Big Data Analytics

Program: **Masters of Information Technology - Artificial Intelligence**
Course Code:
Course Type: **TH+PR**
Year: I
Semester:
Credit Hour: **3**
Contact Hours: 45

1 Course Objectives

By completing this course, you will be able to:

1. understand concepts and challenges in Big Data Analytics
2. Implement Big Data code in Apache Spark (in PySpark)
3. Understand and implement foundational algorithms in Big Data Analytics related to similarity search, mining of frequent item sets, clustering and recommender systems.
4. evaluate and apply appropriate principles, techniques and theories to large scale data problems.

2 Course Description

This course will cover concepts and tools used to analyze Big Data. Specifically, this course will introduce students to industry-standard tools/frameworks to work on large-scale datasets as well as provide graduate-level understanding of the algorithms used by these tools to scale to these datasets and be robust and fault-tolerant via a distributed storage system. The course will introduce students to important mathematical concepts to gain a understanding of these tools, useful for making informed decisions for real-world problems as well as provide foundation for further research in the topic.

The Lab sessions will provide hands-on engagement with how to incorporate these algorithms and tools into a real-world pipeline. Additionally, students will implement some of these algorithms from scratch and execute them on sample datasets. On the other hand, industry-standard tools such as Apache Spark will be used to analyze and apply data-mining to real-world datasets.

3 Course Outcomes

By the end of the successful completion of the course, you will be able to :

1. Explain how to setup a fault-tolerant & distributed large-dataset systems using Apache Spark
2. Use Apache Spark to perform data preprocessing
3. Explain mathematical concepts of Points, Space and Distance and it’s application in similarity search & clustering

4. Explain various algorithms for finding frequent itemsets
5. Explain various algorithms for clustering and their pros and cons
6. Explain various algorithms for Recommender systems and their limitations

4 Course Contents

Introduction to Big Data Analytics [3 hrs] Data Mining Concepts - Bonferoni's Principle, Power Laws, Hash functions, Indexes.

Big Data Frameworks [6 hrs] Distributed File Storage Systems. MapReduce: Map function, Reduction Function. Joins: Multiway Joins, Star Joins. Apache Spark programming (Python and PySpark), Spark Stack and Core Spark. Resilient Distributed Dataset (RDDs), Dataframes and Spark SQL.

Finding Similar Items [9 hrs] Locality-sensitive hashing, Similarity Metrics, Distance function, Shingling, MinHashing, Indexes.

Finding Frequent Items [9 hrs] Market-Basket Model. A-priori algorithm. Approximate Algorithms for large datasets - PCY Algorithm, Multistage Algorithm, Multihashing Algorithm, SON Algorithm.

Clustering [9 hrs] Points, Spaces and Distance. Clustering strategies - Hierarchical vs. Point Assignment. Euclidean vs. Non-Euclidean space. Curse of Dimensionality. Hierarchical Clustering, K-Means and it's variants: BFR Algorithm, CURE Algorithm. Clustering in Non-Euclidean Space: GRGPF ALgorithm.

Recommendation Systems [9 hrs] Content-based vs. Collaborative filtering. Long-Tailed Distribution. Utility Matrix. Content-based: Item Profiles, User Profiles, Feature Extraction. Collaborative filtering.

5 Textbook(s)

Leskovec et al. [2020] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. Mining of massive data sets. Cambridge university press, 3rd Edition, 2020. Available online at http://www.mmds.org.

6 Practicals

Lab sessions may include, but not limited to,

1. Worked-out Examples and Exercises from Jonathan Rioux. PySpark in Action: Python data analysis at scale. Manning Publications Co., 2022. Companion source code https://github.com/jonesberg/DataAnalysisWithPythonAndPySpark

   (a) Given a large text file, count the number of words that start with each letter. You may use Spark functions explode() and split().

   (b) Apply Spark Dataframe API such as groupBy(), agg(), count(), map(), select() to perform meaningful operations on a instructor-chosen dataset

2. Worked-out Examples from Stanford CS246 http://web.stanford.edu/class/cs246/
   (a) Apply Fast Frequent Pattern Mining with the FP-Growth algorithm on the Instacart Orders dataset (or any similar datasets). Use Spark Dataframe API for preprocessing. Explore MLib, Apache Spark ML library, and use it's implementation of FP-Growth algorithm.
   (b) Apply Collaborative Filtering on the MovieLens dataset. Explore, MLib and use the collaborative filtering algorithm available there.

3. Working Prototype for a Movie Recommendation system: If the students have earlier experience in web development, the students may be assigned to create a Web UI for Movie Recommendation system where user can search for movies similar to the search query provided by the user. Students may use the Apache Spark MLib library for building the core recommendation engine and see the algorithms studied in this course in action.

4. Implement Algorithms from scratch to reinforce understandings (in any one of the language students are familiar with)
   (a) Implement MinHash Algorithm
   Sample Implementation https://github.com/go2starr/lshhdc
   (b) Implement K-means & it's variants

7 Recommended Readings

The course instructor may assign papers for reading and presentation from Tier 1 Conferences relevant to Big Data Analytics such as ACM SIGKDD (Int. Conf. on Knowledge Discovery and Data Mining), WSDM (Int. Conf. on Web Search and Data Mining), AISTATS (Int. Conf. on AI and Statistics) etc. to name a few. This may include, but not limited to, example applications of Big Data Analytics or novel methods that the students may be able to grasp ideas based on foundations build in this and earlier courses, or build upon ideas that have been taught during the course. Few suggestions have been listed below.

1. Zheng, Y., Liu, F., & Hsieh, H. P. (2013, August). U-air: When urban air quality inference meets big data. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1436-1444).

2. Brijs, T., Swinnen, G., Vanhoof, K., & Wets, G. (1999, August). Using association rules for product assortment decisions: A case study. In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 254-260).

Additionally, position papers, articles and blogs may also be assigned as readings that provide high-level summary, opinion etc. based on instructor's discretion. Few examples have been listed below:

1. Cate, Fred H. (2014). The Big Data Debate. Science 346(6211): 818-818.

2. Manovich, L. (2011). Trending: The promises and the challenges of big social data. Debates in the digital humanities, 2(1), 460-475.

References

Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of massive data sets*. Cambridge university press, 2020.

Jonathan Rioux. *PySpark in Action: Python data analysis at scale MEAP V06*. Manning Publications Co., 2022. ISBN 9781617297205; 1617297208.

Hanghang Tonge Jiawei Han, Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, London, 2022.

Holden Karau, Andy Konwinski, Patrick Wendell Wendell, and Matei Zaharia. *Learning Spark: Lightning-Fast Big Data Analysis*. O'Reilly Media, 1st edition, 2015. ISBN 1449358624; 9781449358624.

Jean-Georges Perrin. *Spark in Action, Second Edition*. Manning Publications, 2nd edition, 2020. ISBN 1617295523; 9781617295522.