INSA-Accademia Program: Undergraduate/Graduate Projects Catelogue

Information Network Security Agency

Contents

Part I. High Performance Computing

1 Implementation of a minature GFS

GFS (Google File System) is a scalable distributed file system for large distributed data-intensive applications that provides fault tolerance while running on inexpensive commodity hardware [2]. The aim of this project is to implement a miniature version of the GFS by implementing the core parts of the GFS. Hadoop Distributed File System (HDFS) is an open source implementation of the GFS by the Apache Software foundation. Sifting Hadoop's openly available source code may give some clues regarding GFS implementation.

PROJECT_TYPE: Graduate | Undergraduate

2 Design & Implementation of an In-memory database engine

Since the last few years, amount of data on the internet is exploding. To fully exploit the power of this 'big' data, a system that supports ultra-low latency services and real time data analytics is required. From hardware point of view, analysis of big data consists of three components: the processor to perform the calculations, the storage to store the data and the network between a combination of these two. The slowest of these components being the weak link (bottleneck) for the overall performance. Current trend shows that the processor's power is not used to full capacity because the data to be processed is not retrieved fast enough from the hard disks. Hence, to make read requests faster, a storage system that caches the data with an efficient (caching) algorithm (strategy) is required.

In this project, members are expected to design and develop an in-memory (i.e in-RAM) database engine that is optimized for a specific database whose table schemas and application behaviour are going tobe provided by INSA.

PROJECT_TYPE: Graduate

3 Fast Ethiopic Search Engine

Search engines have been around for a while, Google being the quintessential example. However, search engines based on the Ethiopic character set (U+1200 – U+137F) are either unavailable or not efficient enough to be adopted by the body politic. This project, therefore, would consist of the following tasks.

1. Fast web crawler to download contents that contains Ethiopic characters
2. Fast content ranking
3. Building a minature search engine (using (1) & (2)) using the Hadoop ecosystem(or a similar technology that supports parallel processing)

(a) Building a search engine may entail some acquaintance with data mining techniques like TF-IDF

PROJECT_TYPE: Graduate | Undergraduate

4 Fast MRI algorithms implementation

Graphics Processing Units (GPUs) can make advanced Magnetic Resonance Imaging (MRI) reconstruction algorithms attractive in clinical settings, thereby enabling low budget hospitals and clinics to analyze MRI scans in time and in budget. This project will go a long way to enable many people in Ethiopia to get high quality health care in their immediate vicinity by reducing the time and cost of travelling to hospitals located in big cities. Working on this project entails some of the following:

1. Studying existing MRI reconstruction algorithms and identifying those that are pliable to parallelism
2. Creating the parallel version of the algorithm chosen in (1).
3. Identifying the required hardware that solves the problem within acceptable time frame while minimizing the hardware cost on which it runs.

PROJECT_TYPE: Graduate | Undergraduate

References

1. BIG-DATA Platform & BIG-DATA Analytics Research & Development Experimental Development Approaches in HPC and Data Mining
2. Ghemawat, Sanjay, Howard Gobioff, and Shun-Tak Leung. "The Google file system." ACM SIGOPS operating systems review. Vol. 37. No. 5. ACM, 2003.
3. Hindman, Benjamin, et al. "Mesos: A Platform for Fine-Grained Resource Sharing in the Data Center." NSDI. Vol. 11. 2011.

INSA Contact Person(s)
Aradom Tassew - 0936542694 Elham Abdulhai - 0935999075

Part II. Hardware Design

1 Designing a system consists of FPGA, micro-controller and two flash storage devices

Complexity and performance of parallel systems using FPGA processors depends on the selected mode of programming. Parallel slave mode is the best and easiest way of programming Xilinx FPGAs directly from CPU or microcontrollers without using JTAG cables. In this project, it is expected to design a system which consists of a microcontroller (for parallel slave mode programming), and two flash storages (one for program, and another one for data storage).

PROJECT_TYPE: Undergraduate

2 Implementation of AES algorithm in an FPGA using VHDL programming language

Due to the repetitiveness of construct blocks and parallel in nature, implementation of most cryptography algorisms is suitable in FPGA processors. Advanced Encryption Standard (AES) is a symmetric encryption algorithm. AES consists of parallel and repetitive blocks and its implementation has been in intensive discussion since its first publication by National Institute of Standards and Technology (NIST) in 2000. However, the studies of low area, low power and low cost implementations require an FPGA processor. In this project, implementation of an optimized (low area, low power consumption, and high throughput) AES algorithm in FPGA is expected.

