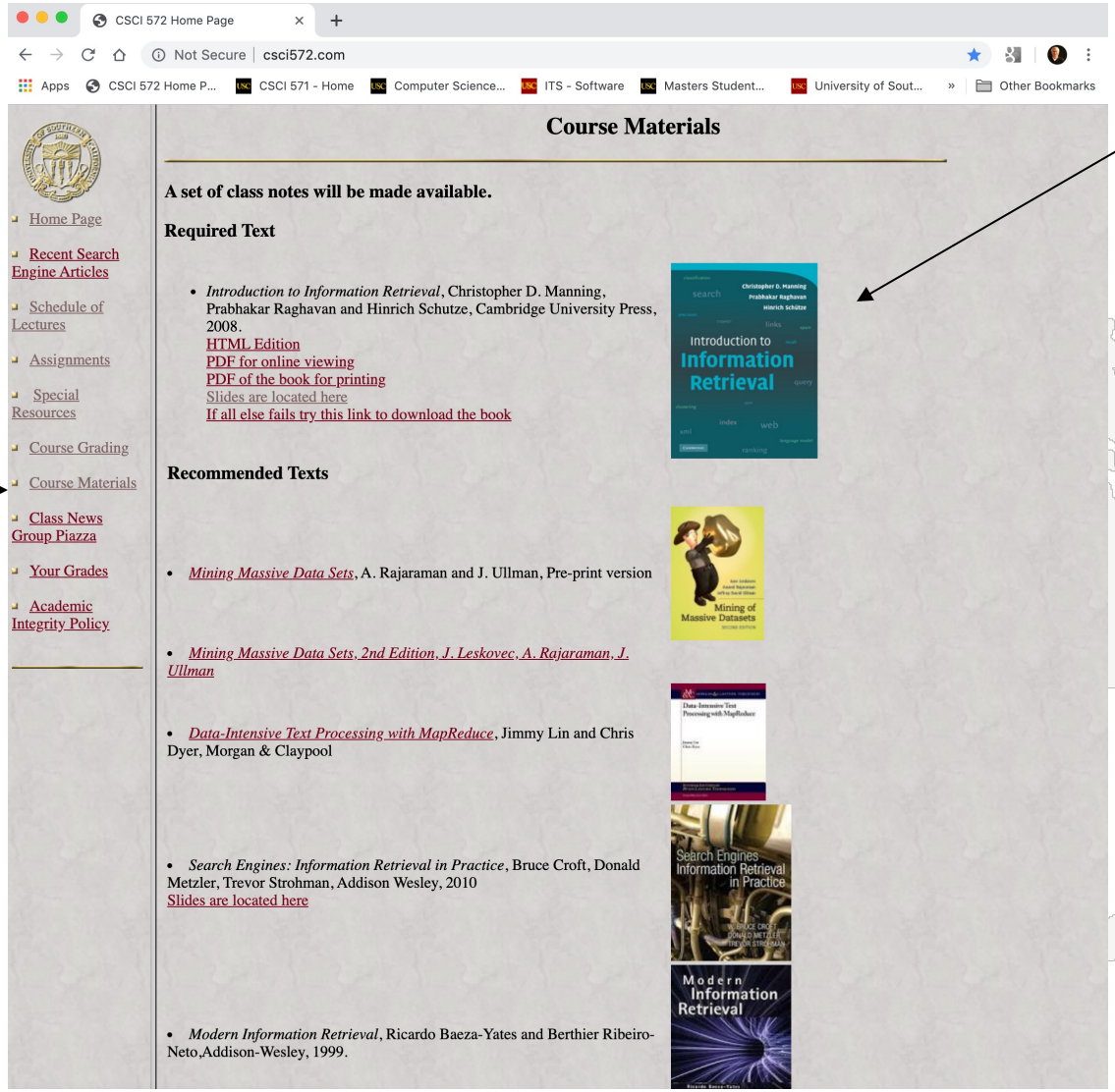


Course Assignments



- 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

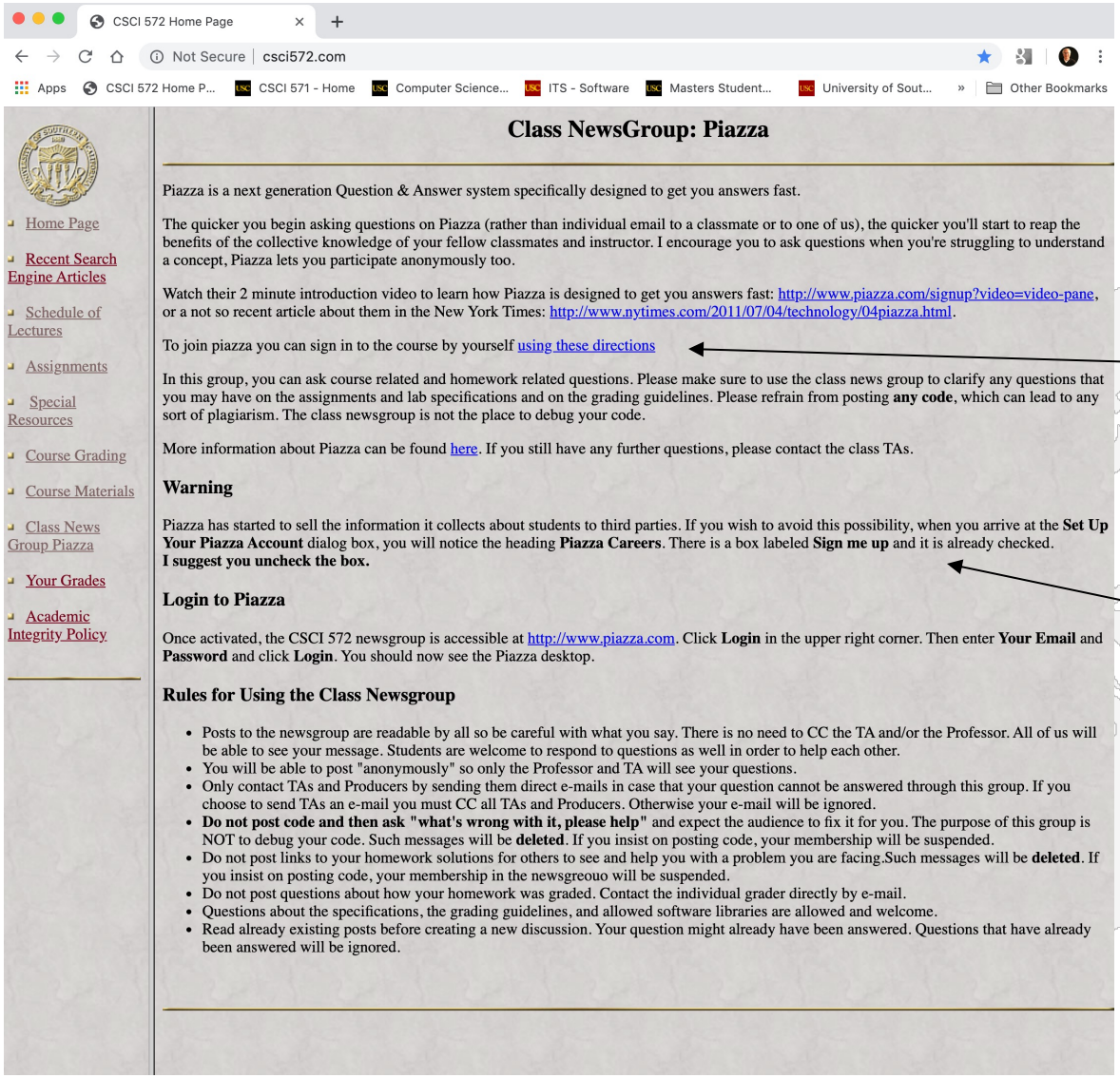


- *Introduction to Information Retrieval* by Manning et al

