**프로젝트 제안서**

**프로젝트명 : AI에 기반한 Windows 실행파일의 카테고리별 분류**

**팀 이름 : 회 광 반 조**

**제출일 : 2020. 04. 06**

**팀 원 : 성호준, 김병렬, 정재현, 이주연**

1. **추진 배경**

* 한국 불법 SW 사용률 32%, 미국은 15% (BSA(세계 소프트웨어 연합), 2018)
* 불법적인 소프트웨어 사용은 저작권자들에게 금전적 또는 정신적 피해 초래.
* 사용자들도 악성코드에 노출될 가능성 높아짐.
* 소프트웨어 분류 기법의 활용

1. 저작권 침해 탐지
2. 악성코드의 분류
3. 소프트웨어 보관소의 소프트웨어 자동분류
4. 소프트웨어 필터링 시스템

* 불법 소프트웨어 사용 억제와 탐지

소프트웨어 필터링 시스템에서 유사도 측정을 통해 불법 소프트웨어를 식별함.

소프트웨어 분류를 활용하여 카테고리를 분류한 다음, 탐색 범위를 축소하면 평균 비교 횟수를 줄일 수 있음.

1. **목표 및 내용**

**목표 1. Machine Learning을 위한 데이터 추출 그리고 Dataset 생성**

* 기본적인 Windows 실행파일에 대한 이해 & 데이터 정보 선정

1. PE 구조의 이해
2. API와 Windows Kernel에서 사용하는 Native API에 대한 이해
3. CFG(Control flow graph), CG(Call Graph)에 대한 이해
4. 실행파일에서 어떤 데이터들을 통하여 분류를 진행하여야 효과적인지에 대한 리서치

* 데이터 추출과 Raw Dataset 생성

1. 데이터를 실행파일에서 추출하여 Raw Dataset을 구성하는 실습

**목표 2. 데이터 전처리 후 Machine Learning 실험 및 고려사항 리서치**

* 데이터 전처리 과정

1. 데이터 정제 이해 후 실습
   1. Machine Learning 속도 개선 고려

* Machine Learning 및 Deep Learning 알고리즘 이해

1. RF(Random Forest)
2. SVM(Support Vector Machine)
3. LMT(Logistic model tree)
4. NB(Naïve Bayes)
5. KNN(K-Nearest Neighbor)
6. DNN(Deep Neural Network)

* Machine Learning 실험
  1. 알고리즘 내 옵션 값 Depth, Tree 수, 등등에 대한 리서치 후 적용

2. 불순도에 대한 리서치 후 적용

3. 과적합에 대한 리서치 후 적용

**목표 3. 교차 검증 및 반복적인 실험을 통한 성능 개선**

* 교차 검증

1. 교차 검증에 대한 이해 후 적용

* 실험 성능 개선 방안 연구

1. 특징 정보

2. Machine learning 알고리즘

**목표 4. 서버 구성과 DB 연동 및 시각화**

* 서버에 대한 리서치 후 서버 선정
* DB에 대한 리서치 후 DB 선정
* 서버, DB와 연동 후 시각화 방법 선택

1. 웹
2. 프로그램
3. **기대 효과**
4. 불법 소프트웨어 사용 억제
5. 필터링 기법을 통한 불법 소프트웨어 탐지를 통한 원작자의 저작권 보호
6. 자동화된 소프트웨어 분류로 인해 사용자의 효율성 증가
7. 악성 코드 탐지에 용이해짐에 따라 사용자들의 안전 보장