CSC 555 Mining Big Data

Alexander Rasin [/about/pages/people/facultyinfo.aspx?fid=1041]
Office: CDM 847
Fall 2018-2019
Class number: 15960
Section number: 710
-
Online Campus
Course homepage: http://d2l.depaul.edu/ [http://d2l.depaul.edu/]

Summary

This is a graduate course in large scale data mining applications. Specific topics to be covered include:

* Fundamentals of distributed file systems and MapReduce (MR) technology
* Advantages of an MR-based system compared to a relational database
* Tuning MR algorithm performance and tools for mining massive data sets
* Hadoop-based tools for clustering, similarity search, classification and data warehousing

Texts

Mining of Massive Datasets, by Anand Rajaraman and Jeffrey D. Ullman, Cambridge University Press. Free download at http://i.stanford.edu/~ullman/mmds.html

Hadoop: The Definitive Guide, by Tom White, O'Reilly Media, 4th edition 2015. ISBN-13: 978-1491901632

Grading

There will be homework assignments given most weeks; assignments (with associated readings) will be posted on the course web site and will be due one week after the day they ar posted, unless otherwise noted. Details of the submission process will be discussed in class; it is your responsibility to verify that your submitted files are readable and submitted in th correct locations. Late assignments will be accepted up to three days late with a 10% penalty for each day or fraction of a day that the assignment is late; these penalties will be assessed uniformly and in full to all assignments submitted at any point beyond the posted due date and time (including those submitted or re-submitted later the same day). The homework assignments will be worth a total of 45% of the course grade. There will be no exams in this class; instead, students will work on a final project to apply the concepts covered in class. The final project will be worth 55% of the grade and will consist of two parts: Part 1 due at midterm mark and Part 2 due on the day of the scheduled final exam.

Prerequisites

((CSC 401 and (CSC 453 or DSC 450) and (DSC 441 or DSC 478)) or (MAT 491 and MAT 449))

Regarding Email Communication

Please begin the subject line of any email to me with "CSC 555", so that I can easily identify your messages. I will reply to email messages within one business day after the day I receive them; therefore questions that are only received by me on an assignment's due date (or late the night before) are not guaranteed replies before the assignment is due. Please plan accordingly and begin the assignments early enough to ask questions and receive answers. If you are having problems, send me a detailed description of the problems you are having; I will try to guide you in locating and solving your problems yourself, rather than simply solve your problems for you.

Regarding Academic Integrity

You are expected to be familiar with and to adhere to DePaul's Academic Integrity Policy, which is available on-line at http://academicintegrity.depaul.edu/AcademicIntegrityPolicy.pdf Violations of the Academic Integrity Policy will be dealt with decisively; penalties may range up to an automatic F in the course and possible expulsion.

Plagiarism includes, but is not limited to: Turning in another person's work as your own (including hiring someone else to complete an assignment for you); Starting with another person's work and modifying it to turn in as your own; Cutting and pasting, or otherwise copying, sections of another person's work into your assignment; Allowing another person (such as a tutor) to write part of your assignment; and so on. (Obviously, any examples that I post qualify as "another person's work".) Supplying such assistance to another student is considered an equivalent violation of the policy. You may feel free to discuss the assignments with other students at a general level. However, when it comes to actually completing your assignment, you must work independently. Your assignments must be entirely your own individual work. If you have any questions or doubts about what plagiarism entails, you should consult me.

Week 1 (Sept 6th): Relational Databases, Hadoop
- Relational database review
- Cloud computing
- Introduction to Hadoop (HDFS, MapReduce fundamentals)

Week 2 (Recorded on Sept 7th instead of 13th!): Hadoop, Distributed Computing
- Hadoop (vs R-DBMS) trade-offs
- Performance issues, parallel computing, intro to hashing

- Implementing relational operators in Hadoop

Week 3 (Sept. 20th): Virtual Instances and Performance Tuning
- Remote/Virtual instances on the cloud
- Matrix multiplication and Hadoop monitoring
- Hadoop performance tuning (replication, compression, distribution)

Week 4 (Sept. 27th): Hadoop Ecosystem
- Hadoop Examples, Hadoop 2.0
- NoSQL data-store systems
- Hive