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

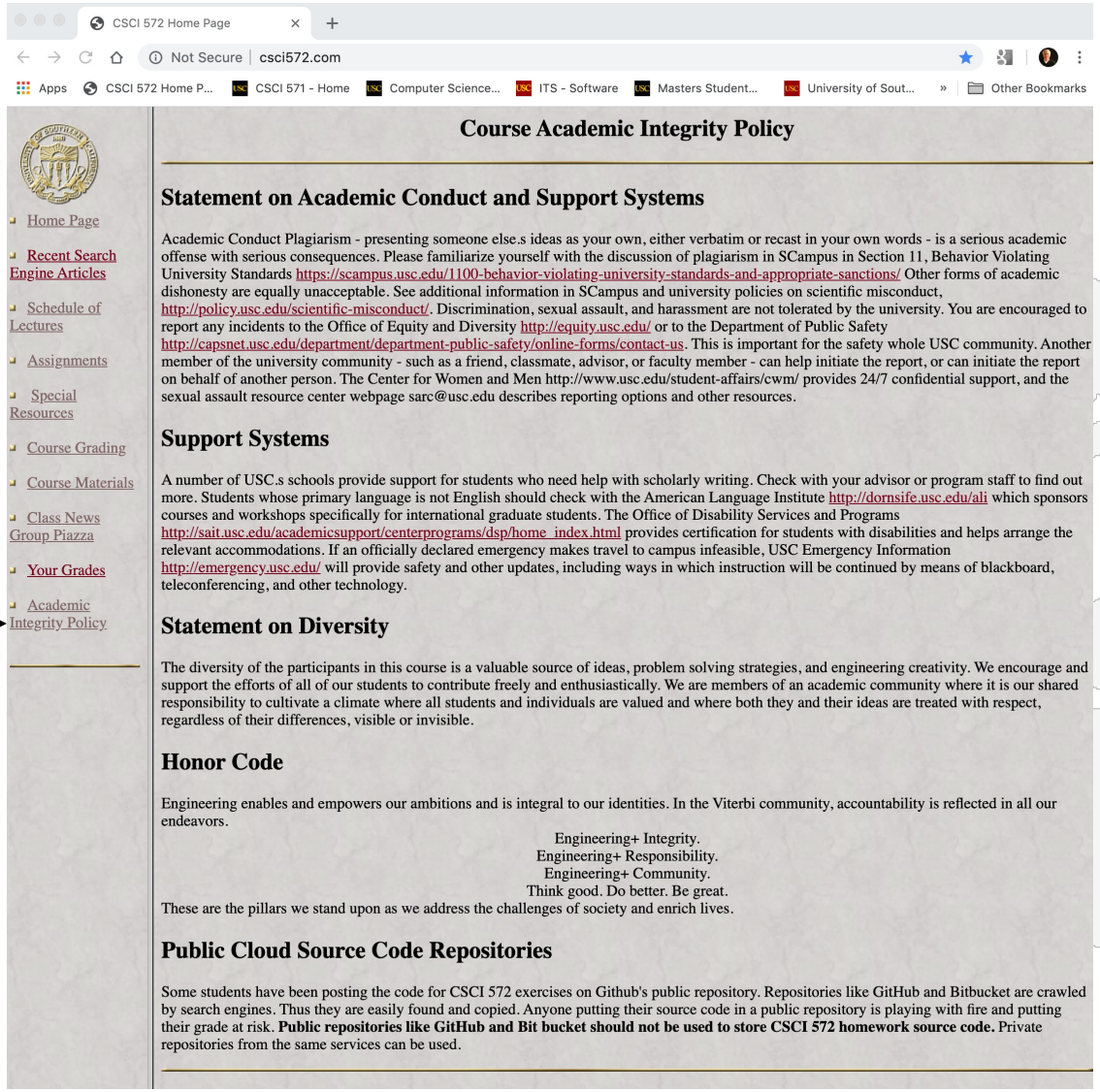


If you are not already signed up you can sign up yourself

BEWARE

Review the rules

