Undergraduate 499/500 Projects

BAM Lab Projects for 2019-2020

Project 1: Unsupervised Clustering of Data Stream Using Apache Spark

Supervisors: Dr. Farhana Zulkernine, BAM Lab.

Objective

As part of an ongoing academic-IBM collaboration research project, we are developing a data analytics infrastructure for multilevel streaming analytics. This project will help building an adapter between IBM Streams and Apache Spark, a popular open source data analytic tool. Data from IBM Streams will have to be passed to Apache Spark and Spark libraries should be used to analyze the streaming data to enable real time learning.

Learning Outcomes

Learn about cutting edge streaming data stream analytic tools such as IBM Streams and Apache Spark, design, implement and validate a mutli-stream multi-modal data analytic system.

Expertise Needed

Knowledge about Apache Spark distributed systems and programming expertise.

Description

The ever-increasing volume and highly irregular nature of data rates pose new challenges to data stream processing systems. One such challenging but important task is how to accurately ingest, process and analyze data streams from various sources. Multilevel data stream processing will allow fast processing of the main data stream using IBM Streams and extracting important information to be sent to Apache Spark to be further processed at the 2nd level using more complex deep learning algorithms. This way the main pipeline will not get clogged by long running algorithms and more insights can be extracted from blocks of data. At the core of this system is the integration of IBM Stream and Apache Spark Streaming tools and implementation of stream clustering algorithms to extract insights from the data.

Data stream sources

Twitter Streaming API, Satori live data channels, and propriety data streams via WebSockets.

Deliverables

A prototype application, source code, and project report.

Support

The student will work closely with Sazia Mahfuz and Dr. Zulkernine and will have access to one-one support and various software stacks necessary for the project.

Project 2: Intelligent Interactive Robotic System for Providing Student Support at Queen's University (multiple sub-projects: conversation tone, emotion and image analytics and generating intelligent responses using deep learning models)

Supervisors: Dr. Farhana Zulkernine, BAM Lab.

Objective

Create an intelligent interactive system equipped with voice communication and image recognition system and able to process data to effectively communicate with a customer. It will build on prior systems and the emphasis will be on the intelligent analytic side to identify a person from video, analyze the emotion from the tone, and find good answers to the questions.

Learning Outcome

Explore existing cutting edge technology for voice, text and image processing (hardware and software) and mine Queen's web data sources to effectively communicate with a student to respond to queries effectively using machine learning, AI and decision techniques.

Expertise Needed

Understanding of AI and data mining techniques, must have taken CISC/CMPE452 or COGS400, experience in developing neural network models, interest in exploring cutting edge voice, video and text analytic techniques.

Description

Smart communication systems are leading this cognitive era. We have interactive chatbots to communicate with using text messages, Google API, Watson and Amazon Alexa to answer user queries and execute commands. This project will explore ways to improve the existing technologies by adding image and text based search and decision techniques to mine information from Queen's web data and other static data sources to identify the student user asking a question and cater to his or her needs. The student will have to combine voice, text and image processing techniques to identify the emotion, tone, identify the student user, understand natural language questions and generate voice or visual responses such as a map. The student can be creative and add a mini projector to display the data or just show it on computer screen. The program should be accessible on the cloud and should use Google API or Watson for the voice communication mechanism.

Data sources

Publicly available Queen's university data.

Deliverables

An intelligent interactive system to understand student queries and generate effective responses, and a project report.

Support

The students will work closely with Yuhao Chen (MSc student) and Dr. Zulkernine. Additional expertise can be sought if advice is sought on hardware or integration of robotic systems (if the student is enthusiastic).

Project 3: Attention based Image Object Detection and Segmentation

Supervisors: Dr. Farhana Zulkernine, BAM Lab.

Objective

Develop and test an image object detector.

Learning Outcome

Learn about image object recognition models, the inner workings of object detectors and deep neural networks as well as the process of testing and refining neural networks.

Expertise Needed

Expertise in programming neural networks, must have taken CISC/CMPE452 or COGS400, an understanding of image processing.