Week 5 (Oct. 4th): Hive and PIG
- Hive, PIG
- Public/private keys and encryption
- NoSQL, Mahout

Week 6 (Oct. 11th): Hadoop and Mahout
- Hadoop Streaming (with python) - Mahout
- Similarity/distance measures (Ch3 from Rajaraman & Ullman)

Week 7 (Oct. 18th): Link Analysis and Clustering
- Hadoop cluster configuration - Link analysis (Ch5 from Rajaraman & Ullman)
- Clustering (Ch7 from Rajaraman & Ullman)

Week 8 (Oct. 25th)
- Frequent itemsets
- Document analysis
- Storm and Spark

Week 9 (Nov. 1st): Mining Data Streams
- Recommender systems (Ch9 from Rajaraman & Ullman)
- Streaming data engines
- Mining Data Streams (Ch4 from Rajaraman & Ullman)

Week 10 (Nov. 8th): Advertising on the Web
- Advertising on the web (Ch8 from Rajaraman & Ullman)
- Mining Social-Network Graphs (Ch10 from Rajaraman & Ullman)

School policies:

Changes to Syllabus

This syllabus is subject to change as necessary during the quarter. If a change occurs, it will be thoroughly addressed during class, posted under Announcements in D2L and sent via email.

Online Course Evaluations

Evaluations are a way for students to provide valuable feedback regarding their instructor and the course. Detailed feedback will enable the instructor to continuously tailor teaching methods and course content to meet the learning goals of the course and the academic needs of the students. They are a requirement of the course and are key to continue to provide you with the highest quality of teaching. The evaluations are anonymous; the instructor and administration do not track who entered what responses. A program is used to check if the student completed the evaluations, but the evaluation is completely separate from the student's identity. Since 100% participation is our goal, students are sent periodic reminders over three weeks. Students do nc receive reminders once they complete the evaluation. Students complete the evaluation online in CampusConnect [https://campusconnect.depaul.edu] .

Academic Integrity and Plagiarism

This course will be subject to the university's academic integrity policy. More information can be found at http://academicintegrity.depaul.edu/ [http://academicintegrity.depaul.edu/] If you have any questions be sure to consult with your professor.

Academic Policies

All students are required to manage their class schedules each term in accordance with the deadlines for enrolling and withdrawing as indicated in the University Academic Calendar [http://oaa.depaul.edu/what/calendar.jsp] . Information on enrollment, withdrawal, grading and incompletes can be found at http://www.cdm.depaul.edu/Current%20Students/Pages/PoliciesandProcedures.aspx [/Current% 20Students/Pages/PoliciesandProcedures.aspx] .

Students with Disabilities

Students who feel they may need an accommodation based on the impact of a disability should contact the instructor privately to discuss their specific needs. All discussions will remain confidential.
To ensure that you receive the most appropriate accommodation based on your needs, contact the instructor as early as possible in the quarter (preferably within the first week of class), and make sure that you have contacted the Center for Students with Disabilities (CSD) at:
Lewis Center 1420, 25 East Jackson Blvd.
Phone number: (312)362-8002

Fax: (312)362-6544
TTY: (773)325.7296

DePaul University
College of Computing and Digital Media
243 South Wabash Avenue
Chicago, IL 60604

(312) 362-8381

Contact Us

- Admission
- Advising
- General
- Website Feedback

CDM Schools

Cinematic Arts [/about/Pages/School-of-Cinematic-Arts.aspx]

Computing [/about/Pages/School-of-Computing.aspx]

Design [/about/Pages/School-of-Design.aspx]

Sign In

MyCDM [https://my.cdm.depaul.edu/authe section=mycti&title=MyCTI&url

D2L [https://d2l.depaul.edu/]

CDM Intranet [https://my.cdm.depaul.edu/cti/ad

Campus Connect [https://campusconnect.depaul.ed

Thanks for visiting! Please only print the pages you need in an effort to conserve. Tip: Next to links to other pages you will see the the url path for your convinience. Remember to add "http://www.cdm.depaul.edu" prefixing the shown path, in the square brackets, for links shown without the "http://".

privacy statement [/Pages/Privacy.aspx]