MWS Cup 課題3のご紹介

株式会社ＦＦＲＩセキュリティ
（東証マザーズ：3692）
https://www.ffri.jp/

2021-06-02

概要

概要

表層解析ログの分析を題材とした問題を出題予定

項目	概要
分析問題	表層解析ログを対象にデータを分析し設問に回答
分類問題	表層解析ログを対象にマルウェアとクリーンウェアに分類

マルウェア検知製品開発における実務的な感覚値や難しさのイメージを獲得して頂く

- 表層解析手法それ自体及び、クリーンウェア・マルウェア双方の見え方
- マルウェア検知の実際的な難しさ

問題構成

1. 表層解析ログ分析

- 課題データを事前に分析しておくことを推奨

2. 表層解析ログ分類

マルウェア/クリーンウェア分類

- 課題データを対象に**分類器を事前に検討しておくことを推奨**
- 良い成果は将来的な論文化やGitHubでの公開を期待

事前公開する課題データ(FFRI Dataset 2021と同形式)を使用

分析問題

小規模データセットに対してデータ分析を行って頂きます

8GibほどのjsonlデータをDuckDBで集計ができる環境をご用意ください

下記環境で分析できる程度のデータ量にする想定です

- Google Colaboratory
- Core i5 16GB GPUなし

使用するデータは事前公開予定

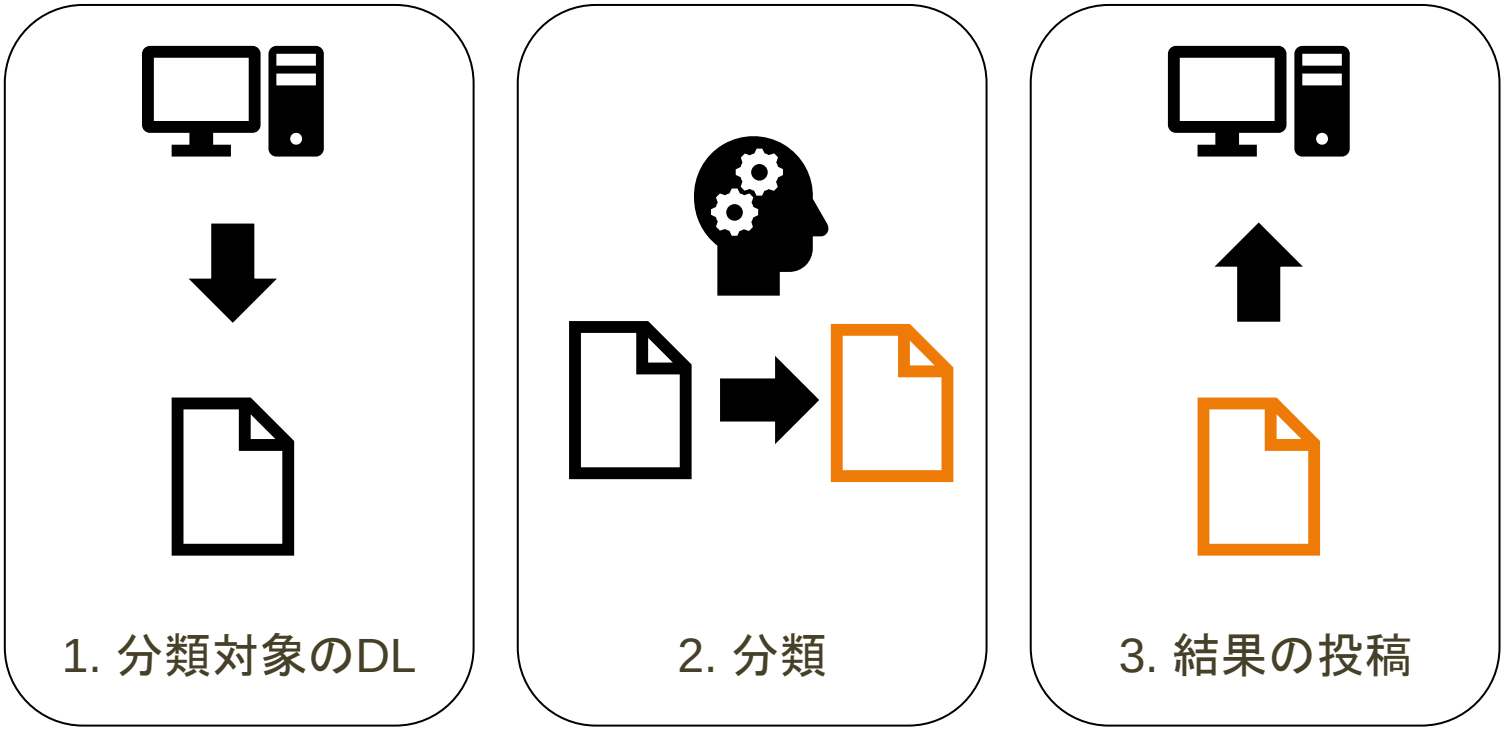
公開する際にはSlackにてご連絡いたします

出来る限り早期に公開したいと思っています

分類問題

マルウェア・非マルウェアの2値分類

- 事前公開する課題データを対象にマルウェアかクリーンウェアかを判別し、回答して頂く形の想定



おすすめ事前学習方法

過去問

MWS Cup 2019以降の過去問で雰囲気を把握しておく

MWS Cup 2019

- MWS Cupの公式ページに解説を掲載
  - MWS Cup - https://www.iwsec.org/mws/mwscup.html

MWS Cup 2020

- 公式ページには未掲載
- MWSのSlackの下記投稿にて解説資料・模範解答を共有
  - **https://mws-workshop.slack.com/archives/CDJSKVCTE/p1603787491113700**

