


# 8조 Do it ! DREAM



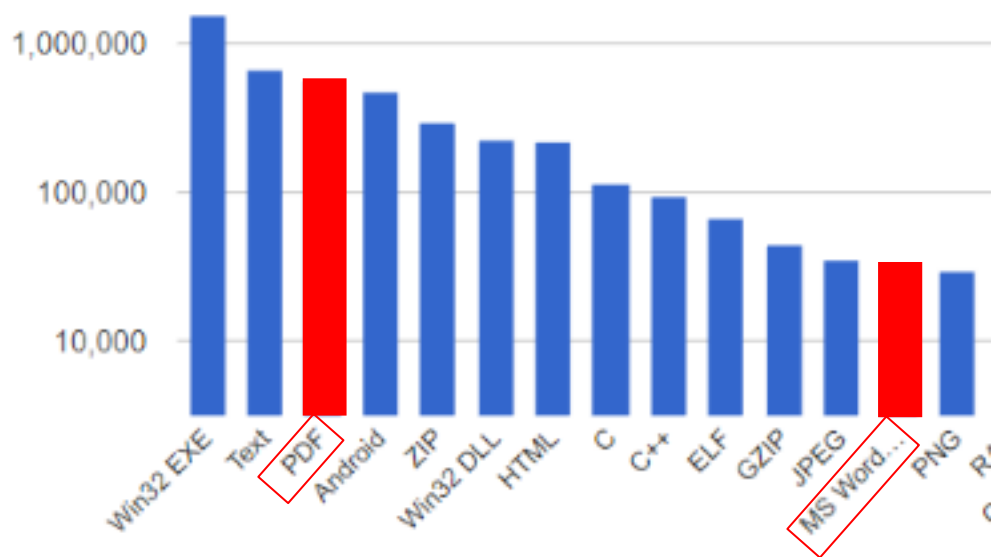
문다민 김기환 김현석  
정혜리 방유한  
윤명근 교수님



# 1. 프로젝트 소개

최근 가장 위협적인 사이버 공격은 PDF, MS Office 등 **문서**에 숨겨진 악성코드

〈 평균 일주일 동안 유입되는 악성 파일 유형 수 〉



출처 : <https://www.virustotal.com/ko/statistics/>



“**악성코드**가 숨겨진 **문서**로 인한 피해는 증가,  
하지만 이를 탐지하는 **백신**은 적음!”

# 1. 프로젝트 소개

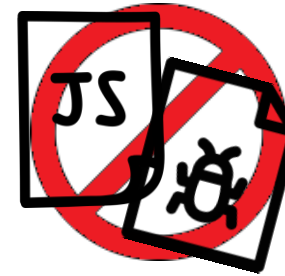
기존 기술의 한계점

## A사 제품 M



- 많은 시간과 비용 소요
- 직접 실행해야 하므로 사전 탐지 불가능

## B사 제품 S



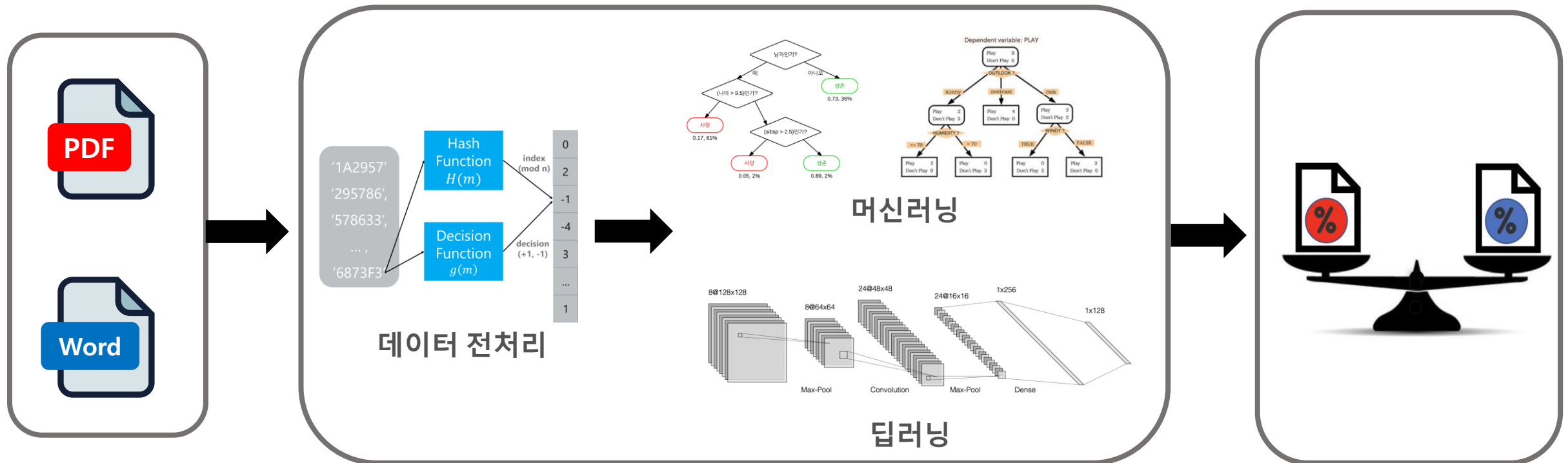
- 의심스러운 액티브 콘텐츠 원천 제거
- 파일 내용 임의 변경으로 정상적 사용 제약

