2020 MITS Practicum Project Descriptions

Sponsor & Project: Dr. Kathleen Carley, Social Media Analysis

Business Motivation:

Develop and test a databased procedure for supporting the study of influence operations in social media during elections. The system is intended to be part of a much larger toolchain.

Project Description:

Base Task:

- Analytics: Analyze who is influencing whom, and their impact in each election. In particular – were news agencies, news reporters, politicians, bots or trolls more influential. How does that change by topic. Which actors were trying to create excitement? Which were trying to create dismay? Are any of these actors relating known disinformation? Which were involved in polarization campaigns? Who were their targets? Actor types are pre-identified.
- Development: Develop a database system for social media data that supports selection of authors and documents of interest. In this case the emphasis is on Twitter. However think in terms of authors (e.g. the tweeter), documents (e.g., a tweet or reddit post), meta-data about author, meta-data about documents. Data selection may be based on author name, author attributes (e.g. number of followers, or is-a-bot), frequency of activity, document attributes (e.g., language, contains a url, contains an emoji), and/or date range. Complex selections should be possible in which the user rules in and out authors or documents of interest or not of interest. The database should be sql compliant and accessible from web-tools and client tools. In addition, built in features should support selecting the data of interest and exporting it into a form to be used with other tools – specifically ORA. Raw json from the two twitter APIs as well as output from ORA and csv files with attributes of authors or documents should be imported into database and fused with other data. Extend the system to support identifying near neighbors of documents – e.g. using word2ve and cosign similarity, or word match.
- Background: You should download and run ORA twitter report to see the networks and attributes extracted and so the kinds of structures that need to be

supported. Tools already exist to measure excitement, dismay, identify bots, trolls, cyborgs and memes. The excite and dismay might be augmented by y=using information from any emoji or emoticon.

Sponsor & Project: Software Engineering Institute, CERT

Business Motivation:

The CERT Executable Code Analysis team develops and maintains a set of program analysis tools, known as Pharos that are based on an increasingly complex infrastructure. There are many tools in the Pharos suite (see https://insights.sei.cmu.edu/sei_blog/2017/08/pharos-binary-static-analysis-tools-released-on-github.html) and they perform a variety of functions to automate and support program analysis and reverse engineering tasks.

Pharos tools typically export results in rudimentary data formats that are then ingested by other analytical tools. Despite generally using the same underlying libraries and infrastructure, the Pharos toolset is not easy to operate, inflexible, and individual tools do not integrate well together or with external tools.

Project Description:

The goal of this project is to develop a control center to unify the pharos tool set and make it easier to manage. This control center will enable program analysts to run, monitor, and export results from Pharos tools using a single interface. The ultimate vision would be to develop a project-based model where each tool can be applied on a common set of artifacts as needed, state is preserved as tools are run, and the results are easy to review and consume. Finally, we also want to support different deployment models (local versus distributed) to enable more sophisticated processing.

Deliverables:

The deliverables will be a prototype system to consolidate the pharos analysis tools and any/all supporting artifacts. This will include source code, test cases, design documentation, requirements documentation, and user manuals and miscellaneous documentation. If the team is successful, then there may also be an opportunity to present their work to the CERT team building the tools.

Sponsor & Project: Professor David Garlan, Explainable Games

Business Motivation:
Security attacks are increasingly being handled by automated mechanisms, often based on AI techniques. While effective in many situations, these approaches suffer from the problem that they are often opaque and hence hard to understand the rationale behind the automated decisions. In this project we plan on focusing on a particular kind of automated security response system to provide visual insight into how and why a system plans to respond to security attacks. This will not only help increase users' trust in the automated mechanisms, but help to design them correctly in the first place.

Project Description:
The purpose of this project is to develop a tool for visualizing and understanding the output off an automated security response planner based on the technique of stochastic games. Such plans are typically represented in the form of a game tree that captures the possible moves and responses of an attacker and defender (resp.). To accomplish this several capabilities are needed including (a) the ability to display a game strategy (generated by an automated tool), (b) the ability to understand how the "moves" in the game affect the system, and (c) ways to explore "why" the game strategy is the way it is, and "why not" some other strategy.

The project will therefore focus on three tasks, in order of importance: visualization, architectural effects, and model exploration. The first will allow users to visualize a game tree (or some other similar representation). The second will relate game moves to changes in a system's architecture (depicted using an existing architecture description tool and language). The third (developed if time permits) will allow the user to ask questions of the model, such as "why is this move taken here?" Or "why not do this other move at this point?".