PROJECT_TYPE: Undergraduate

3 Design a high performance computing (HPC) using multiple FPGA processors (master and slave architecture)

Field programmable gate arrays (FPGAs) are emerging in many areas which require high performance computing (HPC) - either as tailor made signal processor, systolic array, software accelerator or application specific architectures. FPGAs are so flexible and reconfigurable that they are capable of massively parallel operations, explicitly tailored to the problem at hand. There are lot of paradigms to put FPGAs at work in a high performance computing environment. The outcome of this project is system of FPGAs with -

1. Master - which serves as controller
2. Slave - which serve as computing nodes and

3. Communication protocol - which serves as a communication bus between master and slave nodes.

Besides, power estimation and PCB design considerations are part this project.

PROJECT_TYPE: Graduate

4 Design and implementation of software defined radio (SDR) using FPGA processors

Software-defined radio (SDR) refers to the class of reprogrammable or reconfigurable radio via software, in another word, it is a radio in which some or the entire physical layer functions are implemented in software. The use of SDR technology is predicted to replace many of the traditional methods of implementing transmitters and receivers while offering a wide range of advantages including adaptability, re-configurability and multi-functionality encompassing modes of operation, radio frequency bands, air interfaces, and waveforms. The most computationally intensive part of a SDR is the channelizer, which extracts multiple narrowband channels from a wideband input signal. The objective of this project is to implement and design the channelizer using FPGA processor.

PROJECT_TYPE: Graduate | Undergraduate

5 Implementing a system on chip (SoC) Design on an FPGA processor

Design and test engineers face various problems in a SoB (system on board) design, which uses different components. The common problems are signal integrity, huge power consumption and increase in testability challenges. System on chip (SoC) implementation is the solution for this problem. With this, the designer can concentrate on the complete system rather than checking the correctness or performance of the individual components. The objective of this project is to design and develop a system containing a

1. CPU
2. SRAM storage
3. Bidirectional data bus and
4. serial communication interface protocol (UART)

The implementation is expected to be conducted on a single FPGA processor using soft-cores from Xilinx or Altera FPGAs.

PROJECT_TYPE: Graduate

INSA Contact Person(s)
Gebrehiwot Abraha - 0923289075

Part III. Operating Systems

1 Real Time Network Monitoring Tool for Linux

Linux distributions usually incorporate network utili- ties,such as ifconfig, ethtool, iperf and many others, used to monitor networking. These commands, however, does not provide many functionalities in one. Administrators has to use different utilities to get different functionalities. Most of them also lacks some important features and the presentation is not quite suitable for linux administators. The main objective of this project is to develop a real time network monitoring tool, using C programming language ,which altleast consists of the following features:

1. Display network interfaces current speed (link utilization in percent) of every interfaces found.
2. Top 5 applications consuming high traffic bandwidth.
3. Network statistics (total, recieved , transmitted, dropped) packets and bytes of every network intefaces.
4. Link status (up,down) of network interaface
5. Network Controller type (intel, broadcom, realtek)

PROJECT_TYPE: Undergraduate

2 Eclipse Plugin for NOS SDK

NOS SDK is basically developed for network application developers targeting INSA's Network Operating System. The current SDK version supports Linux based host machines only, and it is not integrated into any of IDEs. However,most developers usually prefer to develop applications using IDEs like eclipse. Because IDEs provide easy build and debugging interfaces and, fast and flexible code editors. It also allows program portability across different host architectures and operating systems as well. The objective here is to develop SDK plugin for Eclipse IDE.

PROJECT_TYPE: Undergraduate

3 Embedded Linux Based House Automation

Nowadays, the availability of fast mobile internets (3G and 4G), and also mobile phones currently being inseparable part of our daily life, different remote management systems are being introduced. Smart Home Control System enables users to manage home appliances (light, security cameras ,TV, cooking appliances,refregirator and many others) remotely using mobile phones. Basic requirements of the system are:

1. Get status of different Home appliances through mobile phone
2. Send commands to different appliances from mobile phones.

