Program Objectives

Applications of Big Data & AI transcend Industries. Use of predictive analytics pervades diverse disciplines such as marketing and sales, sports, molecular biology, drug-designing, waste management, finance, healthcare, knowledge products and the list is very long. Smart cities, for example, are the melting pot where variety of big data technologies mesh with one another to transform a city into a semi-intelligent being. In Marketing and Sales, for example, Big Data & AI are fast emerging as a potent tool to gain deeper insights into Customer behavior and thereby act as a strong driver in spurring innovation. In manufacturing, operations managers are employing advanced analytics on historical process data to identify patterns and relationships among discrete process steps and inputs, and then optimize the factors that prove to have the greatest effect on yield.

There are three principal segments to the program: One the Analytics part, second the technological part and third the contemporary Generative AI and Designing LLM products. The analytics part is about Machine Learning Algorithms and implementing them, the technological part is about learning to use spark on Hadoop/NoSQL Databases and work on Streaming Data Analytics using Apache Kafka. As NoSQL databases are important big data storage systems, we cover them in our course. Deep Learning technology is applied in such fields as Healthcare, Personalized Marketing, Financial Fraud Detection, Facial Recognition, Recommendation Systems, Agriculture and others. Generative AI and Large Language Models (LLMs) have a wide range of applications across various industries. As for example:

a) Text Generation: Generating human-like text for content creation, storytelling, and dialogue; Summarizing long documents or conversations; Translating text between languages; Answering questions and providing informative responses.
b) Creative Applications: Generating images, music, and other media content; Assisting with ideation and brainstorming for creative projects; Designing 3D models and virtual environments.
c) Conversational AI: Powering chatbots and virtual assistants for customer service and support; Engaging in open-ended conversations and providing personalized responses; Automating repetitive tasks like scheduling and email management.
d) Analytical Applications: Analyzing large datasets to identify trends and insights; Generating reports, visualizations, and summaries from data; Providing recommendations and predictions based on data
e) Educational Applications: Tutoring and teaching by answering questions and providing explanations; Generating practice problems and grading student work; Providing personalized learning experiences
f) Research and Development: Accelerating scientific discoveries by generating hypotheses and experiments; Assisting with literature reviews and summarizing research papers; Aiding in the development of new products and services

The key difference between generative AI and LLMs is that generative AI encompasses a broader range of models that can generate diverse types of content, while LLMs are specialized in understanding and generating human-like text. The choice between the two depends on the specific application, available data, and resources.

At the end of this course, given a large dataset from any domain, a participant should:

1. Be able to clean, transform and visualize a dataset to gain deeper insights and make it ready for further analysis

2. Be able to engineer features and select a subset of appropriate machine learning or deep learning algorithms that could be applied to get the desired predictive results.
3. Make situation and context specific performance assessment as also interpret predictive models and explain them to clients
4. Apply the knowledge of image processing, image analysis, sensor-data analysis and language modeling to a wide array of disciplines such as health, process control, navigation and others.
5. Design/create knowledge-products using Large Language Models

Further:

6. Should be able to himself install, setup and configure and experiment with a complete hadoop and Kafka ecosystems (that include hive, spark, zookeeper, flume and other important layers).
7. Should be able to install, configure and be sufficiently familiar with the variety of NoSQL databases and decide for himself which one to use, when and how
8. Should be able to install fully functional various deep learning platforms including Tensorflow and PyTorch.
9. Should be able to install Web-User Interface(s) incorporating RAG for managing LLMs.

~~10.~~ This course is project (lab) oriented: All tools, data and platforms including deep-learning platforms, hadoop-ecosystem, Kafka-streaming technologies and LLMs necessary for experimenting are either provided to the participants in advance or we help participants to install them on their laptops. There is a heavy emphasis on open-source technologies universally used almost throughout the industry. Each participant, at the beginning of the course, receives Virtual Machines (VMs) fully equipped with many software platforms, tools, packages and data to work on.

11. Our faculty have experience with several Industrial projects. Our students regularly execute projects on Kaggle—in fact, executing projects on Kaggle, creating a GitHub repository of projects and hosting LLM based web-interfaces on HuggingFace are a must and during the course several projects are implemented on all these platforms.