Skillset Requirements:

- Programming skills.
- Experience with Java.
- Familiarity with some forms of AI are a plus, but not essential.

Deliverables

1. A game strategy visualizer tool for predefined security-related game strategies
2. A tool to show the impact of game moves on an architecture

3. A mechanism for answering user queries about a game such as “why” and “why not” questions (optional, as time and skills permit)

Sponsor & Project: Dollar Shave Club, Facial Recognition

Business Motivation:
We are interested in implementing a facial image processing based product recommendation system.

Project Description:
A deep learning engine capable of taking an input image and providing two tiers of results:

1. A set of identifiable facial features based on pre-set classes. Specifically, there will be a provided set of multi-class feature dimensions and the system should identify for each feature dimension which class the face belongs to.
    a. Sample dimensions:
        i. Beard/No Beard/Stubble
        ii. Beard/Stubble Growth: thin, thick, sparse
        iii. Skin Type: Dry/Oily/Combination
        iv. Wrinkles: None/Minimal/Average/Deep

2. A set of recommended products to purchase selected from among the Dollar Shave Club line of products.

Skillset Requirements:

- Familiarity with Modern Deep Learning Architectures such as Tensorflow, MXNet, PyTorch.
- Familiarity with facial feature detection techniques such as Constrained Local Models, OpenFace/Cambridge Face Tracker, facial landmark detection, face part identification, etc.
- Familiarity with Python and Data Science/ML coding within a python environment.
- Familiarity with 2D Image Processing Techniques such as Filters, Lighting Transforms, Fourier Transforms, etc.

Sponsor: Hasan Yasar, SEI CATS

Business Motivation:

During the 2018/2019 Academic year, the SEI was fortunate to work with a group of MSIT Students to produce the “CATS” (Continuous Authorization as a Service) system, as well as create the corresponding deployment pipeline, infrastructure as code, and a custom plugin for the Jenkins Continuous Integration service.

The SEI would now like to expand and enhance the original v1.0 of the “CATS” system. This system is entirely open-source, and will be a great contribution and utilization of MITS skill sets to give back to the community.

Project Description:
This project will consist of continuing development of the CATS (continuous Authorization as a Service) system. Additionally, students will deploy the enhanced version of CATS to a Kubernetes cluster on Azure.

Deliverables:

1. System to be deployed in an automated fashion to Azure
2. IaC for all Azure resources
3. Updated RMF (Risk Management Framework) controllers integrated into existing CATS codebase
4. Updated Jenkins plugin to work with latest stable release of Jenkins CI service

Skillset Requirements:

- Python 3/Django/PostgreSQL (CA service)
- Java/Maven 3 (Jenkins plugin)
- Git (version control system)
- Docker/Docker-Compose/Kubernetes
- AWS/Azure (Cloud Formation Template)
- Agile SDLC
    - SCRUM/Sprints
- DevSecOps Pipeline
    - Build Server (Jenkins)
    - Issue tracking system
    - Wiki

- Code repository (github)
- Automated testing (Selenium/Katalon)
- Automated Security Testing
- OwaspZap penetration testing
- SonarQube static code analysis
- Sonatype Nexus artifact repository

Sponsor: Home, LLC

Business Motivation:

1. As a recruiting test for potential hires
2. To gain insights on the processes and outcomes for analyzing a large data set
3. To give back to CMU (both founders are CMU alums)

Home.LLC is an private equity fund that will invest in downpayments for residential home buyers in exchanging for retaining the change in home value. Home.LLC will get paid a part of the change in home value when the home buyers sells it in 10 years. In order for Home.LLC to invest in the right homes, we need to build a data science model to predict future home values in 10 years (2030).

Project Description:

For this project – we need to develop a model that can predict home prices for US residential homes in 2030. To develop projections – the model will analyze various parameters that lead to an imbalance in supply and demand for homes in different cities.

The input data will be different variables like new housing starts, salary, etc. sourced from publicly available sources like FRED and US Census data as well as private data sources like Attom. The projected variable will be the home price index (HPI).

Deliverables:

1. Data science model predicting the home price index for selected zip-codes till 2030 (main deliverable)
2. Visual representation of data science model (supporting deliverable)
3. Pitch deck summarizing the process and model (optional deliverable)

Skillset Requirements:

- R Studio
- Python
- Looker/Tableau/Spotfire