The system should be based on embedded linux and embedded hardware boards (e.g. Beagle Board)

PROJECT_TYPE: Undergraduate

4 Security analysis of SDN based networks

Software-defined Networking (SDN) is a current trend in communication networks based on the concepts of control plane and data (forwarding) plane separation, and logically centralized control. Because the technology is now in its early stage, security of SDN based network ,however, is not getting a focus it deserves. The purpose of this project is to evaluate security of SDN network with respect to traditional network and identify possible security holes and vulnerabilities of SDN architecture (centralized and distributed architecure). The experiment shall be conducted based on Opendaylight Controller.

PROJECT_TYPE: Graduate

5 ONOS based campus network

ONOS is a widely adopted SDN controller by many telecom service providers. It suppports distributed architecture which makes it preferable for a large network with lots of subscribers. The Objective here is to emulate a campus network using ONOS controller and develop the following applications:

1. Traffic Analyzer: Displays each users traffic usage
2. Host Management: Forbid black listed users from accessing internet.

PROJECT_TYPE: Graduate | Undergraduate

INSA Contact Person(s)
Yared Taye, 0913140377

Part IV. Biometrics

1 Fingerprint Recognition using wavelets

Fingerprint verification is one of the most reliable personal identification methods and it plays a very important role in forensic applications like criminal investigations, terrorist identification and National security issues. Some fingerprint identification algorithm (such as minutiae extraction) may require so much computation as to be impractical. Wavelet based algorithms may be the key to making a low cost fingerprint identification system. Wavelet analysis and its applications to fingerprint verification is one of the fast growing areas for research in recent year. Wavelet based technique results in high recognition rates. The use of wavelet in image feature extraction lies in the fact that the wavelet transform clearly separates the high frequency and low frequency information on a pixel by pixel basis. Wavelets are well localized in both time and frequency domain and often give a better signal representation using Multiresolution analysis. Wavelet transform based method takes less response time which is more suitable for online verification with high accuracy, reliable and better technique. The fingerprint recognition methods in wavelet transform domain converts the fingerprint images directly to the feature vector form, and then verify it. It preserves a potential of developing a lost cost, and small size fingerprint recognition system.

In this project, a student/group of students will undertake a literature review of the various published studies in the area of wavelet based fingerprint recognition, analyze and select the most promising algorithms, implement the algorithms using C/C++ and finally evaluate their implementation using online free test databases. The project does not need additional materials except a personal computer.

PROJECT_TYPE: Graduate | Undergraduate

2 Liveness detection based anti-spoofing methods in face recognition

Biometrics is an emerging technology that enables uniquely recognizing humans based upon one or more intrinsic physiological or behavioral characteristics, such as faces, fingerprints, irises, voices, etc. However, spoofing attacks are still a fatal threat for biometric recognition systems. Therefore, liveness detection, which aims at recognition of human physiological activities as the liveness indicator to prevent spoofing attacks, is becoming a very active topic in field of biometrics. Although numerous recognition approaches have been presented by the face recognition community, the effort on anti-spoofing methods is still very limited. The most common faking method in face recognition is to use a facial photograph of a valid user to spoof face recognition systems. Nowadays, video of a valid user can also be easily captured by chip everyday cameras for spoofing. Therefore

anti-spoof problems should be well solved before face recognition could be widely applied in our life. Most of the current face recognition works with excellent performance are based on intensity images and equipped with a generic camera. Thus, an anti-spoofing method without additional device will be preferable, since it could be easily integrated into the existing face recognition systems. The aim of the project will be studying, selection, implementation and testing of state-of-the-art liveness detection methods in face recognition. Only a web cam will be enough to successfully finish the project. Alternatively, online free test databases can be exploited for a more scientific and hassle free research project.

PROJECT_TYPE: Graduate | Undergraduate

References

1. Rafael C.Gonzalez and Richard E. Woods, "Wavelets and Multi-resolution processing," Digital Image Processing, Third Edition.
2. A. Graps., “An introduction to wavelets”, IEEE Comput. Sci. Engr., 2(2):50-61, 1995.
3. Online: Robi Polikar, “The engineer’s ultimate guide to wavelet analysis“,
4. U. Uludag and A. K. Jain. Attacks on biometric systems: A case study in fingerprints. In Proc. SPIE-EI 2004, Security, Steganography and Watermarking of Multimedia Contents VI, pp. 622–633, 2004.
5. Hadid, A. Biometrics systems under spoofing attack: an evaluation methodology and lessons learned. Signal Processing Magazine, IEEE. Sept 2015
6. Hadid, A. Face Biometrics under Spoofing Attacks: Vulnerabilities, Countermeasures, Open Issues and Research Directions