データ分析

Kaggleなどを題材にデータ分析手法について学習しておく

弊社エンジニアの
おすすめ書籍

Kaggleで勝つデータ分析の技術：書籍案内｜技術評論社 - https://gihyo.jp/book/2019/978-4-297-10843-4

ドメイン知識

FFRI Dataset 2021

1.取得されているデータの各要素について詳しくなっておく

- データセット生成スクリプトに付属のjson schemaには
  各フィールドの概要が記載されている
- ffridataset-scripts/ffridataset_v2021.json at master · FFRI/ffridataset-scripts –
  https://github.com/FFRI/ffridataset-scripts/blob/master/schema/ffridataset_v2021.json

2. マルウェア・クリーンウェアの傾向について詳しくなっておく

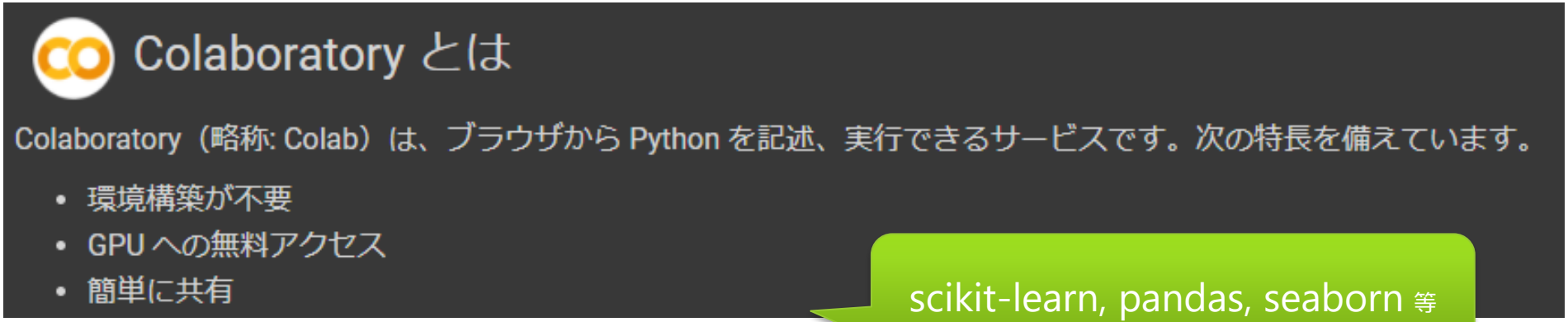
- 出題データはFFRI Dataset 2021のサブセット**ではない**が、参考になる点は多いはず
- 普遍的な特徴が見つけられれば強みになる
  - FFRI Dataset 2020と比較しつつ仮説・検証するのをお勧め
- 可能なら事前に分類器をつくって検討してみる

分析環境

利用する場合、Google Colaboratoryでの分析・分類の手慣らしをしておく

Jupyter Notebookインスタンスを無料で利用可能

https://colab.research.google.com/notebooks/welcome.ipynb?hl=ja



90分ルール・12時間ルールにご注意ください

- https://flat-kids.net/2020/07/28/google-colab-%E3%82%BB%E3%83%83%E3%82%B7%E3%83%A7%E3%83%B3%E5%88%87%E3%82%8C%E3%82%92%E9%98%B2%E6%AD%A2%E3%81%99%E3%82%8B/

課題データを外部サービスへのアップロードすることは原則NGですが、Google ColaboratoryはOKとします。
他のサービスを利用したい場合には、事前に#mwscup2021にて@oshiba(FFRI)までご相談ください！

専用特徴抽出パッケージFEXRD

FEXRDを利用した分類の手慣らしをしておく

FEXRD:

- FFRI Dataset 2021形式からの特徴量抽出を簡易化するパッケージ

FFRI/FEXRD: Feature Extractor for FFRI Dataset - https://github.com/FFRI/FEXRD

```
import json
from fexrd import StringsFeatureExtractor

sfe = StringsFeatureExtractor() # instantiae feature extractor class for the "string" element
fin = open("ffridataset_sample.jsonl", "r")
for l in fin.readlines():
    obj = json.loads(l)
    column_names, vector = sfe.get_features(obj["strings"]) # convert to the vector
```

- 昨年度に比べドキュメントも拡充してます
  - Basic Usage - FEXRD Documentation - https://ffri.github.io/FEXRD/basic_usage/