Домашнее задание

Домашнее задание

В приложенном архиве *macos_correlation_rules.zip* содержатся папки с названием реализуемых тактик, а внутри - готовые правила корреляций по данной тактике. Данный датасет является **тренировочным**.

Структура правила корреляции:

- *i18n/i18n_{en/ru}.yaml* - файлы локализации. Они содержат описания корреляции на соответствующем языке.
- *tests/events_{i}_{j}.json* - структурированные ненормализованные события.
- *tests/norm_fields_{i}_{j}.json* - заполненные SIEM-поля для соответствующего события.
- *rule.co* - правило корреляции.

Домашнее задание

Важные поля в правиле корреляции:

- *category.high* - реализуемая тактика согласно MITRE ATT&CK.
- *category.low* - реализуемая техника согласно MITRE ATT&CK.
- *importance* - важность корреляции с точки зрения важности для ИБ: *low, medium, high.*

```
$importance = "medium"

$category.generic = "Attack"
$category.high = "Discovery"
$category.low = "Credentials from Password Stores: Keychain"
```

Домашнее задание

В приложенном архиве *windows_correlation_rules.zip* содержатся папки с названием по маске *correlation_{number}*, где number - номер корреляции. Все **дальнейшие задания** выполняются для этой директории.

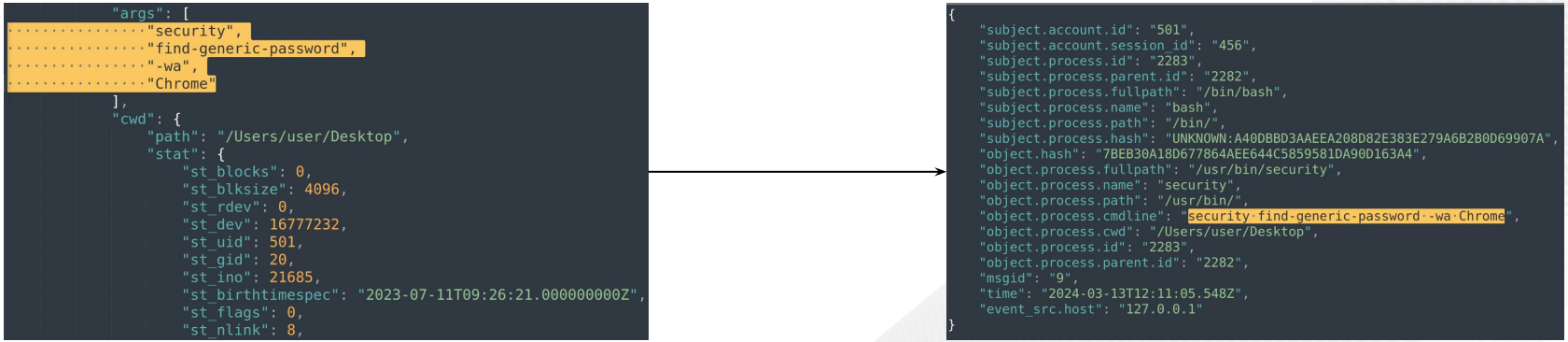
Каждая корреляция содержит файлы с маской *events_{i}_{j}.json*. Что это означает:

- Это группы ненормализованных событий по которым реализуется правило корреляции.
- i - это индекс группы, j - индекс события в группе. Событий в группе может быть от одного и более.

Домашнее задание

🎯 **Задание 1.** Выполнить нормализацию структурированных событий, согласно схеме полей SIEM. Описание полей SIEM на русском и английском языках хранятся в *taxonomy_fields/i18n_en.yaml* и *taxonomy_fields/i18n_ru.yaml*:

- Взять **структурированное событие.**
- Извлечь из него **только важную информацию.**
- Преобразовать такие значения в **единый, плоский формат**, понятный SIEM-системе.
- Привести к **стандартным именам полей** согласно схеме.



Домашнее задание

🎯 **Задание 1.** Ожидаемый ответ студента - для каждой корреляции в *windows_correlation_rules* и файла *tests/events_{i}_{j}.json* выполнить нормализацию и сохранить в файл с маской *tests/norm_fields_{i}_{j}.json*.

Домашнее задание

🎯 **Метрики по оценке задания 1.** Для оценки правильности заполнения нормализованных полей вычисляются метрики **Precision/Recall**.

Для вычисления Precision и Recall нужно сначала посчитать:

- **TP (True Positive) -** поле **правильно извлечено и заполнено**. Это означает, что в вашем ответе и ключ и заполненное значение идентично совпадает с эталоном.
- **FP (False Positive) -** поле **заполнено, но неправильно**. Это означает, что если сгенерированное поле отсутствует в эталоне - оно автоматически FP. Даже если само поле верное, но значение неправильное - тоже FP.
- **FN (False Negative) -** Поле **должно быть заполнено, но оно отсутствует**. Это означает, что если в вашем сгенерированном ответе отсутствуют поля из эталона - это FN.