INSA Contact Person(s)
Animut Yigzaw, 0922175834

Part V. Data Mining

1 OCR Algorithm for Ge'ez Characters

It won't be an exaggeration to claim that Ethiopia's intellectual property is hardly digitized; and is stored in paper – be it in the form of century old parchment paper in monasteries or in the form of file cabinets in various regional and federal offices. Digitizing the plethora of documents by hand is not a feasible affair. To this end, building a system that identifies and digitizes documents made up of Ethiopian characters is a logical move. This system would play pivotal role in preserving the intellectual property of the country and making it easily accessible for the posterity. The project would entail many tasks; including but not limited to the following:

1. Machine learning: to train the system to recognize Ethiopic fonts(Doing this with high fidelity especially for century old hand written documents would be an interesting challenge).

2. Parallelizing the developed (or chosen) OCR algorithm to minimize the amount of time it takes to do the task (GPU based implementation seems the logical approach but other paths can be chosen).

3. Once the document is scanned, the words on the document need to be compared against the list of all known Geez, Amharic and Tigrigna .... Words for error identification/correction. This step may entail trying out different heuristic mechanisms. Besides, indexing all the words in a given language would be a necessary subproject. The project stated in [1] can also be an input to this project.

PROJECT_TYPE: Graduate | Undergraduate

2 Text Analysis Engine (TAE)

When users are looking for something online they go to a search engine first, fewer being the cases where they know the site they want to go to and navigate directly to it by typing their domain into the browser address bar. To this end, we will need a solution that works as a single, authoritative point for obtaining relevant information about the context we need, in this case Ethiopia.

The TAE, is a (set of) algorithm (s) for analyzing information regarding ethiopia and its interests. This analysis will be made on any data collected from internet news, videos comments, social medias, etc. The project would entail many tasks/algorithms; including but not limited to the following:

1. Multi source Data collector and extractor [optional]: Responsible for gathering textual information from different online and offline sources. The data can be - structured data (Eg - XML), unstructured data (Eg - Websites, Blogs and forums ...) or semi structured data.

2. Text Classification: Responsible automatically classifying documents into predefined classes based on their content.

3. Text Summarization: Responsible for reducing a text document into a compact datasets that retains only the most important features of the original document.

4. Sentiment analysis: Responsible for detecting the contextual polarity of text; that is, it determines whether a piece of writing is positive, negative or neutral.

PROJECT_TYPE: Graduate | Undergraduate

3 Data Processor

Data preprocessor is a tool with the objectives of integrating data from multiple sources into a coherent data format, transforming the integrated data into a form that is suitable for analysis and reducing this data into a smaller representation, yet closely maintaining the integrity, of the original data. This tool has methods and algorithms under each category (integration, transformation, reduction) in order to address the different kinds of problems and requirements that mass data has and the analysis requires respectively:

1. Data Cleaning - Attempts to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. Under Data Cleaning we have: Missing Values, Noisy Data and Cleaning.

2. Data Integration - Attempts to combine data from multiple sources into a coherent data format (store) Under Data Integration we have: schema integration, object matching and redundancy control.

3. Data Transformation - Attempts to transform or consolidate the data into forms that are appropriate for analysis. Under Data Transformation we have: Smoothing, Aggregation, Generalization, Normalization, Feature construction.

4. Data Reduction Data - Attempts to obtain reduced representation of the data set that is much smaller in volume, yet closely maintains integrity of the original data. i.e. Analysis on the reduced data set should be more efficient yet produce the same (or almost the same) analytical result. Under Data Reduction we have: attribute subset selection, dimensionality reduction, numerosity reduction, discretization and concept heirarchy generation.

PROJECT_TYPE: Graduate | Undergraduate

4 Event Correlator

Event correlation is a procedure where a stream of events from multiple sources, such as applications like anti viruses, firewalls, intrusion detection systems, servers and so on, is processed with a range of correlation methods in order to detect (and act on) certain event groups that occur within predefined time windows. The event correlator has different techniques and operations to correlate events, in real time, coming either from only same sources or multiple ones, individual events or groups of events, similar events or different kinds of events and so on:

1. Correlation Techniques - Under this category we have - rule based event correlation, finite state machine based, case based reasoning, model based reasoning, bayesian network based event correlation, neural network approaches.

