제목

목차

개요: 간단하게만 말하기

방법론: 전체 pipeline은 다음과 같습니다. 총 4개 phase이며, phase0를 기반으로 phase 1,2,3 설계 및 프로그램으로 만들었다.

데이터분석:

1. 데이터 분석을 통해 1,2,3을 설계
2. 난독화 종류를 확인하였고, 횟수가 어떻게 되어있는 지, 그 결과는 다음과 같다.
3. TFIDF를 통해 토큰별 가중치를 확인, 중요 토큰 확인
   - **A.** 주목할 부분: 정상/비정상 데이터가 다름
   - **B.** 비정상은 악성코드에 사용되는 기능들을 확인 가능
4. 군집화가 잘 되어, 단어 임베딩을 통한 가능성 확인
5. 토큰 개수가 75% 가 1500개 미만 강조
6. 비율 skip
7. 데이터 형태는 간단하게
8. 정리하자면, 난독화 종류와 TFIDF, 데이터 길이, 비율 형태를 종합하여 ~~

난독화 처리:

1. 기존 난독화 처리 도구는 동적 분석은 시간면, 정적도구는 불가능 -> 새로운 도구 만듬
2. 난독화 처리시 가장 키 아이디어는 **iex/invoke-expression** 을 거친다는 것
3. 라인, 블록을 나누는 기준문자들을 고려하여 기준을 세우고
4. 스크립트 블록파싱 ->
   A. 예시들은 다음과 같다.
5. 파란색 네모칸이 iex에 해당하는 부분으로 이부분을 제외하고 echo를 통해 난독화 해제된 스크립트를 얻었다.
6. Echo를 통해 처리
7. 실제로 실행되는건 다음과 같이 된다.

8. 예외도 처리를 해주었다.
9. 이를 완료될 때까지 반복 후 데이터분석을 더 진행하거나, 데이터 전처리 단계로 이동하였다.
10. 정리하자면, 저희는 key idea에서 난독화 처리 툴을 발전시켰고, 난독화를 깔끔히 처리하였다.

여기까지 절반시간

데이터 전처리:

1. 데이터 정규화 과정을 거침 (영문 대소문자, 스타, 달러, 대시 제외 다 공백처리)
2. 100개 미만 스크립트는 제외