



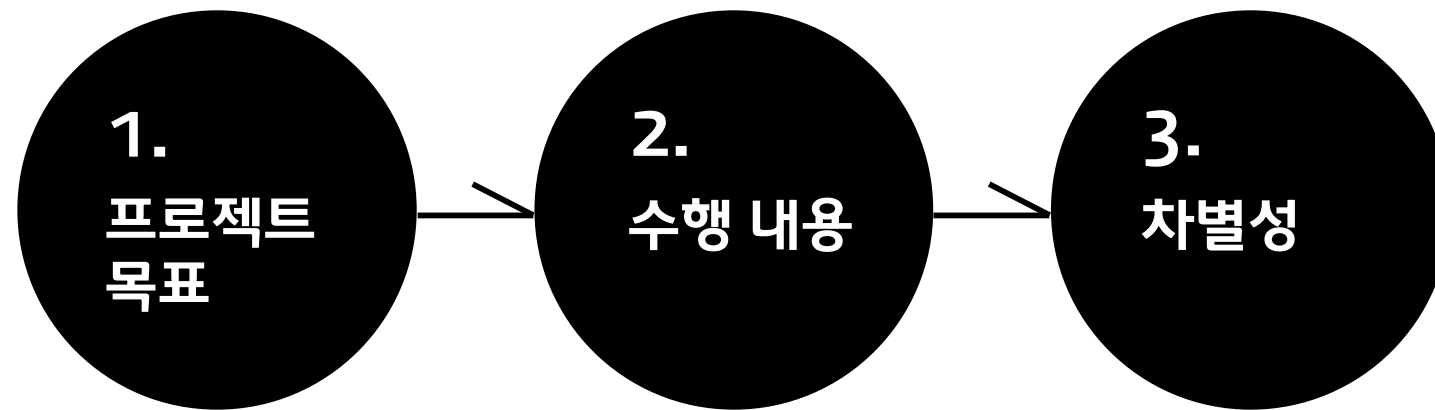
DREAM



문다민 김기환 김현석
정혜리 방유한

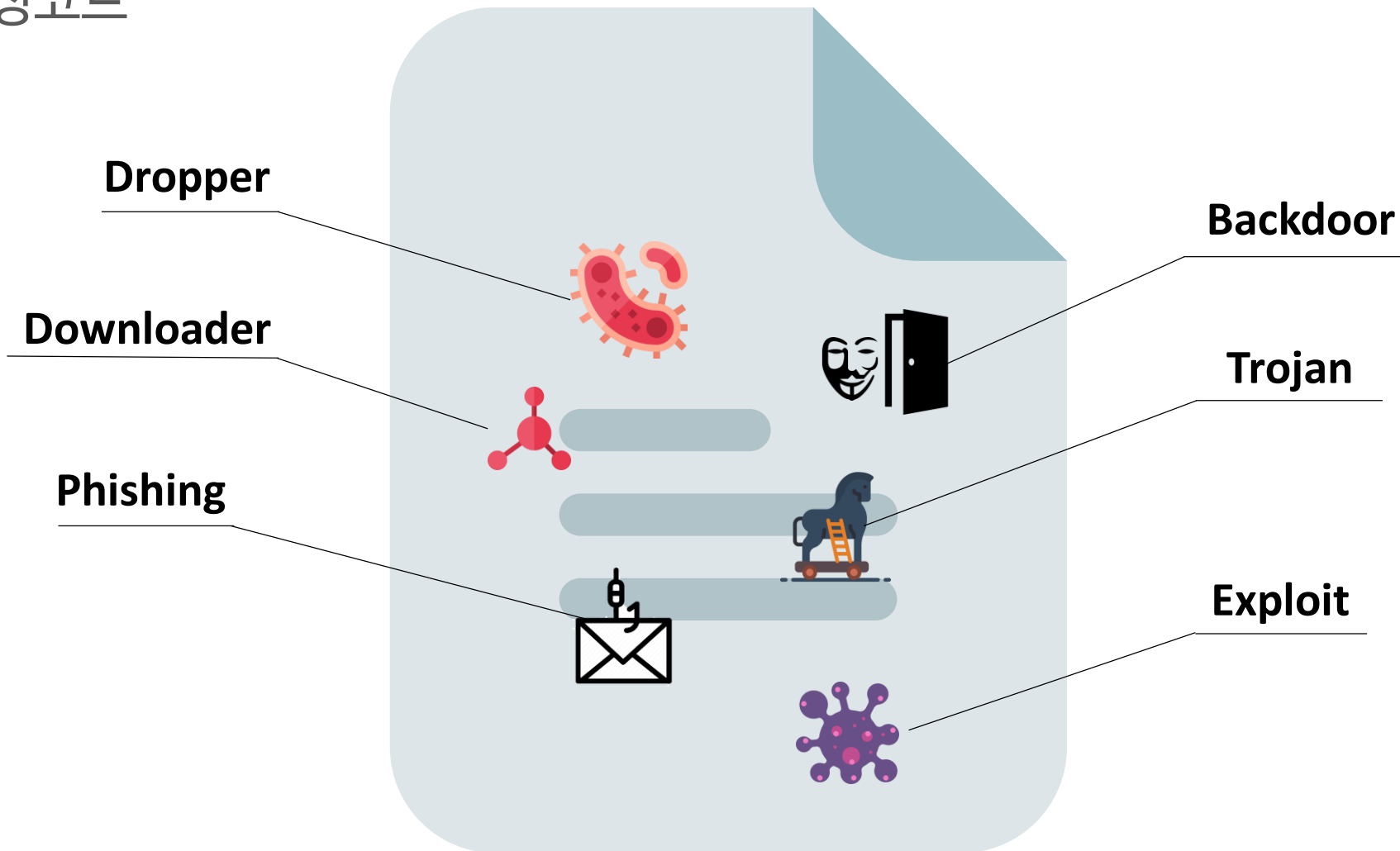
8조 Do it !