2. Correlation Operations - Under this category we have - compression, logical operation, aggregation, filtering, suppression, thresholding, rate limiting, escalation, temporal relations, generalization, specialization, clustering.

PROJECT_TYPE: Graduate | Undergraduate

5 Data Mininig and Statistical Library

Data mining and statistical library is a set of data mining and statistical methods, models and algorithms for analyzing problems or data from different domain areas with the objectives of uncovering hidden patterns in massive data, which could be unimaginable to perform with the traditional and manual means. This set of methodologies, models and algorithms are targeted at making generic analysis for the majority types of data, mainly of network data, in order to detect hidden patterns indicating abnormal network data behaviors, which might inturn indicate the presence of attacks or misuse of the network resources.

PROJECT_TYPE: Graduate

References:

1. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

2. Diapanjan Das, Andre F.T. Martins, A survey on automatic Text Summarization, Nov 21, 2007.

3. Evans D.K Similarity Based Multi-lingual multi-document summarization Columbia University 2005.

4. Feng Hao John Daugman, Piotr Zielinsky, A fast search algorithm for alarge fuzy database, IEE Transaction on information forensic and security, june 2008

5. H. Bast, A,Chitea, F.M. Shutanek and I Weber.Ester : Efficeint search on text,entitites and relations. In SIGIR, Pages 671-678, 2007.

6. Data Mining, Concepts and Techniques (Jiawei Han and Micheline Kamber), Chapter Two

7. Event Correlation Engine (Master's Thesis), Andreas Muller, Spring Term 2009

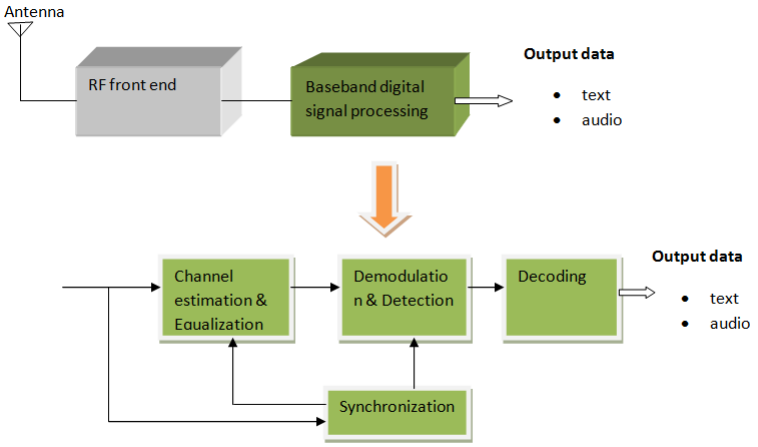
8. Weka Data Mining Software (Open Source)

INSA Contact Person(s)
Samuel Tamirat - 0911013980 Behailu Adugna - 0911346112

Part VI. Radio Communication

1 Software Implementation of a baseband signal processing block

In wireless communication systems the source data is encoded, modulated and transmitted over the unguided channel (air). During transmission, the signal encounters lot of impairments such as noise, fading and interference. At the receiver end, the effect of the channel should be estimated and compensated (equalized) so as to regain the transmitted signal. The received signal will be demodulated and detected into series of symbols which then will be decoded to retrieve the original message information. Hence, as depicted in the diagram below, for reliable communication an efficient receiver structure is required. The main functionalities of this structure will be to - equalize the channel effects, demodulate and detect the received signals over HF, VHF communication channels.



The goal of this project is to select reliable and efficient algorithms for equalization, channel estimation, demodulation and detection to fulfill the functional receiver which retrieve original message from digitally modulated signal. For this, the modulation types are limited to - Phase modulation (QPSK, 8PSK, 16PSK), Frequency modulation (FSK, GMSK), Amplitude and phase modulations (ASK2PSK8). The decoding module is applicable to known coding techniques such as PCM coded audio, or ASCII 7/8 formatted text.

INSA Contact Person(s)
Messele Araya - 0912603226

Part VII. Cryptography

1 Design and Development of a Cryptography Platform

Almost all internet communication is unencrypted and unauthenticated, leaving it completely unprotected against attacks. One might wonder why any programmer would fail to protect communication if free cryptographic libraries are readily available. The reason is often simply that cryptography is too slow; keeping up with high network loads requires many expensive computers with high electricity and maintenance costs. Analogous problems apply to smartphones and tablets, which have smaller network loads but also much smaller central processing units (CPUs) and limited battery life. Sometimes, rather than not deploying cryptographic protection at all, programmers react to performance problems by deploying low-security cryptography. The reason is that many cryptographic libraries allow trade-offs between security and performance.