# 1. 프로젝트 소개

프로젝트 목표

## 머신러닝 및 딥러닝 기반 정적 분석

악성 확률 예측

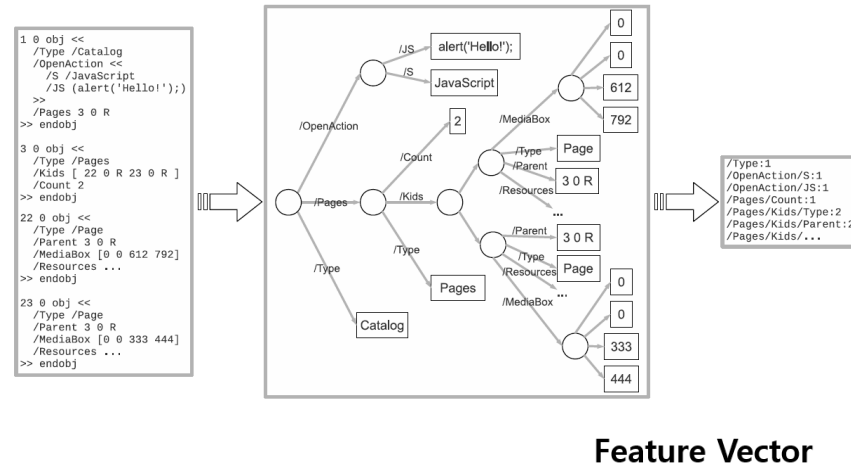


## 2. 수행 내용

### 머신 러닝 기반 PDF 탐지 방법

#### ■ 기존 피처 추출 방법

PDF 내부 태그(tag) 정보로  
피처 벡터 생성  
→ 머신 러닝 입력 데이터

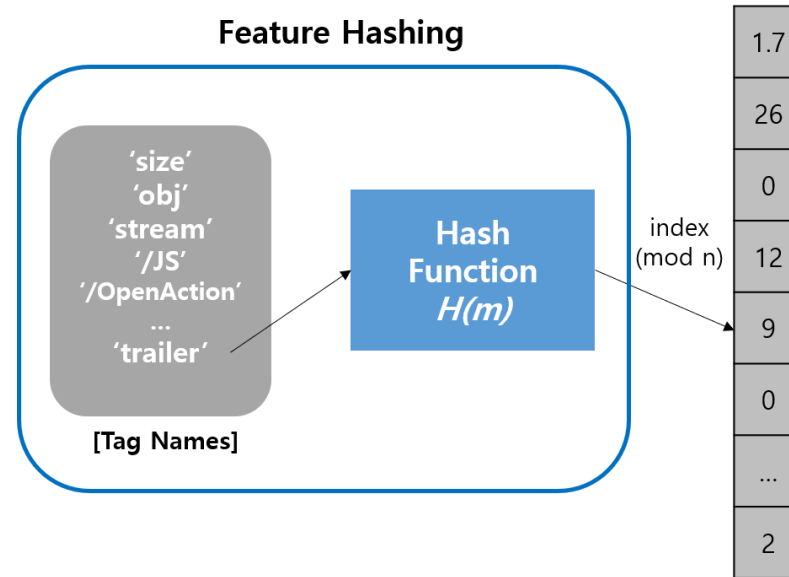


#### ▶ PDF 파일 문서 구조

출처: Nedim Šrndić and Pavel Laskov. Detection of Malicious PDF Files Based on Hierarchical Document Structure. In *Proceedings of the Network and Distributed System Security Symposium, NDSS 2013*

