A machine learning project in support of human rights
Why introduce machine learning into our data pipeline?

TO MAP THE THREATS & HARMS FROM THE USE OF SPYWARE AGAINST VULNERABLE COMMUNITIES & NGOS

Much of the data we use as part of ongoing research to trace and document cyber incidents relating to the misuse and abuse of commercial-grade spyware technologies comes in an unstructured form. With 100s, if not 1000s, of articles to process manual methods are no longer sufficient or efficient.

Our analysts are spending more time processing data than analyzing it!

By introducing machine learning we plan to automate much of the extraction of key information into a format usable by analysts.
Goals: What we seek to achieve

FASTER TIME TO ANALYSIS & THE IDENTIFICATION OF HIDDEN LINKS
Deliverables: What are we building?

TOOLS TO FACILITATE AND SPEED UP THE ANALYSIS PROCESS

Graph-oriented Named Entity Recognition   Table Question Answering with BigQuery

Document Summarization                   Document Question Answering

WITH OUTPUTS ACCESSIBLE TO AN ANALYST WITH LIMITED CODING SKILLS
PHASE 1
- Document-to-Text using Document AI from Google services
- NER using Google services only
- Pushing the structured entities to BigQuery via Apache Airflow

PHASE 2
- Labeling of the unstructured data and training of a custom NER model for ensembling
- Creating Knowledge Graphs using the NER ensemble
- Document Summarization using Google’s open-source Pegasus model
- Document Question Answering with LayoutLM
- Table Question Answering on a finetuned TAPAS model from Google

FOUNDATIONS
- Investigating the types of entities that would provide value within the data
- Configuring the data pipeline to feed into and extract from ML models
- Performing exploratory data analysis to understand the preprocessing requirements of the extracted raw data

→ Delivering in 2 week sprints through the agile methodology

Machine Learning Approach
# Project Timeline

<table>
<thead>
<tr>
<th>PHASE</th>
<th>Activity</th>
<th>Week</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Foundations</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Definition of project road map</td>
<td>03-16</td>
</tr>
<tr>
<td></td>
<td>Initial design of infrastructure &amp; data pipeline</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data exploration &amp; discovery [entity identification]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data exploration &amp; discovery [processing requirements]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Configure data pipeline to feed &amp; extract from ML pipeline</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Phase 1</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ML process [document-to-text]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>User Testing &amp; Feedback loop [doc-to-text]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ML process [NER via Google services]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>User Testing &amp; Feedback loop [NER]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ETL structured entities to data lake</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Phase 2</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Labelling unstructured data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Training custom NER for ensembling</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Creating knowledge graphs using NER</td>
<td></td>
</tr>
<tr>
<td></td>
<td>User Testing &amp; Feedback loop [knowledge graphs]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Document summarization</td>
<td></td>
</tr>
<tr>
<td></td>
<td>User Testing &amp; Feedback loop [document summarization]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Document Question Answering</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Table Question Answering</td>
<td></td>
</tr>
<tr>
<td></td>
<td>User Testing &amp; Feedback loop [Question ^ Answering]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data visualization &amp; analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Deploy into production</td>
<td></td>
</tr>
</tbody>
</table>

**PJMF Deliverables**
The CyberPeace Institute prides itself on the diversity of professional backgrounds and expertise whilst also empowering the younger generation to take part in our projects through internship programs.

<table>
<thead>
<tr>
<th>Role</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Learning Intern</td>
<td>ATA</td>
</tr>
<tr>
<td>Chief Information Security Officer</td>
<td>FLORENT</td>
</tr>
<tr>
<td>Senior Software Engineer</td>
<td>SIEGFRIED</td>
</tr>
<tr>
<td>Full stack web developer</td>
<td>NIKOLAOS</td>
</tr>
<tr>
<td>Chief Research &amp; Analysis Officer</td>
<td>EMMA</td>
</tr>
<tr>
<td>Senior Intelligence Officer</td>
<td>IAN</td>
</tr>
</tbody>
</table>
Success

Outcomes

PHASE 1

- Integrating a simple ML pipeline into the pre-existing data pipeline
- Having moderate success in recognizing entities in documents
- Successfully testing different use cases.
Success Outcomes