Any government needs a trusted cryptographic base from which it can (legally-force to) build all security solutions for its communication and storage systems. The Reasons are -

1. To avoid or close weak links that usually happen as a consequence of a naive choice of cryptographic algorithms or cryptographic systems by ordinary application and system developers with no background in cryptography. Those naive choices are; ofcourse, a consequence different inherent tradeoffs of system attributes including but not limited to - security, simplicity (of implementation and usage), performance (speed) and energy (power consumption in large systems).

2. To Protect a data of diverse nature which can be systematically categorized into the following two classses

   (a) Resident_Data - Any data that is processed by exactly one program. Example - Document processors - word/microsoft-word2007, PDF/adobe- pdf, etc.

   (b) Transmitted_Data - Any data that is processed by two or more programes. Example - 2 programes - client-server/{mozilla, http-apache}, more than 2 programes - P2P/skype, P2P/Torrent/bittorrent, P2P/Tor, etc.

   Note that the above categorization is based on the morphology of the logic (program) that processes the data.

The design of a cryptographic platform shall consider the following characteristics -

1. Type of data - audio, video, image, text.

2. Nature of storage media - semiconductor, magnetic, optical.

3. Nature of transmission media - electronic, electromagnetic (radio), optical (photonic).
4. Nature of program's data access pattern - batch, interactive, real-time, streaming.
5. Nature of systems - Cryptography for different systems need different approach - OS, embedded, server, desktop.

The objective of this project is to design and develop a multilayer cryptographic platform on top of which different development, testing and prototyping works can be conducted by developers and researchers in fields of - cryptography, computing and commuications.

NOTE: Full concept paper of this project will by provided by INSA.

PROJECT_TYPE: Graduate | Undergraduate

Reference

1. Concept Paper: INSA Crypto-Platform, Concept Paper, INSA, Information Assurance Directorate
2. Securing Communication: High-security and high-speed protection for computer networks, white paper. Daniel J. Bernstein, Tanja Lange, Peter Schwabe

INSA Contact Person(s)
--

Part VIII. Quality Assurance

1 Hacking Game for INSA Talent Hunting Program

INSA has developed a talent hunting program to recruit talented professionals to work in job families such as penetration testing, computer emergency and response team (CERT) and cryptanalysis. To recruit the right professionals from universities and high schools, INSA wants to develop its own software game that is able to simulate different scenarios of hackings including social engineering attacks, hacking of client computers, web servers, database servers, applications, wireless targets and network devices (Example - Firewall). Depending on the type of the simulated hack, the game can provide small clues and finally should be able to profile the actions made by the examinee.

PROJECT_TYPE: Graduate | Undergraduate

2 Source Code Analysis Platform

Vulnerabilities in technology that could have been mitigated at the development stage are becoming an easy gateway for hackers and a nightmare for technology owners. Some of those vulnerabilities may not even have a compensating control that can protect the technology after its deployment. Hence it is mandatory to identify the security vulnerabilities at its earliest stage using source code analysis tools.

This source code analysis platform should accept a raw data developed in the most common programming languages such as C, C++, Java, PHP, Ruby, Perl, Python, ASP.net and identify the possible vulnerabilities. The platform should identify the most common vulnerabilities and provide a recommendation to mitigate the vulnerabilities. The platform should be modular enough sothat other developers can extend its capabilities the the following mechanisms -

1. Developing parser plugins for different programing languages
2. Developing/Extending vulnerability databases which contains known programing sequences

The following is a very rough and highlevel state machine of the source code analysis platform -

1. INPUT
    - source code
2. ACTION/LOGIC
    - load the right parser plugin for the input source code
    - load the right vulnerabily(s) database for the right parser plugin

- with a fast/suitable algorithm, search for vulnerabilities within the input source code

3. OUTPUT
   - list of vulnerabilities found within
   - where within the input source code the vulnerabilities occur
   - detail description of the vulnerability
   - Recomendations to mitigate the vulnerabilities

PROJECT_TYPE: Graduate | Undergraduate

INSA Contact Person(s)
Gebrekidan Gebremedhin - 0935999077