#### ■ DREAM 피처 추출 방법

태그 정보 + Feature Hashing  
→ 머신 러닝 입력 데이터



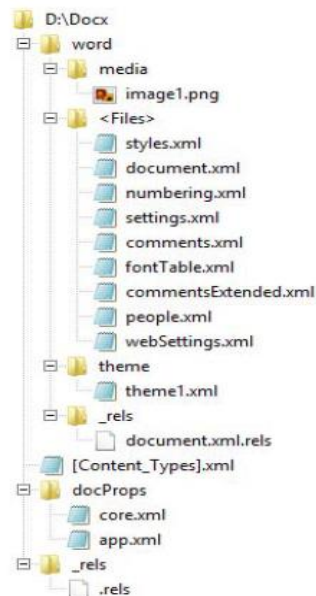
출처: Kilian Weinberger KILIAN, Anirban Dasgupta ANIRBAN, John Langford et.al. *Feature Hashing for Large Scale Multitask Learning*. Proc. ICML. 2009

## 2. 수행 내용

### 머신 러닝 기반 DOCX 탐지 방법

#### ■ 기존 피처 추출 방법

DOCX 구조 정보와 DF(Document Frequency)로 피처 벡터 생성  
→ 머신 러닝 입력 데이터



#### ▶ DOCX 파일 문서 구조

출처: Nissim, N., Cohen, A., Elovici, Y.: ALDOCX: detection of unknown malicious microsoft office documents using designated active learning methods based on new structural feature extraction methodology. IEEE Trans. Inf. Forensics Secur. 12(3), 63146 (2017)

#### ■ DREAM 피처 추출 방법

DOCX 구조 정보와 DF + File size + Entropy  
→ 머신 러닝 입력 데이터

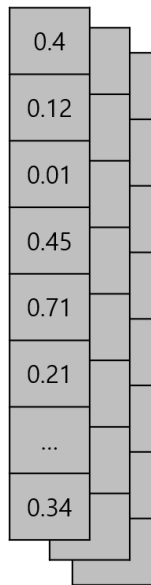
| Feature                      | DF    |
|------------------------------|-------|
| word/media/image1.jpg        | 0.036 |
| _rels/item1.xml              | 0.143 |
| word/_rels/numbering.xml     | 0.021 |
| word/_rels/webSettings.xml   | 0.414 |
| customXml/_rels/document.xml | 0.007 |
| ...                          | ...   |

