Algorithms for Modern Data Sets (COMP 3801)

Weekly Schedule

Assignments

Instructor: Anil Maheshwari
Office: HP 5125b
E-mail: anil@scs.carleton.ca

Lectures: Lectures are **Wednesday and Fridays at 08:35 to 09:55 AM**. See public class schedule for the location.

Office hours: **Wednesdays 10:15-11:45 AM (HP 5125b)**

Please feel free to send me email at anil@scs.carleton.ca

Teaching Assistant:

Course objectives: Algorithmic design techniques for modern data sets arising in, for example, data mining, web analytics, epidemic spreads, search engines and social networks. Topics may include data mining, hashing, streaming, clustering, recommendation systems, link analysis, dimensionality reduction, online, social networking, game theoretic and probabilistic algorithms.

Caution: Note that you need a minimum of B+ in COMP 2804 to register in this course. The contents of this course are fairly broad, and will cover a spectrum of techniques from the design and analysis of algorithms. It is assumed that you have a very good grasp on the analysis of algorithms (O-notation, recurrences, and complexity analysis), elementary probability theory including expectation and indicator random variables (contents of COMP 2804), the knowledge of basic data structures (lists, trees, hashing), and the knowledge of discrete mathematics (counting, permutations and combinations, proof techniques: induction, contradiction, ..). Note that there will not be time to review these material, and to appreciate the contents of this course, you must have a very good grasp on these topics.

Textbooks:

- Mining of Massive Datasets (MMDS) by Leskovec, Rajaraman, Ullman (Online at mmds.org and Cambridge University Press)
- Discrete Structures for Computer Science: Counting, Recursion, and Probability by Michiel Smid

- My Notes on Topics in Algorithm Design (we refer to them as ``Notes'')

Useful References related to various topics:

- Linear Algebra (Eigenvalues, Diagonalization, SVD)
  - Eigenvalues (1 2)
  - Video Lectures and the book of Gilbert Strang
  - Chapter on Matrices in My Notes on Topics in Algorithm Design

- PageRank (Markov Chains - aperiodic and irreducible)
  - Page Rank
  - Original paper by Brin and Page on PageRank: The anatomy of a large-scale hypertextual web search engine (1998).
  - What can you do with a web in your pocket by Brin, Motwani, Page and Winograd.
  - Link to the TED talk of Cedric Villani on "What's so sexy about Math" and the description of PageRank.
- Probability (Linearity of Expectation, Concentration Bounds, Balls and Bins, Applications to Hashing)
  - STAT110 on Youtube
  - Introduction to Probability book by Blitzstein and Hwang
  - Discrete Structures for Computer Science: Counting, Recursion, and Probability by Michiel Smid
  - Chapter on Probability in My Notes on Topics in Algorithm Design
  - Introduction to R
- Set Membership - Bloom Filters
  - Chapter in MMDS Book
  - My Notes on Algorithm Design
- Locality Sensitive Hashing (Documnet Similarity, minHash Signatures, Fingerprint matching)
  - Chapter 3 in MMDS Text Book
  - Chapter on LSH in My Notes on Topics in Algorithm Design
  - Useful references on LSH
  - STOC98 paper of Indyk-Motwani
- Data Streaming (Majority, Heavy Hitters, Count-Min Sketch, Estimating Frequency Moments, Counting 1s in Sliding Window)
  - Mining Data Streams chapter of MMDS
  - Wikipedia Article
  - DGIM Article
  - Count-Min Sketch Article
  - AMS Article (Approximating frequency moments)
  - Flajolet and Martin's Article on number of distinct elements in a stream
  - Some parts are covered in the Chapter in My Notes on Algorithm Design
  - My talk on Bloom Filters and Count-Min-Sketch
- Adwords, Online Bipartite Matching, b-Matching and the Balance Algorithm
  - Chapter on Advertisement on the Web from the MMDS Book.
  - For a detailed analysis look at My Notes

- Collaborative Filtering (UV Decomposition)
  - Chapter in MMDS Book

- SVD/CUR/Dimensionality Reduction:
  - Chapter in the MMDS book on Dimensionality Reduction
  - My Notes
  - Try a few SVDs using Wolfram Alpha.
- Randomized Load Balancing & Perfect Hashing
  - http://pages.cs.wisc.edu/~shuchi/courses/787-F09/scribe-notes/lec7.pdf
  - Kleinberg&Tardos Algorithm Design Book, Chapter 13.
- Clustering
  - My slides on the course web-page
  - k-means++ paper (Arthur and Vassilvitskii, 8th ACM-SIAM Symposium on Discrete algorithms, 2007)
- Max k-coverage problem
  - Maximizing the Spread of Influence through a Social Network by Kempe, Kleinberg and Tardos
  - Turning Down the Noise in the Blogosphere by El-Arini et al.
  - https://en.wikipedia.org/wiki/Maximum_coverage_problem
  - See Exercises 14.9 and 14.10 in Notes