Description

With the advent of large visual computing datasets and advances in visual computing in the machine learning industry, object detection and segmentation continues to grow as an interesting field in computer science. The object detection problem is defined as detecting instances of semantic objects of a certain class (human, building, car, etc.) in digital images and video. In neural networks, this is often done by using a box region to define the object and using the neural network to propose these boxes. This can be done in two methods, through two phase object detection or single-phase object detection. Two phase object detection methods detect regions of interest first and then classify those regions. Single phase object detection methods detect and classify regions of interest at the same time. Segmentation focuses on extracting the boundary of objects. This a growing field and research in this field is both extremely important and of increasing interest.

Data sources

The students can use open source image data sources and available codes and extend that.

Deliverables

An image object detection model, validation of the model, and a project report.

Support

The students will work closely with MSc student, Mohammad Gasmallah, and Dr. Zulkernine extending his prior work on image object detection.

Project 4: Autonomous Vehicular Data Analytics

Supervisors: Dr. Farhana Zulkernine, BAM Lab.

Objective

Develop deep learning models to understand driving patterns of smart vehicles.

Learning Outcome

Learn about developing deep learning models for IoT data analytics from smart vehicles.

Expertise Needed

Expertise in programming neural networks, must have taken CISC/CMPE452 or COGS400.

Description

Smart vehicles are equipped with numerous sensors that collect real time streaming data to ensure vehicle and driver safety. To have autonomous or self-driving vehicle in the near future we need more research to analyze the data fast and combine data from not only the vehicle but from traffic, road and weather sources to create a safe driving strategy. This project will explore some of the data collected from smart vehicles and address a problem such as driver identification, identifying safe driving styles, learning driving strategy based on road and traffic conditions etc.

Data sources

The students can use open source vehicular data and available codes and extend that.

Deliverables

A deep learning model that address one of the above mentioned problems, validation of the model, and a project report.

Support

The students will work closely with PhD student, Priyanka Trivedi, and Dr. Zulkernine.

Project 5: A Hospital Surge Predictor

Supervisors: Dr. Farhana Zulkernine, BAM Lab.

Objective

Build a predictive model to forecast surge at the hospital emergency department based on real time patient registration data collected from multiple hospitals in Ontario. The work will be done in close collaboration with KFL&A Public Health, Kingston.

Learning Outcome

Design, build and validate predictive analytics algorithms to forecast surge at hospital emergency department based on hospital emergency patient registration data including location, time of year, weather condition, and other news and geographical data.

Expertise Needed

Expertise in databases, data analytics, machine learning and software development.

Description

Depending on the weather condition and break outs of disease conditions due to various reasons, hospitals face a surge condition at the emergency department which can be difficult to handle due to limitations in staff and hospital resources. A predictive model can help alleviate the problem by allowing the hospitals to have warnings in advance to prepare for the surge and if necessary, to redirect patients to other care facilities to distribute the load. KFL&A developed a real time streaming data collection system that currently gathers data from hundreds of hospitals across Ontario. They also have real time data monitoring system that apply descriptive analytics techniques to notify increasing patient registrations crossing predefined thresholds. However, currently no predictive model exists to predict surge 2-3 days or months in advance. We need a predictive model to be built using the data collected over the last 15 years to accurately predict surge at the hospital emergency departments.

Deliverables

A prototype predictive model with validation outcomes measured in precision, recall, sensitivity and specificity, application source code, and project report.

Support

The student will work closely with Dr. Hasan Zafari (Postdoctoral Fellow), Dr. Zulkernine, and KFL&A Public Health personnel. Confidentiality agreements have to signed to access the data and the student may have to work occasionally at the KFL&A Public Health office (across St. Lawrence College on Portsmouth St.).

Project 6: Automatic Case Identification for COPD

Supervisors: Dr. Farhana Zulkernine, BAM Lab.

Objective

As part of an ongoing academic-industry collaboration research project, we are developing a data analytics pipeline to examine existing datasets to identify key features that could be useful in the diagnosing cases of Chronic Obstructive Pulmonary Disease (COPD) in primary care settings.

Learning Outcome

Design, build and validate machine learning algorithms to identify diseases from medical data.

Expertise Needed

Expertise in programming neural networks, must have taken Data analytics, CISC/CMPE452 or COGS400.

