DREAM

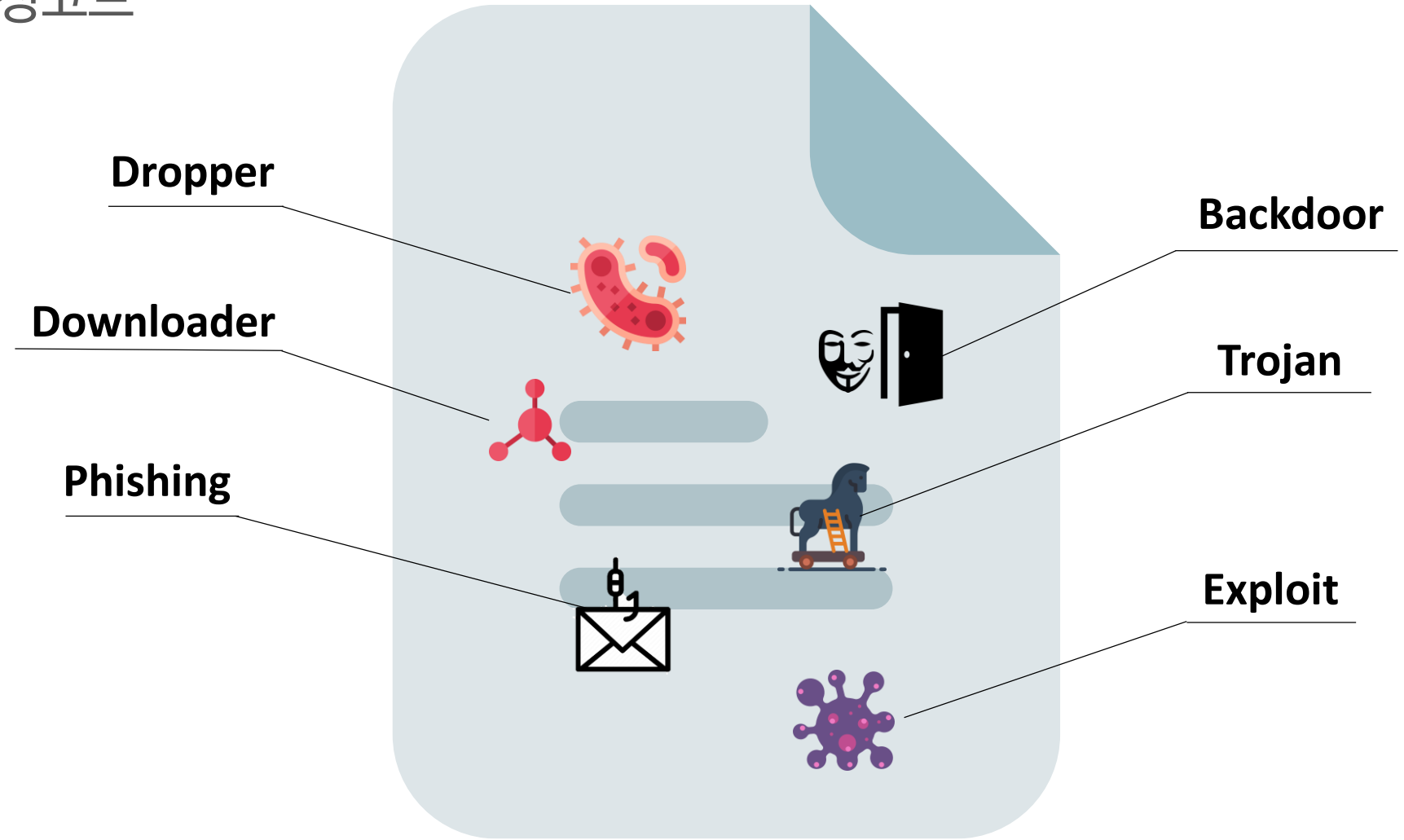
문다민 김기환 김현석
정혜리 방유한

8조 Do it !