Topics

We will likely cover parts of MMDS Chapters 3, 4, 5, 7, 8, 9, 11, and possibly 10 and 6. In addition to this there will be more material on Data Streaming from some research articles. In broad terms, some combination of the following topics:

Link Analysis
Mining Data Streams - Frequency and Moment Estimates
Finding Similar Items - Locality Sensitive Hashing
Advertising on the web - Adwords & Online matching
Recommendation Systems - Collaborative Filtering
Dimensionality Reduction - Eigenvalues, PCA, SVD
Clustering - K-Means
Mining Social Networks - Community Detection, Partitioning of Graphs, Dynamic Graph Algorithms
Frequent Itemset
+ some probability+linear algebra as and when required.

Grading Scheme (Tentative):

- Assignments: 3x12%=36%

Please only refer to class notes and the reference material listed on the web-page and/or during lectures for solving assignment problems. Please do not collaborate. Please cite all the references used for solving each of the problems. All assignments need to be submitted electronically using the brightspace system.

Assignment #	Due Date
I	September 27
II	October 29
III	November 29

- Test: 20%: Scheduled during the class time slot on November 1.

- Final Exam: 44% Scheduled by the Exam Services

Note: Final exam will consist of several problems from various topics in the course. Questions will be similar to what you have seen in the assignments. The problems will use the ideas directly from the lectures. Please review the references, class notes, and problems mentioned in the assignments and/or notes. For each topic covered in the class - try to recall the main idea, the primary technique, and how the stated performance bounds were derived.

Schedule for FALL 2024

Sep 04: Introduction + Online Learning

- Introduction to Multiplicative-Weight Update Method
- Arora, Hazan and Kale, The multiplicative weights update method: a meta-algorithm and applications, Theory of Computing 8(1): 121-164, 2012.
- Section 11.1 of notes.

Sep 06:

Sep 11:

Sep 13:

Sep 18:

Sep 20:

Sep 25:

Sep 27:

Oct 02:

Oct 04:

Oct 09:

Oct 11:

Oct 16:

Oct 18:

Oct 30: Solutions to Assignment problems

Nov 01: MID-TERM

Nov 06:

Nov 08:

Nov 13:

Nov 15:

Nov 20:

Nov 22:

Nov 27:

Nov 29:

Dec 04: Review + Solutions to Assignment and Mid-term problems

Previous Term Schedule

- **W1: Introduction + Review: Probability + Linear Algebra**

  - Course Logistics
  - Probability Basics (Sample Space, Random Variable, Linearity of Expectation, Markov's Inequality) - Section 2.1+2.2 in my notes and/or COMP 2804 Textbook.
  - Matrices (Product - Product as sum of rank 1 matrices, RREF, Rank of a Matrix) - Section 4.1 in notes.

- **W2.1: Probability (Geometric Random Variable and Coupon Collector Problem) + Matrices (Eigenvalues)**
  - Probability (Geometric Random Variable - See Section 6.6 of Michiel's Notes; Coupon Collector's Problem - See Exercise 2.28 in my notes and *Coupon's Collector Problem*)
  - Matrices (Eigenvalues - Section 4.2 of my notes)
  - See a quick review

- **W2.2: Matrices**
  - Eigenvalues and Eigenvectors of of A, A^2,...,A^k
  - Markovian Matrices (Definition, Recurrent/Transient States, reducible/irrreducible, aperiodic/periodic)
  - Eignevalues and Eigenvectors of Recurrent, irreducible, aperiodic Markov Chain/Matrix.
  - Perron-Frobenius Theorem
  - *Eigenvalues (1 2)*
  - *Video Lectures and the book of Gilbert Strang*
  - Section 4.7.1 of Notes for Markov Chains

- **W3.1: Markov Chains and Matrices, Balls-and-Bins Problems**
  - Convergence, Steady State of Markov Chains and connection to principal eigenvectors
  - See Exercises 2.35 and 2.36 in Notes.
  - See summary on Balls and Bins

- **W3.2: Pagerank Algorithm**
  - *Page Rank*
  - *Original paper by Brin and Page on PageRank: The anatomy of a large-scale hypertextual web search engine (1998).*
  - *What can you do with a web in your pocket by Brin, Motwani, Page and Winograd.*
  - *Link to the TED talk of Cedric Villani on "What's so sexy about Math" and the description of PageRank.*
  - Section 4.7.2 of Notes
  - See summary on Page Rank