+

문서파일 내부 데이터의  
min/max/mean **Entropy**  
min/max/mean **File Size**



Feature Vector

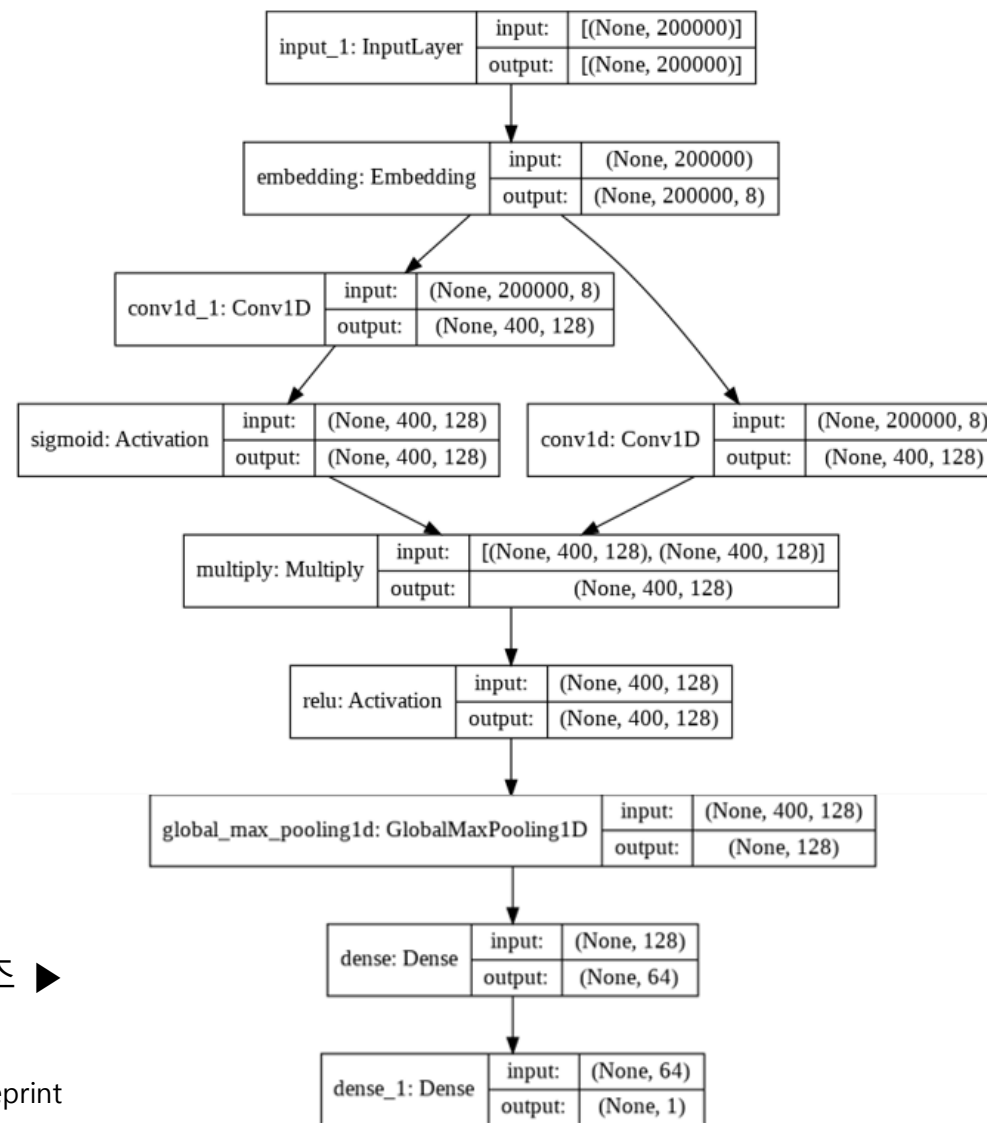


## 2. 수행 내용

딥 러닝 기반 PDF, DOC 탐지 방법

- 딥 러닝은 주어진 데이터의 주요한 특징을 추출하는 능력이 탁월한 것으로 알려져 있음
- 악성 파일의 특징을 추출하여 찾아낼 것이라 기대하여 딥 러닝을 사용해 실험을 진행하였음

딥 러닝 모델 구조 ►



출처 : E. Raff, J. Barker, J. Sylvester, R. Brandon, B. Catanzaro, and C. Nicholas, "Malware detection by eating a whole EXE", arXiv preprint arXiv:1710.09435, 2017.