목차

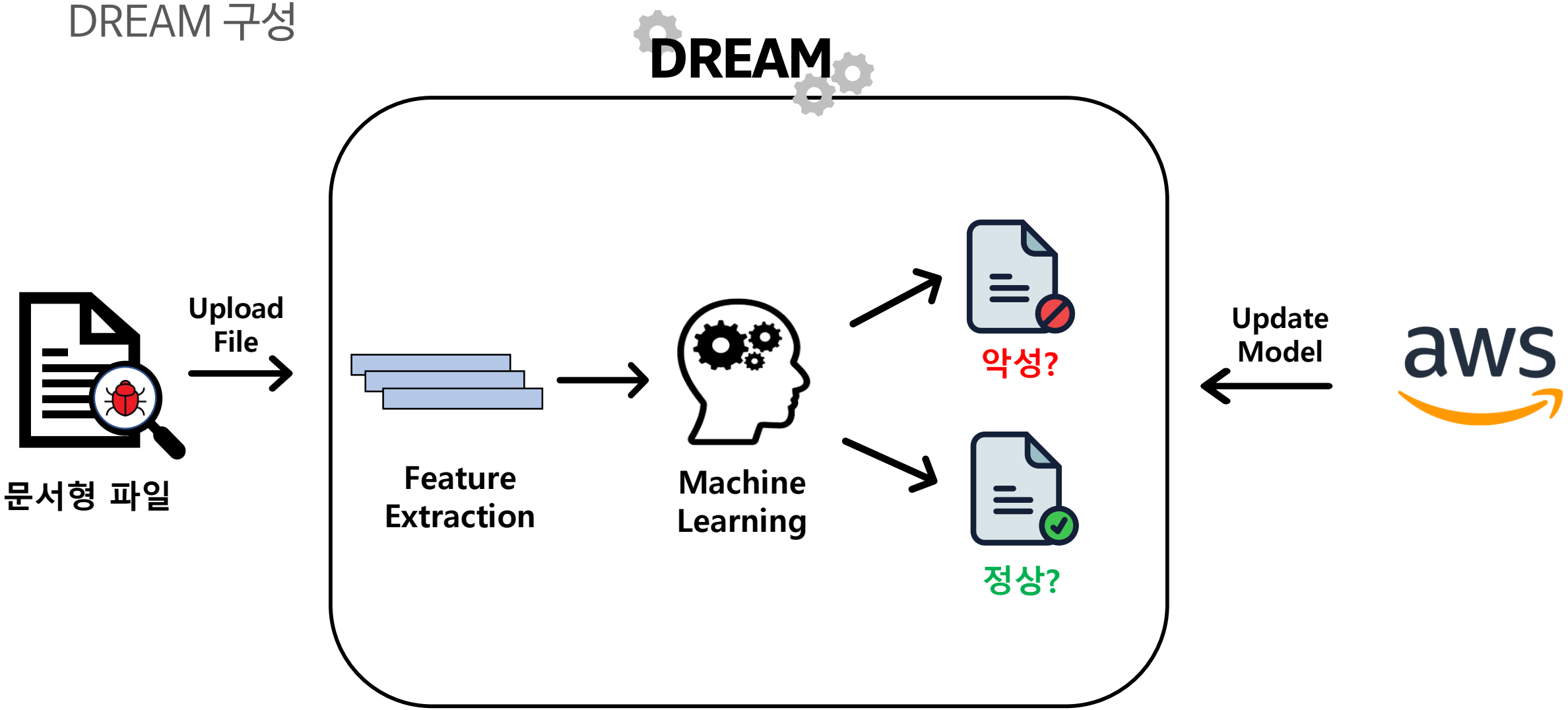
1. 프로젝트 소개

문서형 악성코드



1. 프로젝트 소개

DREAM 구성



1. 프로젝트 소개

프로젝트 목표

1. 문서형 악성코드 유포 방지

문서형 악성코드 유포 방지로 사회 문제 해소

2. 오픈소스 소프트웨어

여러 개발자들이 참가하여 개발함으로써 엔진 발전에 도움

3. 높은 확장성

메일 서버, 웹 서버 등에 엔진 사용 가능

2. 수행 내용

데이터 수집

Virus Share

contagio

VirusSign

HYBRID ANALYSIS

ViruSign

Google

2. 수행 내용

데이터 라벨링

KASPERSKY lab

F-Secure

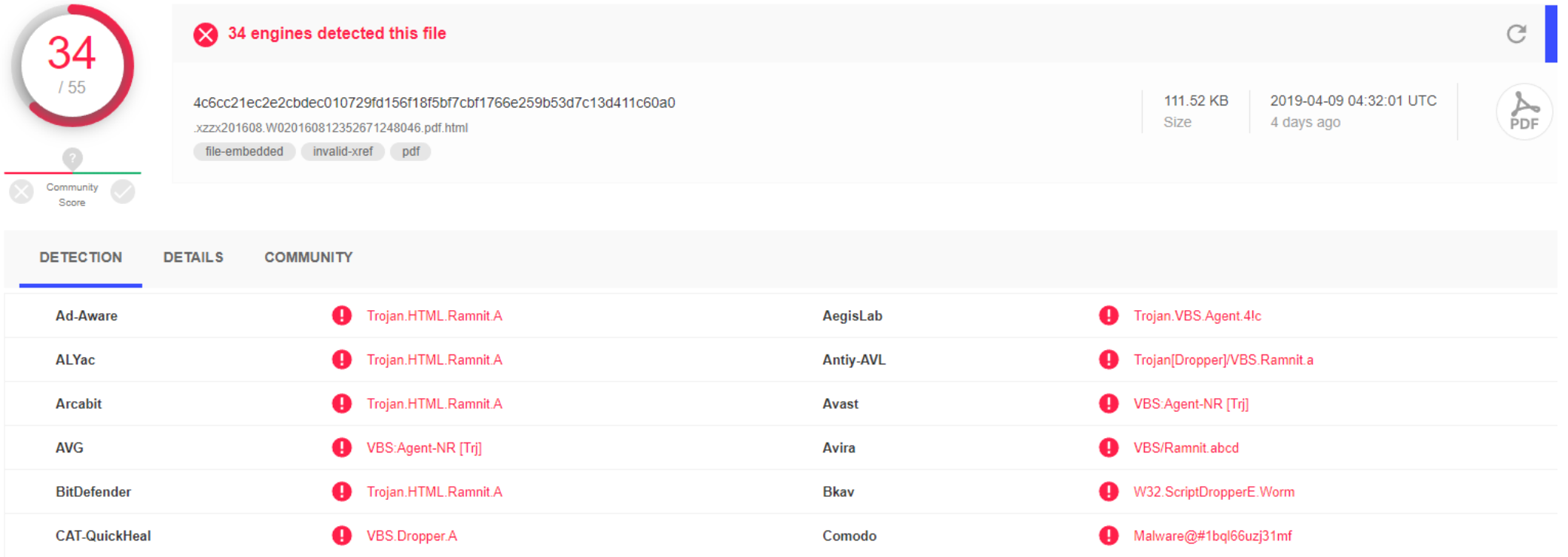
Symantec

글로벌 안티바이러스 테스트 기관 'AV-TEST'의 성능 부문 평가에서 수상한
안티바이러스 Kaspersky, F-Secure, Symantec

2. 수행 내용

데이터 라벨링

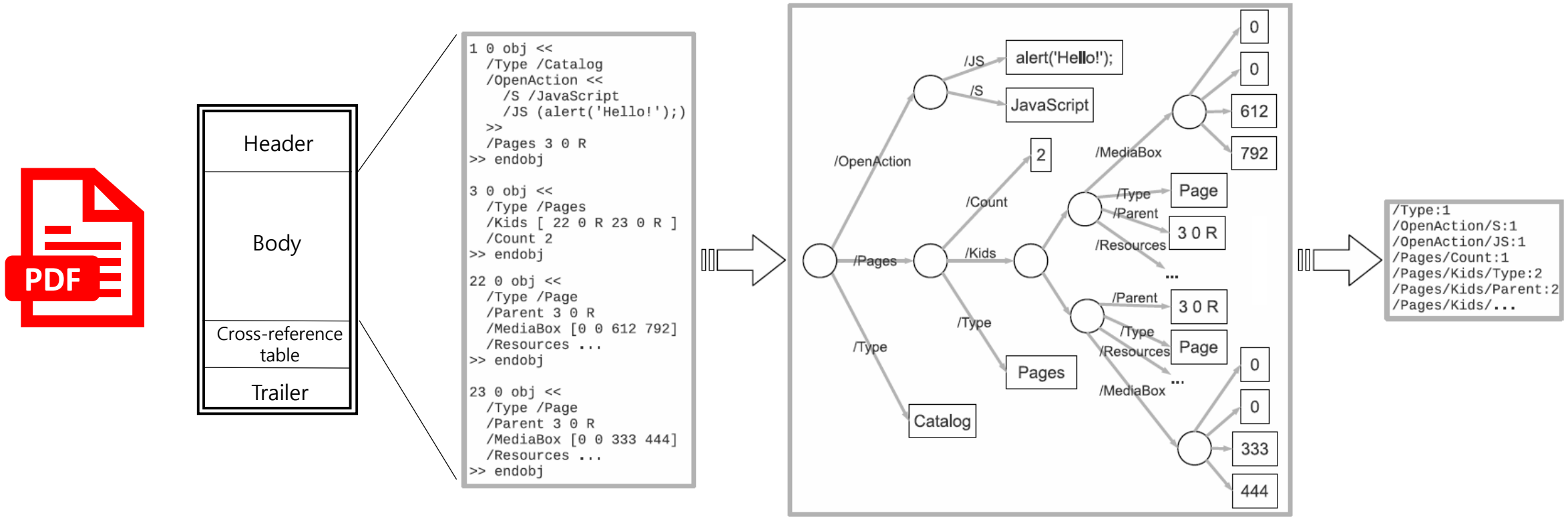
VirusTotal



선정한 안티바이러스 3개 중 1개 이상이 탐지한 파일을 악성이라고 판단!

2. 수행 내용

특징 추출 – 기존 연구 방법



[PDF 파일 문서 구조]

출처: Nedim ˇSrndi´c 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*

2. 수행 내용

특징 추출 – 기존 연구 방법

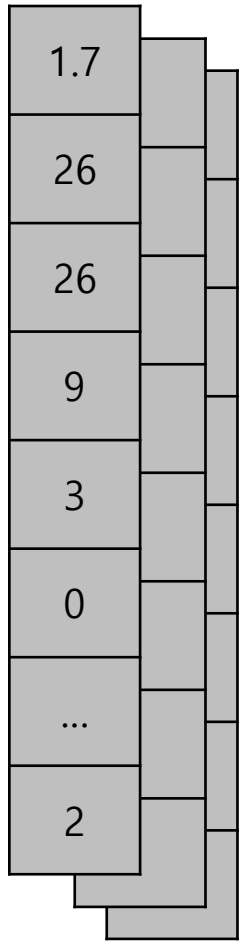
Count Tag Names

```
PDF Header: %PDF-1.7
obj                    26
endobj                 26
stream                  9
