Course Name: Scalable Data Science

Course Number: CS-563

Credits: 3-1-0-4

Prerequisite: Data Structures and Algorithms (CS202), Probability, Statistics and Random Processes (IC210), Algorithm Design and Analysis (CS403).

Students intended for: B.Tech.(3rd/4th year)/M.S./Ph.D.

Elective or Core: Elective.

Preamble: Recent advancements of the WWW, IoT, social networks, e-commerce, etc. have generated a large volume of data. Algorithms that help in gaining insights in order to make wiser decisions in large data sets are ubiquitous in several applications. Many naive algorithmic techniques may not be able to cope up with such a large volume of data by: a) running out of memory, and b) having a large running time. This course attempts to address these challenges by a) offering scalable algorithms, b) reducing the size of the data set such that the result of an algorithm on the reduced datasets is very close to its result on the original dataset. Algorithms covered in the course fit well in the intersection between theory and practice, and have guarantees on their accuracy and efficiency, and can be easily implemented.

Learning outcome: Topics covered in the course are fundamental subroutines (or sub-problems) that are required to build large scale systems. After taking this course, students will become familiar with their classical as well as state-of-the-art algorithms. They will have a theoretical understanding of the algorithms along with their practical implementation.

Course Modules:

1. **Dimensionality reduction algorithms:** Johnson-Lindenstrauss Lemma; Random Projections; Spectral Projection, and their applications [5]. (4 hours)

2. **Sketching algorithms for large data stream:** Reservoir sampling; Frequent element detection – Misra Gries algorithm; probabilistic counting – Flajolet and Martin Sketch; Set membership problem – Bloom filters and Cuckoo filters; Frequency estimation– Count Min- Sketching [8, 5]. (7 hours)

3. **Algorithm for large scale search:** Introduction to Locality Sensitive Hashing (LSH) and its variants: LSH for Jaccard Similarity – Minwise Independent Permutations (MinHash) [6] and its recent advancements (b-bit MinHash [14], One Permutation Hashing [15]); LSH for Cosine Similarity – Signed Random Projections (SimHash) [7]; LSH for Euclidean Distance [12]; LSH for Hamming distance [10]. (8 hours)

4. **Application of LSH:** Faster duplicate detection, clustering the web, large scale itemset mining, model compression. (3 hours)

5. **Mining massive graphs and applications:** Algorithms for page rank; community detection; finding overlapping communities and connected components; partitioning of graphs; counting triangles. Learning embedding of nodes with applications in link prediction, node classification. (7 hours)

6. **Clustering algorithms for large data:** Sampling algorithms for k- means clustering – k-means++ [1], scalable k-means++ [2]; spherical k-means clustering [9]; k-mode clustering [11]; spectral clustering [5]. (6 hours)

7. **Miscellaneous Topics:** Learning representation of text – word2vec [16, 13] and images – spectral hashing [17, 5] and its connection with Matrix Factorization; Topic modelling and Topic labelling [4, 3]; Building Recommendation System – a) Collaborative Filtering, b) Content based recommendation. (7 hours)

Similarity Content Declaration with Existing Courses: Following consists of a comparison between the proposed course and existing courses CS-561 (MapReduce and Big Data) and CS-660 (Data Mining and Decision Making).

1. CS-561 is more of a hands-on course which provides a thorough understanding of the MapReduce paradigm, and implementation of various algorithms on Big data platforms. However, scope of the proposed course is different and focuses on developing simple and practice algorithms with provable performance guarantees for several fundamental data science problems.
2. There are some intersections between the CS-660 and the proposed course on topics such as: a) Clustering, b) Principal Component Analysis, c) Association Rules. The intersection is less than 20%. Furthermore, the approach of covering these topics in CS-660 is different as compared to the proposed course. For these topics heuristics are covered in CS-660, while in the proposed course algorithms with provable guarantee on their accuracy and efficiency will be covered.

Proposed by: Rameshwar Pratap **School:** SCEE

Text Books

[i] Anand Rajaraman, Jure Leskovec, and Jeffrey D. Ullman. Mining Massive Datasets. 2014.

[ii] A. Blum, J. Hopcroft and R. Kannan, Foundations of Data Science, Cambridge University Press, 2020.

References

[1] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In Proceedings of the Eighteenth Annual ACM- SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007, pages 1027–1035, 2007.

[2] Bahman Bahmani, Benjamin Moseley, Andrea Vattani, Ravi Kumar, and Sergei Vassilvitskii. Scalable k-means++. PVLDB, 5(7):622–633, 2012.

[3] Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. Automatic labeling of topics with neural embeddings. In COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan, pages 953–963, 2016.

[4] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. the Journal of Machine Learning research, 3:993–1022, 2003.

[5] Avrim Blum, John Hopcroft, and Ravindran Kannan. Foundations of data science, 2015.

[6] Andrei Z. Broder, Moses Charikar, Alan M. Frieze, and Michael Mitzenmacher. Min-wise independent permutations (extended abstract). In Proceedings of the Thirtieth Annual ACM Symposium on the Theory of Computing, Dallas, Texas, USA, May 23-26, 1998, pages 327–336, 1998.

[7] Moses Charikar. Similarity estimation techniques from rounding algorithms. In Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada, pages 380– 388, 2002.

[8] Graham Cormode. Sketch techniques for approximate query processing. In Synopses for Approximate Query Processing: Samples, Histograms, Wavelets and Sketches, Foundations and Trends in Databases. NOW publishers, 2011.

[9] Inderjit S. Dhillon and Dharmendra S. Modha. Concept decompositions for large sparse text data using clustering. Machine Learning, 42(1/2):143–175, 2001.

[10] Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Similarity search in high dimensions via hashing. In VLDB'99, Proceedings of 25th International Conference on Very Large Data Bases, September 7-10, 1999, Edinburgh, Scotland, UK, pages 518–529, 1999.

[11] Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 2(3):283–304, Sep 1998.

[12] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In Proceedings of the Thirtieth Annual ACM Symposium on the Theory of Computing, Dallas, Texas, USA, May 23-26, 1998, pages 604–613, 1998.

[13] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, pages 2177–2185, 2014.

[14] Ping Li and Arnd Christian König. b-bit minwise hashing. In Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010, pages 671–680, 2010.

[15] Ping Li, Art B. Owen, and Cun-Hui Zhang. One permutation hashing. In Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States, pages 3122–3130, 2012.

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States, pages 3111–3119, 2013.

[17] Yair Weiss, Antonio Torralba, and Rob Fergus. Spectral hashing. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, Advances in Neural Information Processing Systems 21, pages 1753–1760. Curran Associates, Inc., 2009.