## 2. 수행 내용

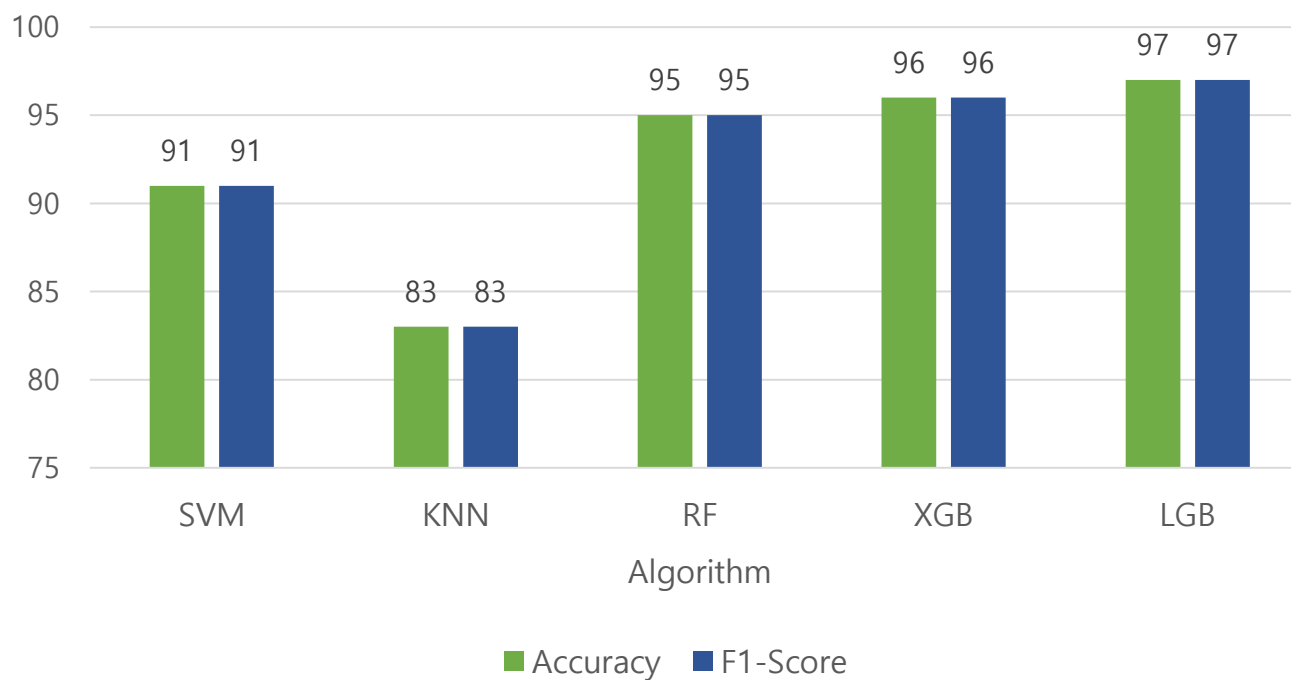
PDF 탐지 실험 결과

- 머신 러닝
- 딥 러닝

머신 러닝 탐지 ▪ PDF  
▪ DOCX

딥 러닝 탐지 ▪ PDF  
▪ DOC

PDF 탐지 성능 비교



- 학습 데이터
  - 악성: 100,000 개
  - 정상: 120,000 개
- 검증 데이터
  - 악성: 약 10,000개
  - 정상: 약 10,000개