endstream               9
xref                    2
trailer                 2
startxref               2
/Page                   2
/Encrypt                0
/ObjStm                 3
/JS                     0
/JavaScript             0
/AA                     0
/OpenAction             0
/AcroForm               0
/JBIG2Decode            0
/RichMedia              0
/Launch                 0
/EmbeddedFile           0
/XFA                    0
/URI                    0
/Colors > 2^24          0
```

[pdfid.py 실행 화면 일부]

출처 : https://blog.didierstevens.com/programs/pdf-tools/

Feature Vector(FV)



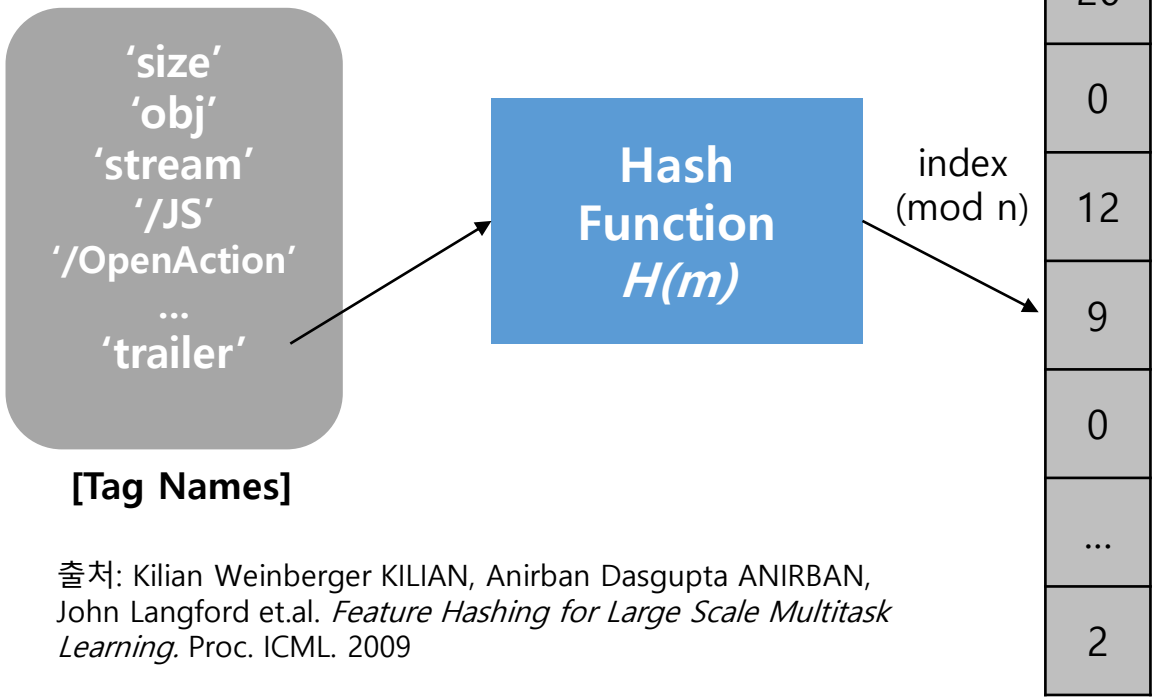
2. 수행 내용

특징 추출 - 제안 방법

PDF Header: %PDF-1.7	
obj	26
endobj	26
stream	9
endstream	9
xref	2
trailer	2
startxref	2
/Page	2
/Encrypt	0
/ObjStm	3
/JS	0
/JavaScript	0
/AA	0
/OpenAction	0
/AcroForm	0
/JBIG2Decode	0
/RichMedia	0
/Launch	0
/EmbeddedFile	0
/XFA	0
/URI	0
/Colors > 2^24	0