Description

The ever-increasing volume of medical data in the form of electronic medical records (EMR) are of great value in primary care system. Patient medical records contain vast amount of information regarding patient conditions that could be used to help screening patients and helping in automatically case identification using data mining and machine learning algorithms. Challenges include small data size considering the fact that many diseases affect only a small subset of the population and hence unbalanced data with errors, missing data, and ambiguity in textual reports.

The goal of this project is to choose and implement analytical methods and approaches for integrating, processing, and interpreting healthcare data for COPD case detection with the highest predictive power. To this end, we will use the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) data as well as the Canadian Institute for Health Information (CIHI) data. These data will be processed and converted as needed. That is, Natural Language Processing (NLP) or Information Extraction (IE) tools and algorithms will be applied to convert unstructured free text part of the data to a structured format. Subsequently, to identify key features that are informative in predicting COPD, feature engineering and selection will be performed. At the core of this system, machine learning algorithms will be applied to create a supervised predictive model. This is a classification algorithm that will accept the features and produce the prediction (COPD case or not).

The performance of these algorithms will be evaluated in terms of sensitivity and specificity of detection. For this purpose, a subset of patients' data along with their labels (case or control) will be reserved and used in the evaluation phase. Students will be required to sign a confidentiality agreement to access the anonymized medical data.

Deliverables

A prototype application, source code, and project report.

Support

The student will work closely with Dr. Hasan Zafari (Postdoctoral Fellow) and Dr. Farhana Zulkernine, with guidance from KFL&A Public Health and CPCSSN as needed for domain specific information. The student will have access to one-one support and various software stacks necessary for the project.

Project 7: Natural Language Question Classification and Validation Using Machine Learning

Supervisors: Dr. Farhana Zulkernine, BAM Lab.

Objective

Develop a Machine Learning (ML) model to train a classification algorithm to identify the Natural Language Question (NLQ) category. Building an SQL query validation and optimization workflow against a Relational DataBase (RDB).

Learning Outcome

Learn about Natural Language Processing (NLP) tools and techniques in addition to some ML algorithms through hands on applications. Learning how to work with relational databases and

SQL. Students will also build strong problem-solving skills during the process of configuring the process of validating SQL queries.

Expertise Needed

Familiarity with NLP tools and classification ML Algorithms (MLA). Familiarity with DBs, SQL, MySQL server or IBM I2B2 or similar DB tools.

Description

MLAs have a wide range of applications in dimensionality reduction, clustering, classification, and multilinear subspace learning. In an NLP, MLA is used to extract NLQ patterns to improve the response time and narrow down predicted sets of SQL templates into a temporary cache memory. Classification MLAs could be either rule based or hybrid approaches for features identification and selection or rules processes. Classification algorithms could be bundled in a library and integrated with query and analytics systems. Main ML properties include scalability, distributed execution and lightweight algorithms.

The query validation module is performed right after receiving the SQL query and before its execution against the RDB. The purpose of this module is to test query correctness percentage that will be conveyed to the user as a confidence ratio. This could be done through any of the machine learning or neural network learning approaches for SQL ranking and classification based on a weighting scheme or an error/correctness rate.

Data sources

The students will receive a TPC benchmark database to build the model with, test it and validate it. GitHub available codes can be used and extend.

Deliverables

A ML model that classifies NLQ categories, validation of the model, and a project report.

Support

The students will work closely with PhD student, Ftoon Kedwan, and Dr. Zulkernine.

Project 8: Optimal Information Retrieval Model Design Using Distributed Storage Systems

Supervisors: Dr. Farhana Zulkernine, BAM Lab.

Objective

Building an optimal distributed and weighted Graph DataBase (DB) model to store and query Big Data using algorithmic community detection and distance functions.

Learning Outcome

Learn to work with Graph DBs and their variate SQL-based query languages in addition to some basic Machine Learning Algorithms (MLA).

Expertise Needed

Familiarity with Graph DBs, Neo4j or Pregel or similar Graph DB tools, in addition to basic MLA background.

Description

The industry needs a storage system that stores and queries tens of thousands of entities and their descriptions, and retrieve them using Knowledge Representation Languages, such as Graph DBs. A variety of storage systems exist today but they haven't been evaluated based on their storage distributability performance, scalability, robustness, security, and functionality.