## 2. 수행 내용

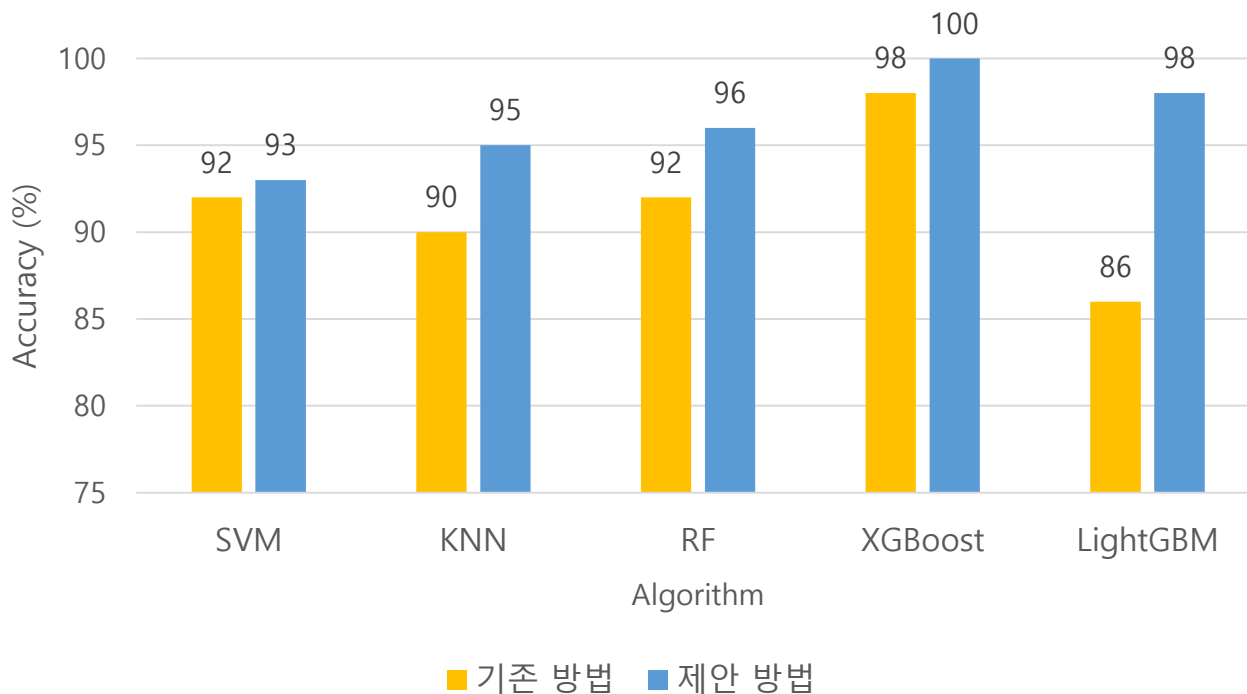
DOCX 탐지 실험 결과

- 머신 러닝
- 딥 러닝

머신 러닝 탐지 ▪ PDF  
▪ DOCX

딥 러닝 탐지 ▪ PDF  
▪ DOC

악성 DOCX 파일 탐지 방법 정확도 비교



- 학습 데이터
  - 악성 : 162 개
  - 정상 : 4,882 개

- MS Word는 수집한 데이터가 많지 않아 10-CV 을 통해 성능을 확인
- 기존의 특징 추출 방법보다 DREAM의 특징 추출 방법이 대체로 더 높은 정확도를 보임

## 2. 수행 내용

PDF, DOC 탐지 실험 결과

- 머신 러닝
- 딥 러닝

머신 러닝 탐지 ▪ PDF  
▪ DOCX

딥 러닝 탐지 ▪ PDF  
▪ DOC

### PDF

- 학습 데이터
  - 악성 : 약 100,000개
  - 정상 : 약 120,000개

(출처: 바이러스사인 2018.11월~2019.1월)
- 검증 데이터
  - 악성 : 약 10,000개
  - 정상 : 약 10,000개

(출처: 바이러스토탈 2017년도, 2018년도)

➡ 탐지 정확도 : 87.5 %

### DOC

- 학습 데이터
  - 악성 : 약 6,000개
  - 정상 : 약 16,000개
- 검증 데이터
  - 악성 : 800개
  - 정상 : 300개

(출처: 바이러스토탈 2017년도, 2018년도)

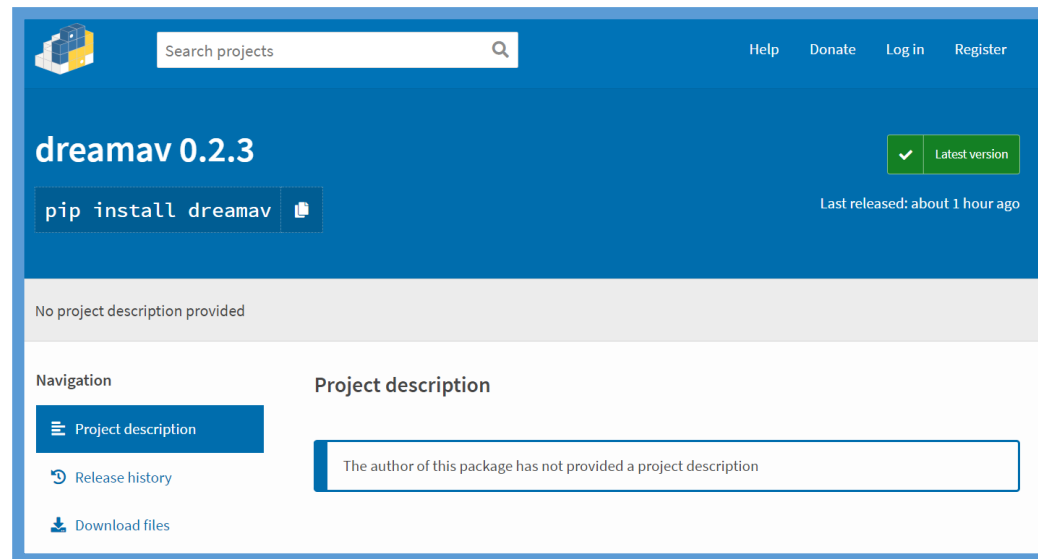
➡ 탐지 정확도 : 91.3 %

## 2. 수행 내용

DREAM 엔진 개발 특징



[엔진 DREAM의 서버 구조]