Who should attend

1. Data being ubiquitous, the program cuts-across job or academic profiles. The techniques taught are generic in nature. These will be valuable to anyone who wishes to analyze data to advance his/her knowledge. Specifically, the course will be useful for:
2. *Executives*
3. Ambitious Executives (from Private/Public sectors) looking forward to sharpening their skills and making sense of data in order to innovate and add more value to their organization and to society.
4. *Academicians*
5. Lecturers and Professors for extending the horizon of their knowledge through deepening their research skills.
6. *Data Scientists/ Developers*
7. Techniques taught to them will have applications in a broad range of disciplines.
8. *Healthcare professionals*
9. Healthcare professionals stand to immensely benefit from the extensive coverage of Deep Learning and LLM technologies and how these are applied in medical case studies.
10. *Students/Research Scholars*
11. IInd year students currently enrolled in Engineering / MBBS / PGDM / MBA or any graduate or post-graduate program who have had an introductory course in statistics.

These students can look forward to better placement opportunities with an added skill set.

Module wise details

The complete course is divided into five distinct modules. Module 1.1 is about Machine Learning Algorithms. In this module, we use a variety of python based libraries. Module 1.2 is about Hadoop and Kafka eco-system: we learn to work on Hadoop and its layers; perform data extraction, build data pipelines as also push it into analytics engine. Analyzing streaming data is a major subject in its own right: in this respect we experiment with Apache Kafka and related technologies. Module 1.3 relates to NoSQL and Graph databases. The new millennium and the explosion of web content has marked a new era for database management systems. A whole generation of new specialized databases have emerged, all categorized under the name of NoSQL databases with focus on "task-oriented" database management system, selecting the right tool for the job depending upon its characteristics, nature and requirements. We cover, in depth, some often used NoSQL databases. Module 1.4 pertains to the exploding field of deep learning and AI. Deep Learning distinguishes itself from classical machine learning by its ability to work with unstructured data, like text and images, without the need for extensive pre-processing. Deep learning models, consisting of multiple layers of interconnected nodes, automatically learn and improve from data, enabling them to recognize patterns, classify phenomena, and make predictions with high accuracy. In this part we cover deep-learning technologies using very popular libraries tensorflow and PyTorch. Module 1.5 is about Generative AI. Generative AI refers to artificial intelligence systems that can generate new content such as text, images, audio, or video based on patterns learned from training data. Generative AI models learn the underlying structure and patterns in their training data, which could be large datasets of images, text, audio, or other modalities. We learn how these models work and perform projects based on Large Language Models.

Pedagogy

We strongly believe that a course in data analytics can only be practice-based rather than theory based. We also believe that a practice-based course requires constant interaction with the teacher during lecture hours in real time. As it is a distance online course, the teaching pedagogy is like this: First the algorithm (or theory part) is conceptually explained without getting into mathematics and then a project is undertaken to implement the techniques. Datasets for implementation are made available in advance and so also a copy of code (or hints on it) that we need to execute. The code is numbered and copiously commented so that long after the lecture has finished, students can go back through the code/comments and refresh their knowledge. During the lecture, we execute this code (or prompt students to fill in the gaps), line-by-line and explain the steps. At his end, the student executes the required code on his laptop. Consequently, results are available at our end as also with the Students immediately. In short, both the teacher and students are working on their respective laptops simultaneously; students solve their problems and ask any questions to clarify. The whole experience is just as if everyone is sitting in a laboratory and working together.

Top

Machine Learning is about data cleaning, data transformation, feature engineering and predictive analytics. It is about knowledge discovery in datasets--structured or unstructured--searching through large volumes of raw data to find useful information-patterns. Data Scientists and decision makers can use this information for new sources of advantages and differentiation or for developing new business models. Broadly speaking the module objectives are three-fold:

- Generate familiarity with Big Data, Data Visualization and Data Mining algorithms: In generating this familiarity there is special emphasis on conceptual understanding of techniques rather than on mathematics. Analytics is a creative process, and students are encouraged to be creative.

- Develop skills to set up predictive models with numerous types of disparate data sets. This is intended to bring home the point that predictive analytics offers a generic set of tools that can be applied on different types of datasets, no matter what be the discipline or the Industry.
- Think differently: Expose students through projects to novel ways of applying Big Data technologies among shifting business models.

##############