Add Tag Names

Feature Hashing

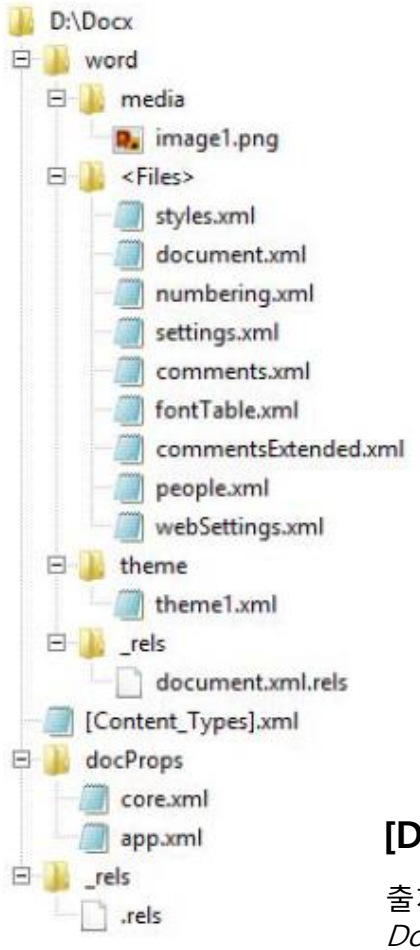


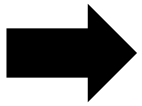
출처: Kilian Weinberger KILIAN, Anirban Dasgupta ANIRBAN, John Langford et.al. *Feature Hashing for Large Scale Multitask Learning.* Proc. ICML. 2009

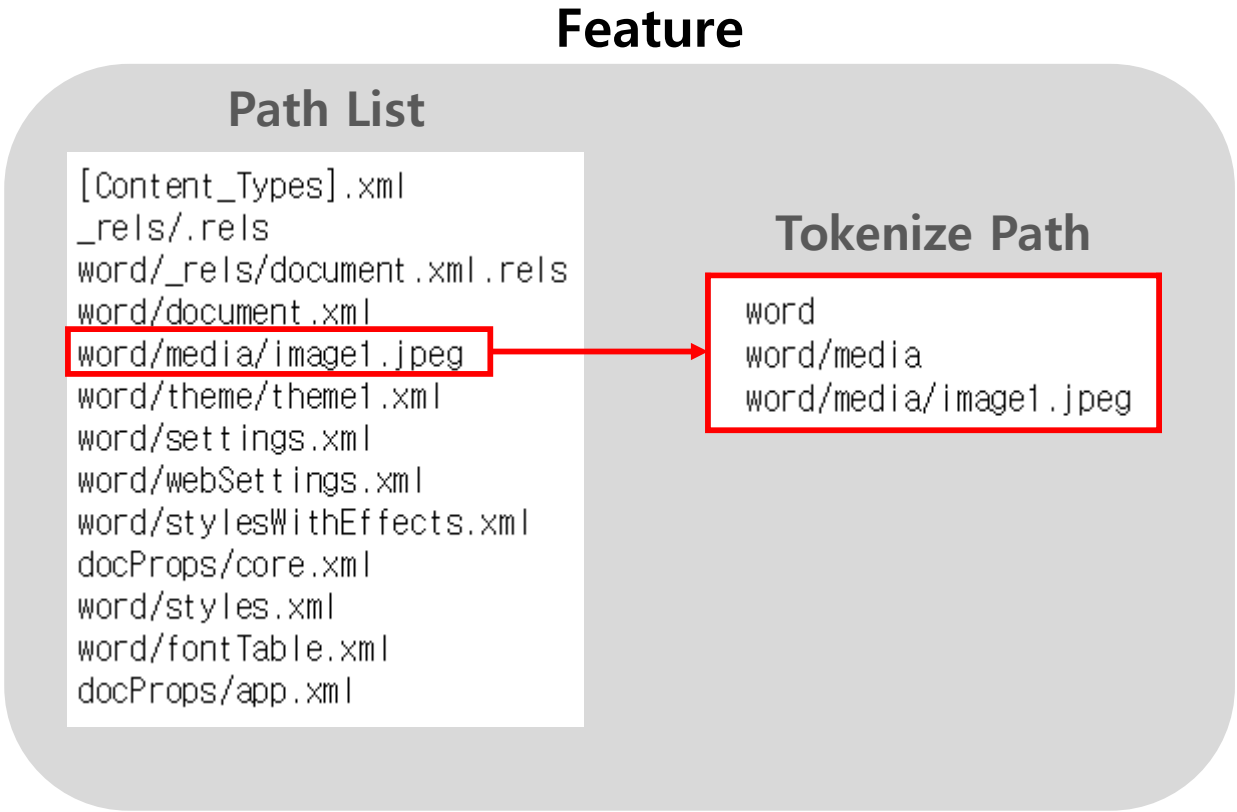
FEATURE_VECTOR[$h(tag\ name)\ mod\ SIZE\_OF\_FV$] +=Num_of_feature

2. 수행 내용

특징 추출 – 기존 연구 방법







[DOCX 파일 문서 구조]

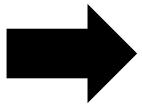
출처: Nir Nissim, Aviad Cohen, *ALDOCX: Detection of Unknown Malicious Microsoft Office Documents Using Designated Active Learning Methods Based on New Structural Feature Extraction Methodology* (n.p.: IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, n.d.), p4.