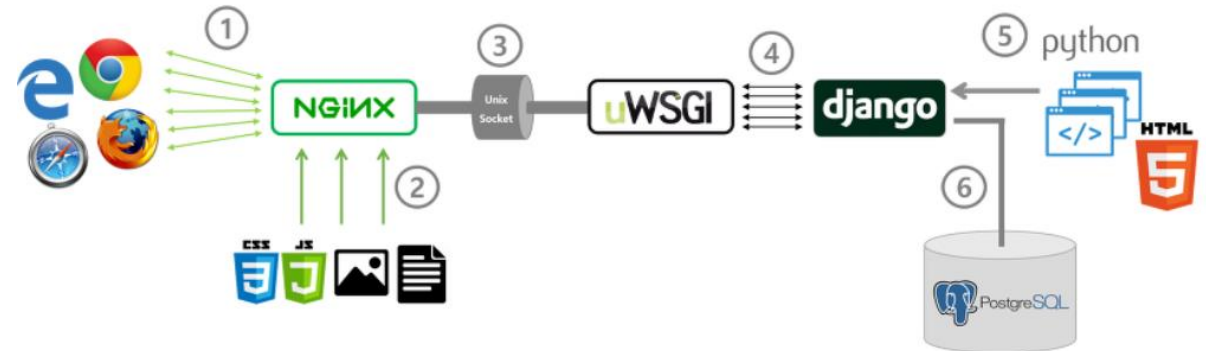
[PyPI 패키지 등록]

문서형 악성코드를 탐지할 수 있는 확장성 높은 엔진 **DREAM** 개발

## 2. 수행 내용

### 파일 공유 사이트

| Download                           |                                      |                                  |         |                           |
|------------------------------------|--------------------------------------|----------------------------------|---------|---------------------------|
| <input type="text" value="Word:"/> |                                      |                                  |         | LOG OUT                   |
| File                               | Name                                 | MD5                              | size    | type                      |
| Download                           | 000e207f32bf2e43c90d702f2b05be7e.pdf | 000e207f32bf2e43c90d702f2b05be7e | 178 KB  | PDF document, version 1.5 |
| Download                           | 0014546bcf4ad55d4d8971f80a063bc8.pdf | 0014546bcf4ad55d4d8971f80a063bc8 | 3037 KB | PDF document, version 1.5 |
| Download                           | 0016640ed91200002873d3deef721af0.pdf | 0016640ed91200002873d3deef721af0 | 28 KB   | PDF document, version 1.7 |
| Download                           | 001a25e187f1f7b5c06890720f86f934.pdf | 001a25e187f1f7b5c06890720f86f934 | 665 KB  | PDF document, version 1.4 |
| Download                           | 00253120a070485b3c87b572d3448cf2.pdf | 00253120a070485b3c87b572d3448cf2 | 52 KB   | PDF document, version 1.4 |
| Download                           | 0025400b48ddfdcf51045c4ce494a14.pdf  | 0025400b48ddfdcf51045c4ce494a14  | 7045 KB | PDF document, version 1.6 |
| Download                           | 003246cce8249f376e660b50c5c09648.pdf | 003246cce8249f376e660b50c5c09648 | 37 KB   | PDF document, version 1.7 |
| Download                           | 003765bb5d19193804a3533a6136a7ce.pdf | 003765bb5d19193804a3533a6136a7ce | 207 KB  | PDF document, version 1.7 |
| Download                           | 003be6251a3980ba65ae34180d0f4608.pdf | 003be6251a3980ba65ae34180d0f4608 | 153 KB  | PDF document, version 1.7 |
| Download                           | 003bf93b1839f774c68e0b3c4643bd39.pdf | 003bf93b1839f774c68e0b3c4643bd39 | 11 KB   | PDF document, version 1.7 |



문서형 악성코드 파일 공유 사이트 구현

## 3. 기대효과

### 문서형 악성코드 유포 방지



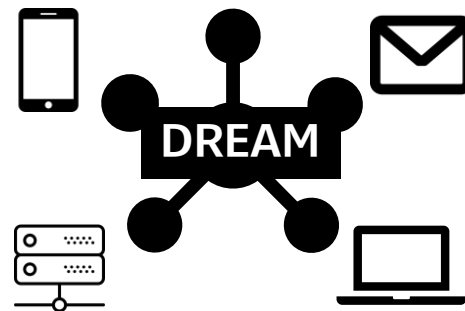
- 문서형 악성코드 조기 탐지
- 악성코드 유포 방지

### 오픈소스 소프트웨어



- 오픈소스 공개
- 커뮤니티를 통한 개발 참여

### 높은 확장성



- 메일, 웹, 모바일 등 다양한 환경 지원
- 편리한 설치 및 연동 지원

### 기존 탐지 방법 한계 극복



- 기존 동적 분석 탐지 기법 한계 극복
- 문서형 악성 파일 실행 전 단계 탐지 가능



국내 문서보안 최고 기업 😊 지란지교시큐리티 와 공동 연구 진행

# 감사합니다

8조 Do it !