PHASE 2

- Extracting Knowledge Graphs out of unstructured documents
- Adding summarization and labeling functionalities to increase analysis efficiency
- Enabling the filtering of details by incorporating Document Question Answering capabilities
- Facilitating analysts to interact with knowledge graphs and extract information through natural language via Table Question Answering
# Project Components

## DATA

**Structured** Data - for learning  
**Unstructured** Data - to process and analyze

## CORE INFRASTRUCTURE

- Cloud Based Infrastructure  
  - **EC**: Kubernetes  
  - **Storage**: GCP Technology  
  - **Storage**: AWS + GCP  
  - **ETL**: Airflow, Cloud Dataflow

## MACHINE LEARNING

- **NER**: Custom models, GCP Services  
- **Labeling**: Custom models, Open-source  
- **Summarization**: Open-source  
- **Question Answering**: Hugging Face

## END USER TOOLS

- **Graph analytics**: GraphXR, Maltego  
- **Dashboarding**: Kibana / Looker / Grafana  
- **Search & query**: BigQuery
Our Infrastructure | today
Our Infrastructure in 6 months
Our Data

Structured Data
- Cyber incidents / events
- Entity-based information
  - Organizations [e.g. companies registration information]
  - Persons
  - Locations
- Exports or extracts from databases

Unstructured Data
- Media articles
- Blog posts / online articles
- Social media profiles [e.g. LinkedIn]
- Lengthy reports
- Exports from online databases
- Leaked documents or emails
- …
Our Data Pipeline
In 6 months
How do we achieve this

Though we still exploring the technologies and algorithms we will use; we currently expect to experiment with the following three pillars of Machine Learning.

1. **Knowledge Extraction**
   Extract all points-of-interest from unstructured text.

2. **Knowledge Organization**
   Create data structures so as to create structured outputs from unstructured data and potentially including structured data as well.

3. **Knowledge Search**
   Be able to search the organized knowledge and automate obtaining points-of-interest from the knowledge base.
EXTRACTING POINTS OF INTEREST

This pillar of the work deals with extracting points of interest from unstructured data namely reports in the form of pdfs, online articles and message forums.

This takes the form of three different problems:

- Named Entity recognition
- Document Labeling
- Document Summarization

Here machine learning can do all the tasks. A lot of work has been done in these areas of work and should be explored comfortably within the organization.
Knowledge Organization

CREATE EXPLORABLE DATA

This pillar takes the data from the knowledge extraction pillar, combines it with structured datasets and organizes it into a data structure that is explorable. This can take the form of knowledge graphs that seems to be the obvious fit.

This work requires some experience before being taken up. If achieved, however, this should lead to significant improvements in the workflow of analysts and make their work much more efficient.
Knowledge Exploration

KNOWLEDGE GRAPHS

This pillar of work comes after knowledge graphs have been created. This doesn’t need to, however. The whole area of information retrieval subfield in natural language processing is made for such cases. If the knowledge graph area is explored then it can be integrated with bayesian inference methods and LLMs for search that makes it extremely robust as compared to what exists today.
Project Risks & Mitigations

**KNOWLEDGE & EXPERTISE**

- Academic level experience in ML with limited professional experience.
  - Lean on PJMF & cohort expertise for support
- Introducing custom models into our data pipeline will be dependent on expert guidance from outside PJMF
  - Actively use the Slack channel to ask for help
- With limited experience in the team putting ML into a professional environment and data pipeline, our proposed phases and ML components may turn out to be overly ambitious given then project timeframe.
  - Revisit the project roadmap on a regular basis and update it accordingly

**DATA & INFRA CHALLENGES**

- Dataset may be too small to provide sufficient data for ML.
  - Acquire additional datasets using allocated budget.
- Variety of data sources could cause challenges when preprocessing the data and summarizing files.
  - Invest sufficient type at the Foundation stage to explore the existing data.
- Testing the accuracy of results from the chatbot will be difficult.
  - Involve different team members & receive recommendations from the cohort.
- We are yet to define our tolerance for false positives / false negatives with possible implications on delivery timeframe.
  - Be flexible as part of the PoC and consider stricter rules as the project matures.
- Latency issue connected to the runtime of ML model inference.
  - Executing the ML pipeline once fresh data is included to ensure that the extracted information is prepared for analysis instantly