2. 수행 내용

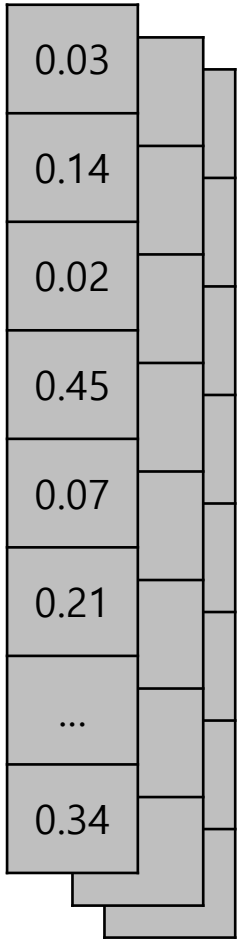
특징 추출 – 기존 연구 방법

Document Frequency

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
…	…
word/header3.xml	0.029



Feature Vector(FV)

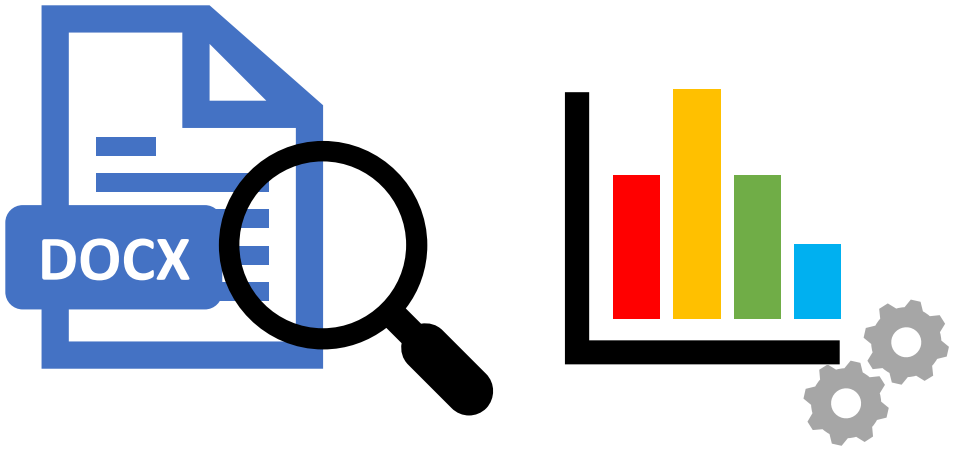


2. 수행 내용

특징 추출 - 제안 방법

DF +

DOCX 파일 내부 데이터의 min/max/mean **Entropy**
min/max/mean **File Size**



Feature Vector(FV)