- **W4.1: Bloom Filters and Count Min-Sketch**
  - Bloom Filters
    - Section 9.2 of Notes
    - MMDS Book Section 4.3
    - See summary on Bloom Filters
  - CMS

        - Section 9.1 of Notes
        - See summary on CMS (Note: It has more than what we covered in the class)

- **W4.2: CMS + Estimation of Frequency Moments**
    - Estimating Frequency Moments F_0
        - Section 9.3 of Notes + Chapter 4 of MMDS
        - AMS Article (Approximating frequency moments)
        - Flajolet and Martin's Article on number of distinct elements in a stream

- **W5.1: Analysis of F_0 + DGIM Algorithm for Estimating 1s in a Sliding Window**
    - **Sliding Window:**
        - Section 9.4 of Notes
        - Chapter 4.6 of MMDS.
        - DGIM Article

- **W5.2: Analysis of F_0, Counting 1s in Sliding Window.**
    - Summary on Estimating Frequency Moments
    - Summary on Sliding Window

- **W6.1: Sliding Window, Estimating F_2, Maximum k-coverage problem**
    - **Max-k-coverage** (No notes)
        - Maximizing the Spread of Influence through a Social Network by Kempe, Kleinberg and Tardos
        - Turning Down the Noise in the Blogosphere by El-Arini et al.
        - https://en.wikipedia.org/wiki/Maximum_coverage_problem
        - See Exercises 14.9 and 14.10 in Notes

- **W6.2: Analysis of Max k-coverage, Locality-Sensitive Hashing.**
    - **LSH:**
        - Chapter 3 in MMDS
        - Useful references on LSH
        - STOC98 paper of Indyk-Motwani
        - Sections 8.1, 8.2, 8.3 and 8.6.5 in Notes

- **W7.1: LSH Continued.**
    - Documents, k-shingles, sets, characteristic matrix, signature matrix, bands, expression for f(s).
    - Finding Similar vectors
    - Finding Similar Fingerprints
    - Sections 8.1, 8.2, 8.3 and 8.6.5 in Notes
    - Summary on LSH

- **W7.2: Review - Assignment 1**

- **W8.1: Mid-Term (In class).**

- **W8.2: LSH - Fingerprint + Bipartite Matching Problem**

  - **Online bipartite matching**
    - Chapter on Advertisement on the Web from the MMDS Book (Chapter 8).
    - For a detailed analysis look at Chapter 10 of My notes.

- **W9.1: Balance Algorithm**
  - Proof of 1/2-competitiveness of the online Greedy matching algorithm (see MMDS Book as my notes has the proof using the Linear Programming).
  - Online Advertisement Model = Adwords Problem
  - Balance Algorithm with advertisers
  - Lower Bounds for b-matching
  - How to handle non-uniform bids and budgets
  - Remarks on how to show the upper bound of 1-1/e for the competitive analysis of the Balance Algorithm

- **W9.2: Balance Algorithm (contd.)**
  - Analysis for 2-advertisers
  - Remarks on N-advertisers
  - Non-uniform bids

- **W10.1: Remarks on Balance Algorithm + Eigenvalues of Symmetric Matrices + SVD**

- **W10.2: Singular Value Decomposition with Applications**
  - Low Rank Approximations
  - Recommendation Systems
  - Section 4.5 of notes
  - Chapter 11.3 and 11.4 of MMDS
  - Summary on SVDs

- **W11.1: Low Rank Approximations (contd.) + Clustering**

- **W11.2: Clustering**
  - k-means (Llyod's algorithm, MMDS Section 7.3)
  - k-means++ paper (Arthur and Vassilvitskii, 8th ACM-SIAM Symposium on Discrete algorithms, 2007)
  - Summary on Clustering

- **W12.1: Online Learning**
  - Introduction to Multiplicative-Weight Update Method
  - Arora, Hazan and Kale, The multiplicative weights update method: a meta-algorithm and applications, Theory of Computing 8(1): 121-164, 2012.
  - Section 11.1 of notes.
  - Randomized MWU (without proofs) - see Section 11.2 of notes.
  - Summary on MWU Method

- **W12.2: MWU Method + Problems from Assignment 2**

- **W13.1: Problems/Review + Problems from Assignment 3**

- **W13.2: NO CLASS (Office Hours only)**
- **Final Exam:** See the Exam Schedule for Room Location.