Academic Integrity Policy

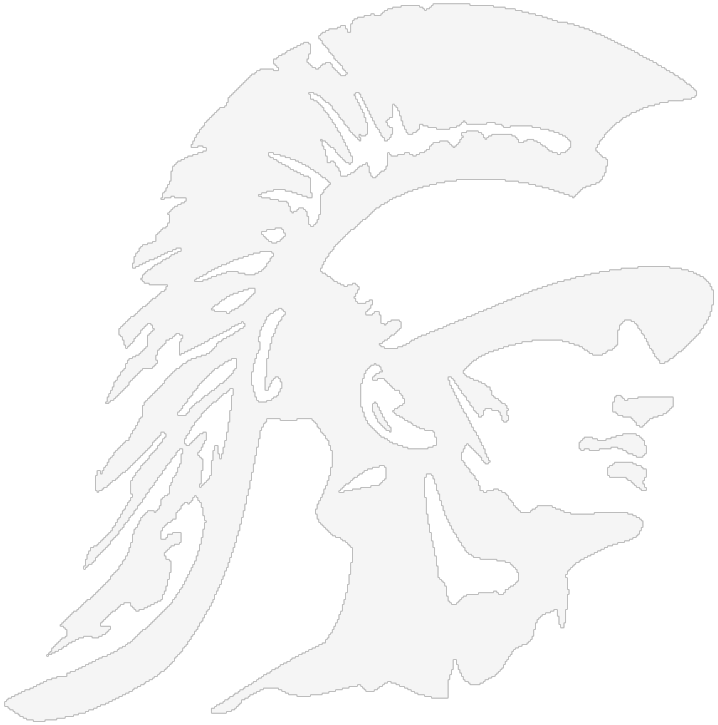


Is this Course Just About Google?

- 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 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