Graph DBs traversing and computations are computationally intensive and requires a good amount of CPU cycles. Hence, this project aims at using some basic MLAs related to weighted nodes relationships and distance functions to build the best Big Data distributed storage model that would retrieve data using the least CPU cycles possible. The graphs in Neo4j are mapped to the OS memory so that the indexes and cached data can be traversed/searched from the memory itself and do not require any expensive I/O operation on disks. Anything that cannot be kept in the OS memory will be flushed out to disks and will be required to read and load back into RAM on subsequent user's request. The lower the I/O requests the better the response time. Neo4j supports an in-memory location or data center specific page cache, other than the traditional JVM-based caching strategies. Stoppable and resumable hot and point-in-time backups are supported as well, while graph DB is still running.

Graph DBs are relatively new, they are getting popular as they facilitate heterogeneous linked data analytics, representation and reasoning with powerful query language and visualization support. Graph DBs with index-free adjacency can search BigData connections in real-time speed with constant performance regardless of the DBs volume or complexity. Some of the storage systems have supporting data analytic libraries for domain-specific data extraction and inferencing. Information processing and retrieval and knowledge representation can be enhanced using inductive learning techniques including symbolic ID3 learning, genetic algorithms, and simulated annealing.

The mainstream industry is focusing on expressiveness and data analytics support in distributed storage systems such as Neo4j, Google Pregel, OrientDB, or Apache Spark which are leveraged with in-memory graph processing, besides Gradoop framework that is used to analyze collections of Graph KBs. The Hadoop Distributed File System (HDFS) enables the underlying storage of Hadoop clusters. It divides the data into smaller parts and distributes them across servers. GraphBase, Apache TinkerGraph, and Titan are also graph DBs that are used widely to explore linked knowledge for searching and descriptive analytics with NoSQL stores such as HBase and Cassandra, and document stores such as MongoDB. Distributed cloud computing architecture is also used to analyze big data due to their scalable storage capacity.

Data sources

The students will receive a TPC benchmark database to build the model with, test it and validate it. GitHub available codes can be used and extend.

Deliverables

Build the information retrieval model, validate it, and document it in a project report.

Support

The students will work closely with PhD student, Ftoon Kedwan, and Dr. Zulkernine.

Project 9: Query Disambiguation and Optimization Model Design

Supervisors: Dr. Farhana Zulkernine, BAM Lab.

Objective

Building a domain independent SQL ambiguity and uncertainty resolution system through fuzzy constraints to ensure data retrieval integrity and reliability. Then, the resulted query will be optimized for better response time and accuracy.

Learning Outcome

Students will learn how to work with Relational DatBase (RDB) functions and the DB query language SQL. Students will experiment with some machine learning or neural network algorithms. Students will apply methods of SQL query optimization to the most optimal SQL form possible

Expertise Needed

Familiarity with RDB, SQL and some background on machine learning or neural network tools and techniques. Familiarity with MySQL Server or IBM I2B2 or similar DB tools.

Description

SQL parameters disambiguation is an intermediate analysis process and is done through circumstantial and relativity analysis. When the system cannot execute the SQL due to some ambiguity, it asks the user for further input in case of the presence of more than 1 match for a particular SQL clause or if the SQL query contains unrecognizable or unknown expressions. Even if the system proceeds with the presence of any of these mistakes, the system will produce incorrect answers. Engaging the end user is solely to clarify a certain ambiguity in the input by choosing from a list of suggestions of similar words (synonyms) present in the lexicon. In the worst-case scenario, the system should ask the user to try entering a different SQL query or redefine any of its clauses.

The main milestone of this project is solving the vague and imprecise SQL clauses parameters through fuzzy DBs that store fuzzy attribute values and fuzzy truth values. Vague SQL parameters can be imprecise expressions such as “Young” instead of “10 years old” and “Very” or “Almost” instead of “%75” or expressed in time-stamped forms with prepositions (i.e. on, during, since) which can be identified and solved through a temporal system architecture for extracting, representing and reasoning temporal information such as TimeText. Another method is deploying a temporal analyzer component (i.e. CliniDAL) that employs a 2-layer rule-based method to interpret the temporal expressions of the query, whether they are absolute times or relative times/events. The Temporal Analyzer automatically finds and maps those expressions to their corresponding temporal entities of the underlying data elements of the DB’s different data design models. Another option is adopting a temporal tagger (i.e. HeidelTime) that uses a hybrid rule-based and ML approach for extracting and classifying temporal expressions.

