SAND2018-7586D

The Center for Cyber Defenders

Expanding computer security knowledge

Expansion and Deployment of NLP Tools

• Ben Greene • Carleton College • Northfield MN

Project Mentors: Steven Elliot and Steven Morrison

Abstract:

Text analytics is a vital tool for extracting information from large sets of data, using machine learning and graph theory to categorize texts and gauge both sentiment and meaning. The software tools developed at Sandia combine UI elements and scripting to sift through data. Their capabilities include buzz entropy, document discovery, and trending analysis. In addition, the toolkit, Citrus, has a large Java library. The goal of this project is to expand and adapt aspects of Citrus to increase the use cases for the existing software and expand its foreign language application.

Case 1

In order to identify and mitigate cyber proliferation activity, vast amounts of data is gathered from various darknets. The format of the existing tools enables accurate analysis of English language data, but data in foreign languages poses a challenge because it requires the analyst to have a full understanding of the relevant language.

Approach and Solution:

There are two approaches to take. One is to provide immediate results on existing data using an analyst who knows the requisite language. The other is to develop a toolkit built to handle FL applications, like FL lexing, parsing, and stemming, as well as more accurate stopword and boilerplate removal. This will enable analysts who don't have a deep understanding of the language at hand to still be able to extract relevant and accurate information from FL data.

Case 2

The same text analytics tools are wrapped into software that aids cross organization analysis. Oftentimes analysts discover new ways of manipulating data but aren't able to implement them, which requires a new feature request. In addition, the current software is desktop based.

Approach and Solution:

By deploying the existing tools in a cloud environment, analysts can communicate with developers who can code up solutions, which are immediately deployed and can be used within hours of conception. In addition, this format will enable additional organizations to use these tools.

One way to do this is to use Notebook web frameworks like Jupyter or Zeppelin as frontend bases to query databases hosted on clusters using ElasticSearch. This enables both developers and analysts to make tools on the job.

Sandia National Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND2017-XXXXC.

U.S. DEPARTMENT OF ENERGY | NNSA National Nuclear Security Administration