0.4
0.12
0.01
0.45
0.71
0.21
...
0.34

2. 수행 내용

학습 모델 평가 지표

Accuracy

F1 score

2. 수행 내용

사용한 학습 모델

LightGBM

Linear Regression

Decision Tree

Support Vector Machine

Random Forest

Naïve Bayes

KNN

2. 수행 내용

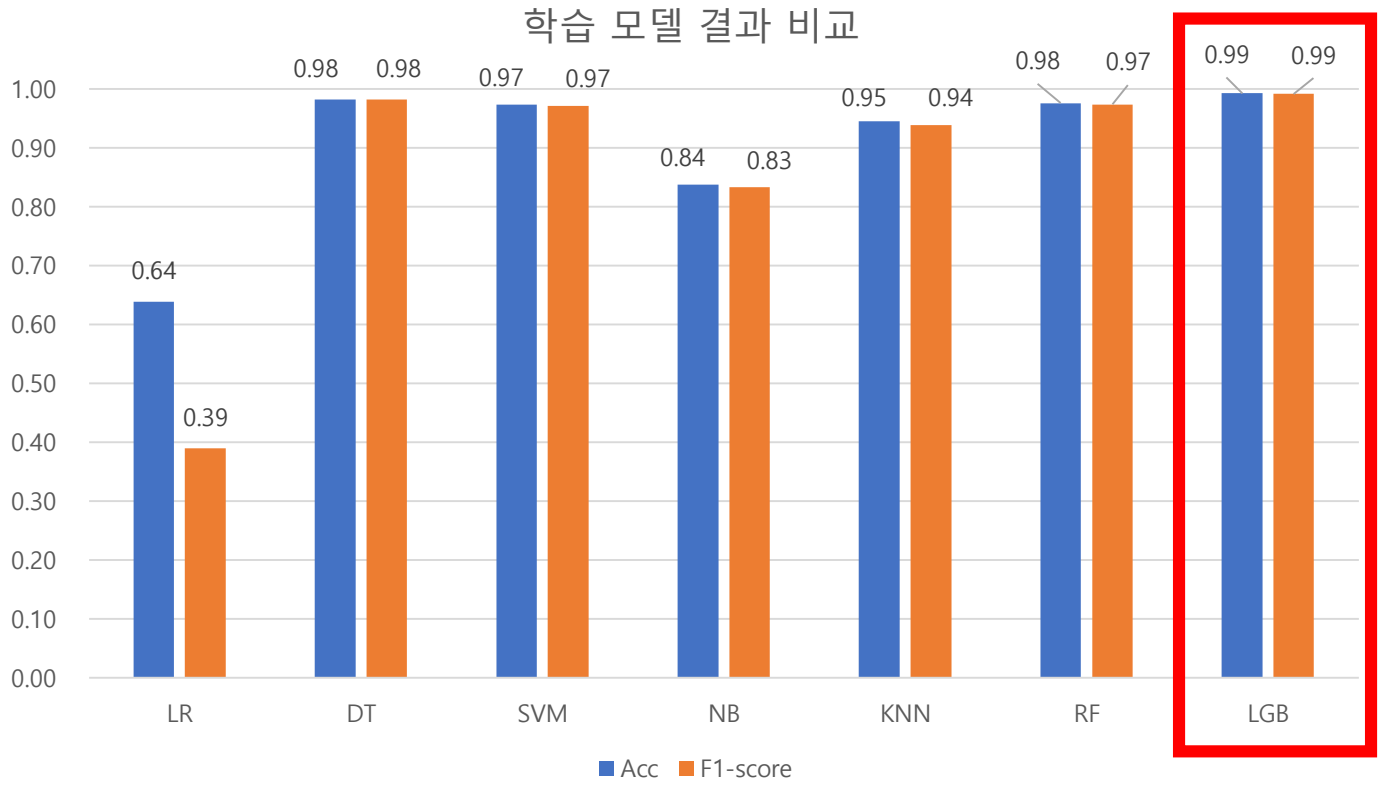
학습 모델 – PDF

학습 데이터

악성 PDF 65,000개, 정상 PDF 110,000개

검증 데이터

악성 PDF 10,000개, 정상 PDF 10,000개



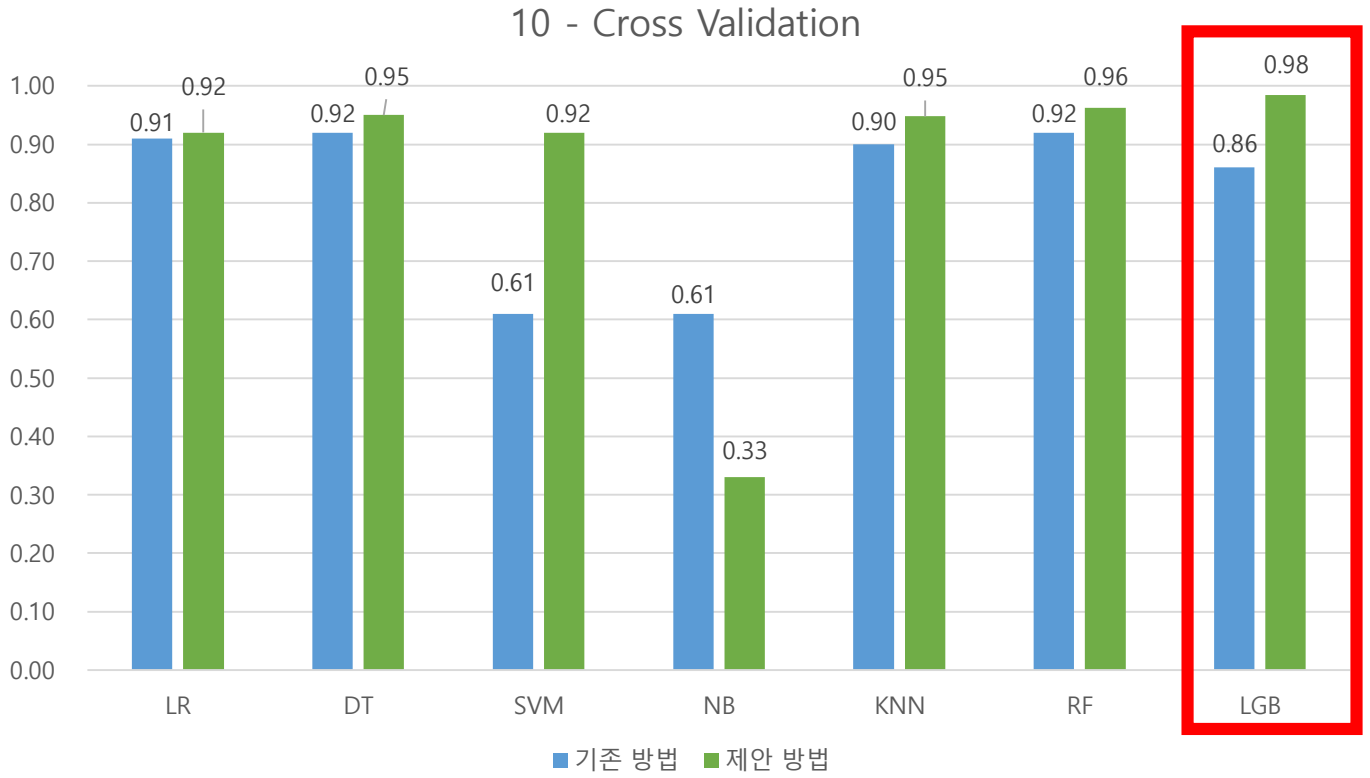
2. 수행 내용

학습 모델 – DOCX

학습 데이터

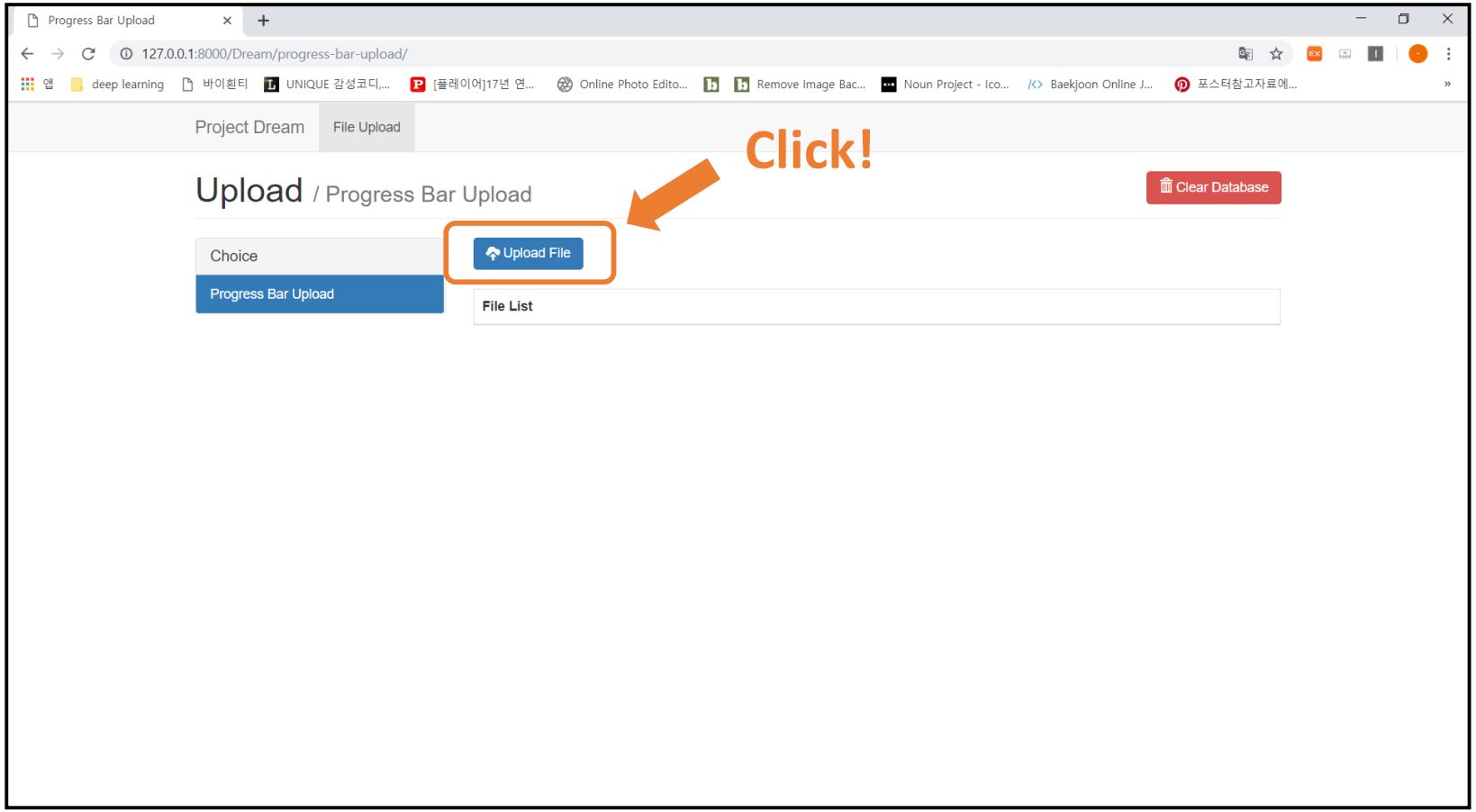