Домашнее задание

🎯 **Метрики по оценке задания 1.**

Precision (Точность) – какая доля сгенерированных полей правильные?

$$Precision = \frac{TP}{TP + FP}$$

Recall (Полнота) – какую долю из всех эталонных полей мы смогли извлечь правильно?

$$Recall = \frac{TP}{TP + FN}$$

📌 Сравнение полей и заполненных значений будет производится над строками, приведенными в нижний регистр.

Домашнее задание

🎯 **Задание 2.** После успешной нормализации событий у вас на руках есть структурированные данные в формате SIEM. Теперь нужно **автоматически определить**, какое **правило корреляции** можно сгенерировать на основе группы событий (по индексу *i*) — то есть:

- Какая **тактика MITRE ATT&CK** (например, *Credential Access, Execution*) соответствует поведению?
- Какая **техника MITRE ATT&CK** (например, *T1555 — Credentials from Password Stores*) применима?
- Какой **уровень важности** (*low, medium, high*) следует присвоить событию?

Домашнее задание

🎯 **Задание 2.** Ответ для каждой корреляции *correlation_{number}* должен быть записан в JSON-файл *answers.json*, который нужно записать в соответствующей директории.

```
{
        "tactic": "Credential Access",
        "technique": "Credentials from Password Stores",
        "importance": "high"
}
```

Правильное название тактик и техник необходимо писать согласно https://attack.mitre.org/. Если для техники генерируется и сабтехника, то ее нужно написать через двоеточие. Например,

```
{
        "tactic": "Privilege Escalation",
        "technique": "Access Token Manipulation: Make and Impersonate Token",
        "importance": "high"
}
```

Домашнее задание

🎯 **Метрики по оценке задания 2** – Оценки за задание будет рассчитываться как accuracy по каждому из полей (tactic, technique, importance) на всём наборе корреляций. Итого во внимание будут приняты метрики по каждому из значений.

$$Accuracy = \frac{\text{Число полностью совпавших значений}}{\text{Общее число корреляций}}$$

Домашнее задание

🎯 **Задание 3.** Генерация файлов локализации на английском и русском языках.

На основе анализа нормализованных событий, а также определённых значений MITRE-техники и уровня важности, необходимо сгенерировать файлы локализации — *i18n_en.yaml* (английский) и *i18n_ru.yaml* (русский) — для каждой корреляции.

Цель — создать локализованные описания, которые будут понятны и полезны аналитикам на обоих языках, сохраняя техническую точность и соответствие стилю существующих правил.

Домашнее задание

🎯 **Задание 3.** Генерация файлов локализации на английском и русском языках.

Рекомендации по стилю и структуре:

Изучите примеры файлов локализации в директории *macos_correlation_rules/*
Там вы найдёте эталонные шаблоны, включая:

- Какие поля должны присутствовать (*Description, EventDescriptions, LocalizationId*)
- В каком стиле и тоне пишутся тексты (технически точный, краткий, ориентированный на аналитика SOC)
- Какие нормализованные поля событий используются для контекста (например, *object.process.name, object.process.cmdline* и др.)

Домашнее задание

🎯 **Метрики по оценке задания 3.** Оценка качества сгенерированных файлов локализации будет производиться **не по строгому совпадению строк**, а с помощью **семантического сходства эмбеддингов — BERTScore**.

- Для каждого из файлов локализации будет вычислена метрика F1 и сравнена с соответствующим эталонным файлом локализации.
- Оценка за задание считается как средняя оценка F1 по всем локализациям на русском языке и на английском языке. Итого во внимание будут приняты две оценки.

Рекомендации по выполнению заданий

Реализуйте **код (скрипт/ноутбук)**, который:

1. **Принимает** входные данные.
2. **Формирует промпт** для LLM, содержащий контекстную информацию.
3. **Использует LLM** одним из способов:
   - **Zero-shot** — просто спросить модель без примеров
   - **Few-shot** — дать 2–3 примера “событие → тактика/техника/важность”
   - **Prompt engineering** — тонкая настройка шаблона запроса
   - **LoRA / дообучение** — если есть возможность — дообучить модель на тренировочных данных (опционально, для продвинутых)
   - **RAG (Retrieval-Augmented Generation)** — использование релевантных фрагментов из базы знаний (MITRE ATT&CK, схема полей SIEM, примеры локализаций и другое)

Рекомендации по выполнению заданий

Технические требования:

- Язык: Python (рекомендуется)
- Можно использовать любую доступную LLM:
  1. OpenAI API
  2. Локальные модели через Hugging Face (Llama 3, Mistral, Qwen и др.)
  3. API-провайдеры с доступом к LLM:
     - 3.1. https://openrouter.ai/
     - 3.2. https://deepinfra.com/
  4. Ollama, LM Studio, vLLM и т.д.

Результат домашнего задания

```
window_correlation_rules/
    ├── correlation_1/
    │   ├── i18n/
    │   │   ├── i18n_en.yaml
    │   │   └── i18n_ru.yaml
    │   └──tests/
    │   │   ├── events_1_1.json
    │   │   └── …
    │       └── norm_fields_1_1.json
    │       └── …
    │   ├── answers.json
...
```

📌 Результат выполнения домашнего задания необходимо организовать согласно структуре директорий выше и собрать результат в zip-файл.
📌 Линк на соревнование в Github Classrooms