In regard to SQL query optimization, the developed model should enforce the best practices of SQL query optimization. Such practices include, but not limited to, the following:

Properly indexing all SQL parameters in the appropriate SQL clauses, avoiding functions in parameters, avoiding wildcards (%) at the beginning of a parameter, omitting unnecessary columns in the SELECT clause, replacing outer join with inner join if possible, only using DISTINCT and UNION if necessary, adding ORDER BY clause if the resulted data should be sorted.

For the DB search mechanism, it can be optimized by indexing data and then using reinforcement learning techniques to adjust relationships' strengths and weights by propagating rewards through a graph.

Supervisors: Dr. Farhana Zulkernine, BAM Lab.

Data sources

The students will receive a TPC benchmark database to build the model with, test it and validate it. GitHub available codes can be used and extend.

Deliverables

Build the SQL query disambiguation and optimization model, validate it, and document it in a project report.

Support

The students will work closely with PhD student, Ftoon Kedwan, and Dr. Zulkernine.

Project 10: Identifying PTSD by processing physician's narrative notes

Objective

Develop a supervised case detection model based on deep learning that will use physician's narrative text data in Electronic Medical Records (EMR) to determine if a patient has Post-Traumatic Stress Disorder (PTSD) or not.

Description

PTSD is a mental disorder that happen as a result of a traumatic event like car accident, war, earthquake etc. Identifying PTSD is challenging because of variations in the symptoms for different patients, and misdiagnosis due to symptoms being shared with other conditions.
The growing amount and availability of EMR data present new opportunities for discovering new knowledge about diseases. Therefore, the objective of this project is to determine if a patient has PTSD or not using unstructured text notes of EMR which the physicians enter during patients' visits. This project will employ deep learning methods to classify each patient to PTSD positive or PTSD negative. To this end, physicians note text data will be processed to extract informative features out of them. Then these features will be fed into a supervised classification algorithm. This will result in a model that would be able to identify PTSD caseness based on patient's note. One challenge is that free text narrative notes are littered with spelling mistakes, acronyms, extra whitespaces. Also, there are many repeated sentences in the narrative notes that may pose a problem in the process.
The performance of these algorithms is evaluated in terms of accuracy of classification. For this purpose, a subset of patient's data along with their class labels will be reserved and used in evaluation phase.

Deliverables

A prototype application, source code, and project report.

Support

The student will work closely with Dr. Hasan Zafari (Postdoctoral Fellow) and Dr. Farhana Zulkernine and will have access to one-one support and various software stacks necessary for the project.

Project 11: Classifying patients based on their medical narrative notes

Objective

As part of an ongoing academic-industry collaboration research project, we are developing a weakly supervised document classification model that would be able to determine if a patient belongs to military/veteran group, is a member of military/veteran personnel, or a civilian patient based on its physician note.

Description

With the development of information technology Electronic Medical Record (EMR) has been popularized in which patient record including texts and other digital information are recorded. The Manitoba Primary Care Research Network (MaPCReN) data is a EMR repository that reviewed more than two million encounter records representing clinical information of about 56,000 patients. These patients belong to different classes including (i) military members and veterans, (ii) family of military and veterans, and (iii) civilian community.

The problem is that, despite being useful for better understanding the symptoms and the quality of patient care for many illnesses, this class information of patients is not recorded in the data. Therefore, in this project we focus on determining the class label for the patients in the MaPCReN data based on the unstructured text notes which the physicians enter during patients' visits.

As there is no labeled data for this problem, a semi-supervised classification method will be applied on this problem. To this end, the news and other documents that are related to military/veteran subject will be processed to identify a set of keywords that could be used as seeds to find a set of notes that are military patient records with high confidence. Then, these documents will be used as training data in a supervised classification model.

The performance of these algorithms is evaluated in terms of accuracy of classification. For this purpose, a subset of labeled data along that acquired in the previous step will be reserved and used in the evaluation phase.

Deliverables

A prototype application, source code, and project report.

Support

The student will work closely with Dr. Hasan Zafari (Postdoctoral Fellow) and Dr. Farhana Zulkernine and will have access to one-one support and various software stacks necessary for the project.