악성 DOCX 3,500개, 정상 DOCX 300개

MS Word는 수집한 데이터가 많지 않아
10-CV 을 통해 성능을 확인



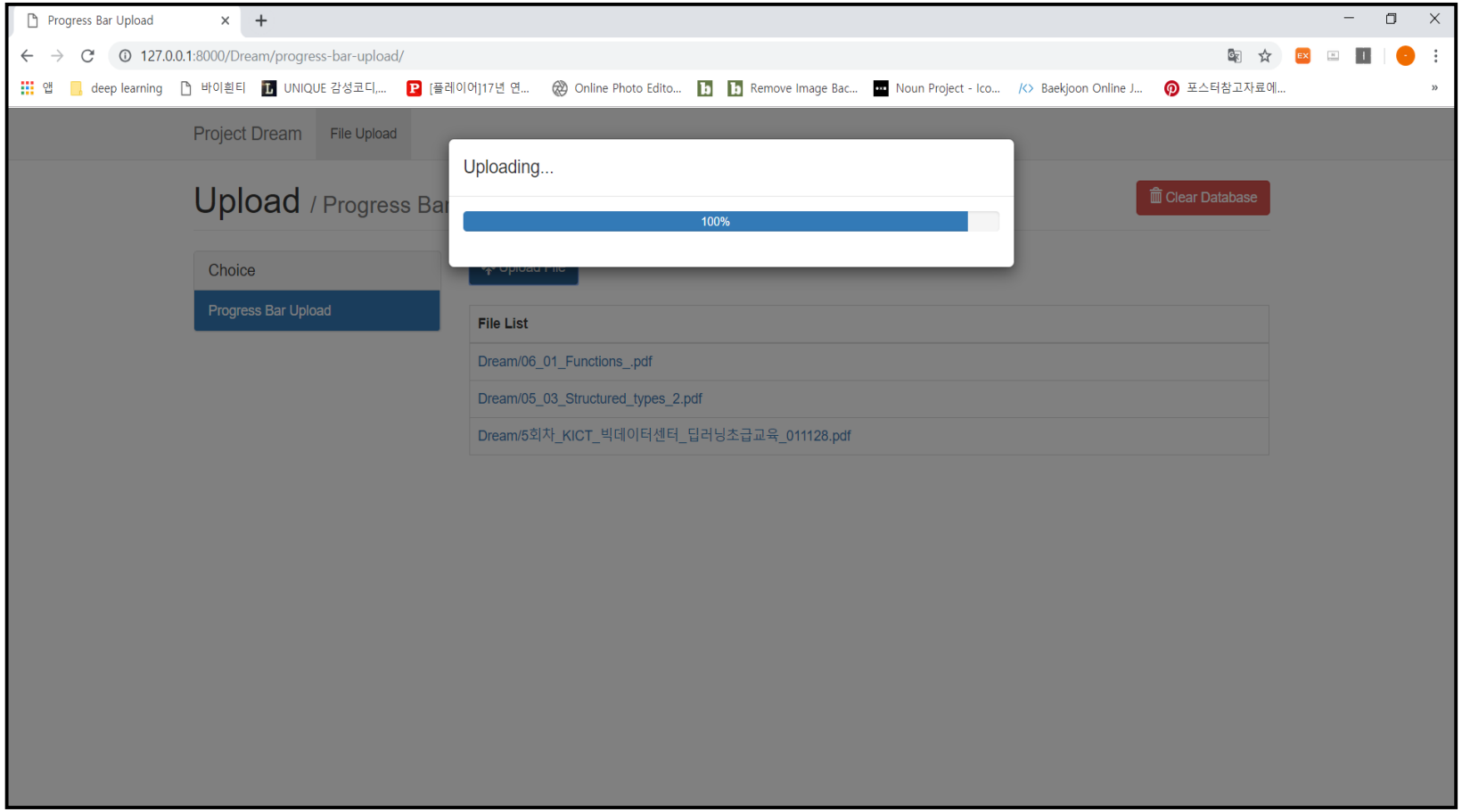
2. 수행 내용

웹 - 파일 업로드가 가능한 페이지 구현



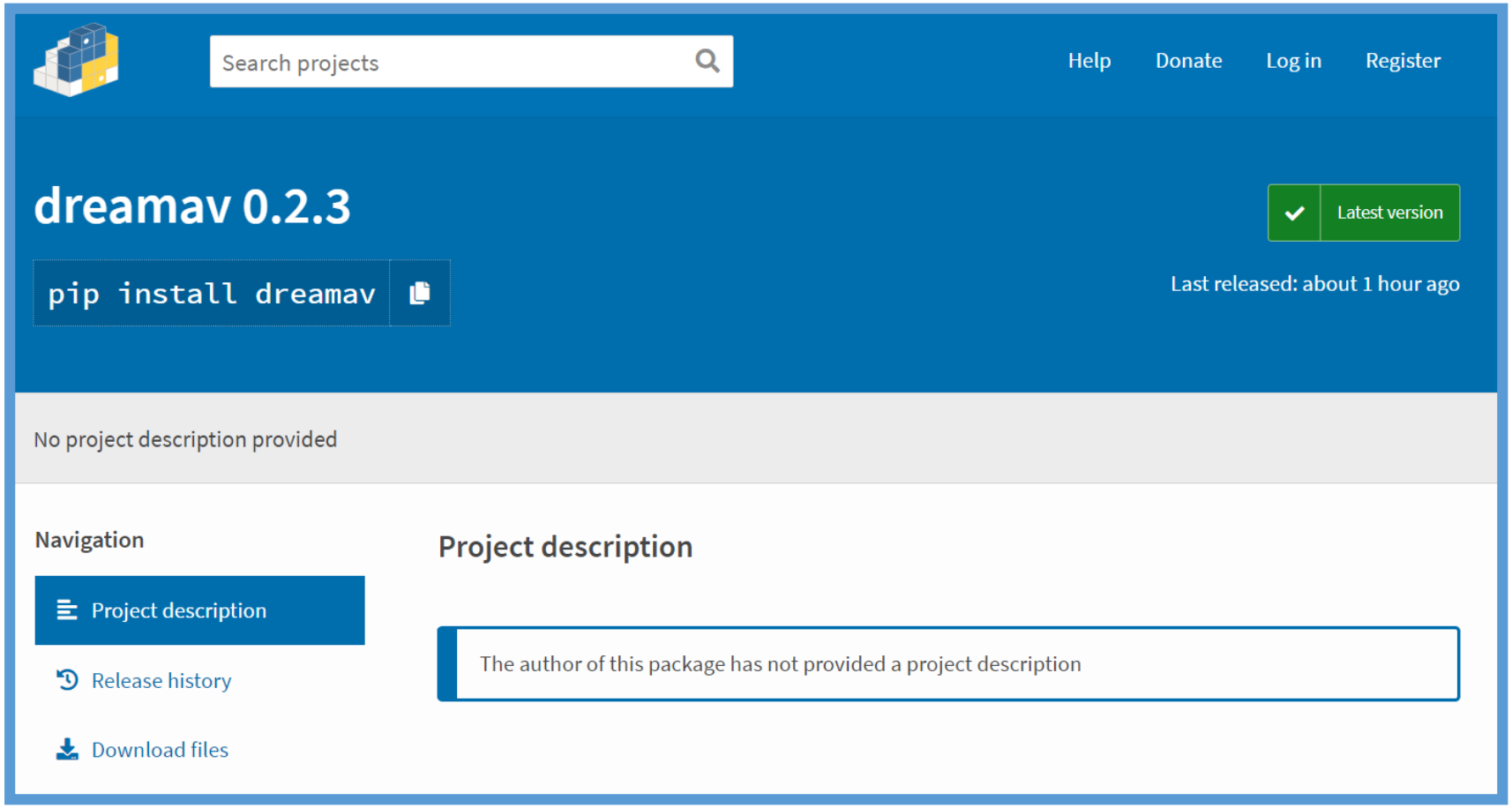
2. 수행 내용

웹 - 파일 업로드 방법



2. 수행 내용

PyPI 패키지 등록



DREAM 을 간편하게 사용할 수 있도록 PyPI를 통해 배포

Q&A

문다민 김기환 김현석
정혜리 방유한

8조 Do it !

3. 차별성

기존 기술과의 차별성

파일 업로드 가능한 환경에 연동 가능

기존 기술의 한계점 보완

3. 차별성

기존 기술과의 차별성

유사한 엔진 'ClamAV' 보다 더 정확한 탐지

[2018년도 문서형 악성 코드 500개]

4. 향후 계획

1. 데이터 수집

기계학습에 사용하기 위해