Important Considerations:

Late assignments are not accepted. Assignments submissions are handled electronically and there is no "grace period" with respect to a deadline. Technical problems do not exempt you from this requirement. You are advised to:

- periodically upload your progress (e.g. upload your progress at least daily)
- attempt to submit your final submission at least one hour in advance of the due date and time**.**

Undergraduate Academic Advisor:

The Undergraduate Advisor for the School of Computer Science is available in Room 5302 HP; or by email at scs.ug.advisor@scs.carleton.ca. The undergraduate advisor can assist with information about prerequisites and preclusions, course substitutions/equivalencies, understanding your academic audit and the remaining requirements for graduation. The undergraduate advisor will also refer students to appropriate resources such as the Science Student Success Centre, Learning Support Services and Writing Tutorial Services.

University Policies

Carleton is committed to providing academic accessibility for all individuals. Please review the academic accommodation available to students here: https://students.carleton.ca/course-outline/. We follow all the rules and regulations set by Carleton University, Faculty of Science, and the School of Computer Science regarding accommodating students with any kind of need(s). Please consult with the appropriate authorities to see how you can be accommodated and please follow their procedures. For information about Carleton's academic year, including registration and withdrawal dates, see Carleton's Academic Calendar. Following is a standard list of recommendations that we have been advised to provide you with respect to whom to contact for the appropriate action(s):

Pregnancy Obligation. Please contact your instructor with any requests for academic accommodation during the first two weeks of class, or as soon as possible after the need for accommodation is known to exist. For more details, visit Equity Services.

Religious Obligation. Please contact your instructor with any requests for academic accommodation during the first two weeks of class, or as soon as possible after the need for accommodation is known to exist. For more details, visit Equity Services.

Academic Accommodations for Students with Disabilities If you have a documented disability requiring academic accommodations in this course, please contact the Paul Menton Centre for Students with Disabilities (PMC) at 613-520-6608 or pmc@carleton.ca for a formal evaluation or contact your PMC coordinator to send your instructor your Letter of Accommodation at the beginning of the term. You must also contact the PMC no later than two weeks before the first in-class scheduled test or exam requiring accommodation (if applicable). After requesting accommodation from PMC, meet with your instructor as soon as possible to ensure accommodation arrangements are made. For more details, visit the Paul Menton Centre website.

Survivors of Sexual Violence. As a community, Carleton University is committed to maintaining a positive learning, working and living environment where sexual violence will not be tolerated, and survivors are supported through academic accommodations as per Carleton's Sexual Violence Policy. For more information about the services available at the university and to obtain information about sexual violence and/or support, visit: carleton.ca/sexual-violence-support.

Accommodation for Student Activities. Carleton University recognizes the substantial benefits, both to the individual student and for the university, that result from a student participating in activities beyond the classroom experience. Reasonable accommodation must be provided to students who compete or perform at the national or international level. Please contact your instructor with any requests for academic accommodation during the first two weeks of class, or as soon as possible after the need for accommodation is known to exist. For more details, see the policy.

Student Academic Integrity Policy. Every student should be familiar with the Carleton University student academic integrity policy. A student found in violation of academic integrity standards may be awarded penalties which range from a reprimand to receiving a grade of F in the course or even being expelled from the program or University. Examples of punishable offences include: plagiarism and unauthorized co-operation or collaboration.

Plagiarism. As defined by Senate, "plagiarism is presenting, whether intentional or not, the ideas, expression of ideas or work of others as one's own". Such reported offences will be reviewed by the office of the Dean of Science. More information and standard sanction guidelines can be found here: https://science.carleton.ca/students/academic-integrity/. **Please note that content generated by an unauthorized A.I.-based tool *is* considered plagiarized material.**

Unauthorized Co-operation or Collaboration. Senate policy states that "to ensure fairness and equity in assessment of term work, students shall not co-operate or collaborate in the completion of an academic assignment, in whole or in part, when the instructor has indicated that the assignment is to be completed on an individual basis". **For this course, the following holds:**

- Students are **NOT** allowed to collaborate on assignments. Please avoid using search engines to look for answers etc. Just a word of caution - in theoretically oriented courses, it is important to come up with your own ideas for the proof/an algorithm/a contradiction/etc. Sometimes these are like logical puzzles - if somebody tells you a solution then they are trivial and hard part is to come up with a solution. What we want to learn is how to solve them ourselves.
- Past experience has shown conclusively that those who do not put adequate effort into the assignments do not learn the material and have a probability near 1 of doing poorly on the exams/tests.
- Penalties for academic offences can be found on the ODS webpage: https://science.carleton.ca/academic-integrity/.

Important Dates: For important academic dates and deadlines, refer to Carleton's Academic Calendar.

Announcements: Please attend classes to know any course related announcements.
