2020-2 임을규 교수님 컴퓨터보안

Assignment #2. Malware Classification

과제 제출: 모델 구현 중 작성한 소스코드 및 보고서를 본인의 Git repository에 업로드

제출 기한: 11월 20일 금요일 23:59까지. 제출 기한에서 한 시간 단위로 10%씩 감점, 최소 0점

문의 사항: 장준영 조교, lartist@hanyang.ac.kr (제출 관련 문의 등)

**과제 내용**

10주차 이론수업 강의에서 소개된 내용 중

정적 분석 결과로 얻을 수 있는 정보 중 하나인 opcode sequence

동적 분석 결과로 얻을 수 있는 정보 중 하나인 API sequence

중 한 가지를 선택하여 이를 활용하여 sequence가 주어졌을 때 해당 sequence가 악성코드인지 정상파일인지 분류하는 모델을 구현하고, 해당 모델의 정확도를 계산하여 보고서를 작성하시오

모델의 정확도를 계산하는 과정이 반드시 들어가야 함 (ex. 샘플의 일부분을 테스트용으로 사용하여 정확도 계산)

모델에서 사용하는 데이터 처리 기법 및 알고리즘은 자율 선택이며, 알고리즘의 고도한 정도 또는 분류 정확도와 같은 결과물은 채점 기준이 아님 (만점을 기준으로 본인의 모델에서 사용한 도구 및 알고리즘에 대해 설명이 미흡한 경우 감점이 발생)

**참고사항**

TF-IDF(Term Frequency – Inverse Document Frequeny) 기법
https://ko.wikipedia.org/wiki/Tf-idf

N-gram 기법
https://en.wikipedia.org/wiki/N-gram

Cosine Similarity
https://ko.wikipedia.org/wiki/%EC%BD%94%EC%82%AC%EC%9D%B8_%EC%9C%A0%EC%82%AC%EB%8F%84

머신러닝 모델 사용은 권장하나 필수사항이 아니며 추가점수로 반영되지 않음

Python의 train_test_split 모듈
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

Python의 여러 머신러닝 분류 모델
https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

**그 외 주의사항**

코드 작성은 본인이 직접 할 것. 소스코드 유사도 검사하여, copy한 과제는 0점 처리

프로그래밍 언어 제한 없음

보고서
분량 제한 없음
모델에 대한 설명 필수 (사용 모듈, 알고리즘 등에 대한 순서도를 넘버링하여 글로 설명하거나 그래프로 정리하여 작성)