더 많은 PDF, DOCX 파일 수집

2. 데이터 전처리

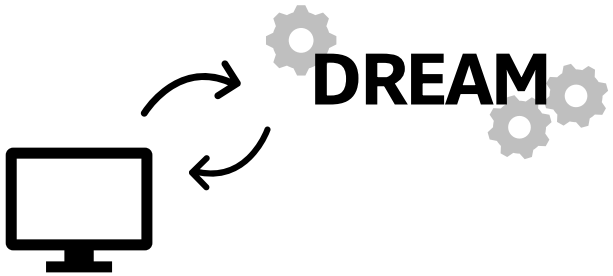
자바스크립트, 매크로
등 추가적인 특징 연구
및 추출

3. 학습 모델

딥러닝 및 앙상블 모델
구현

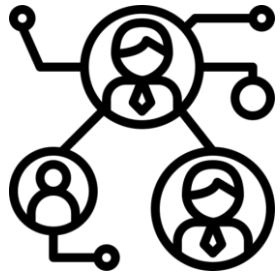
4. 향후 계획

4. 웹 서버 연동



기계 학습에 사용하기 위한 더 많은 PDF, DOCX 파일 수집

5. 오픈소스 커뮤니티 형성



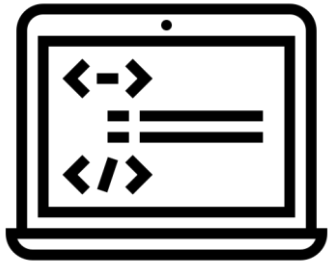
Slack 등 오픈소스 사용자들 간의 네트워크를 형성 할 수 있도록 커뮤니티 구성

4. 향후 계획

6. 파일 요청 API 추가

파일 검사 사이트 ‘바이러스 토탈’에 파일을 요청 할 수 있도록 API 추가

7. 파일 공유 사이트 구현



사용자들이 문서형 악성코드 파일을 공유 할 수 있는 사이트 구현