— 목차



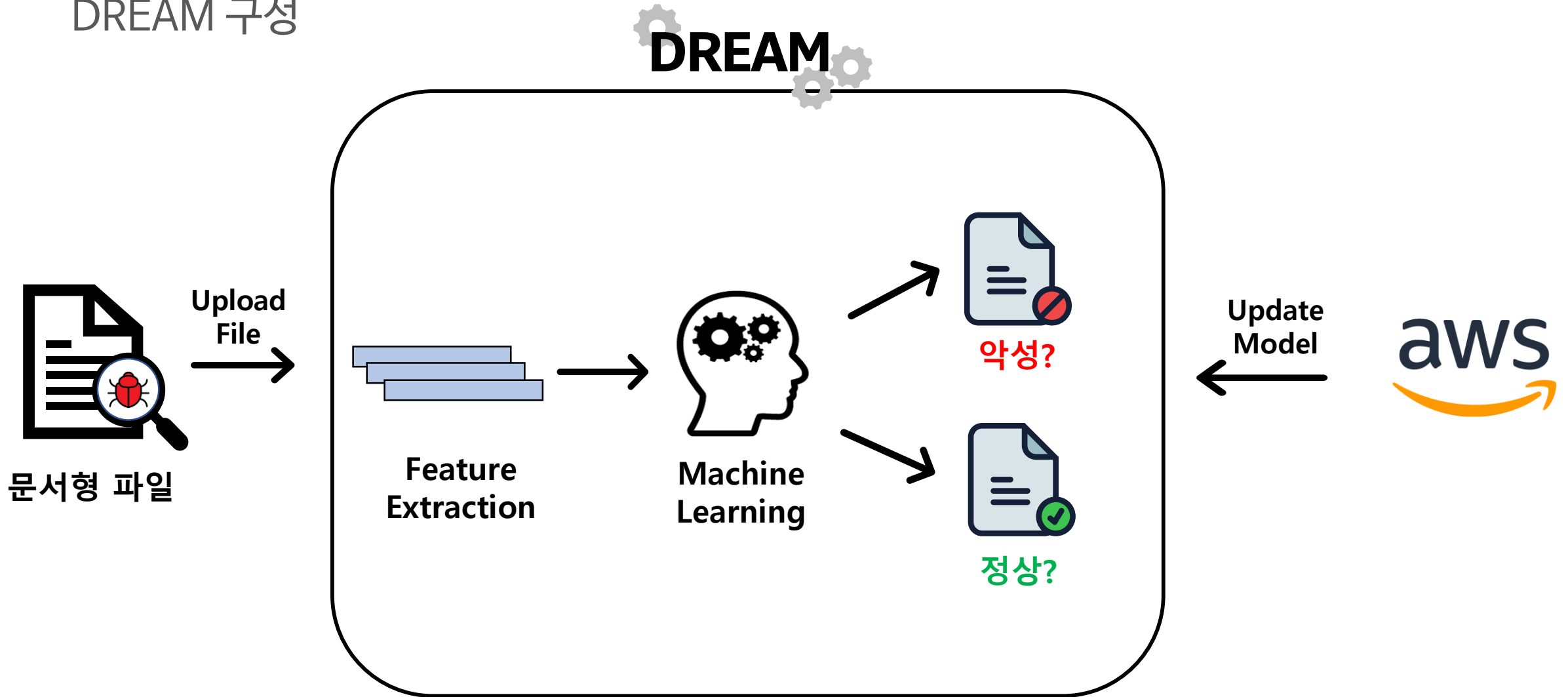
1. 프로젝트 소개

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 개발함
으로써 엔진 발전에 도움



3. 높은 확장성

메일 서버, 웹 서버 등에 엔진
사용 가능

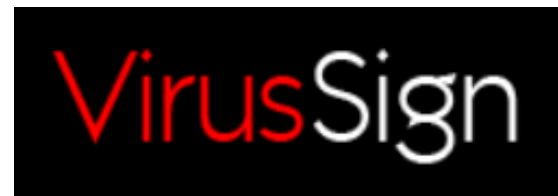


2. 수행 내용

데이터 수집



**Virus
Share**



2. 수행 내용

데이터 라벨링



KASPERSKY 



F-Secure 

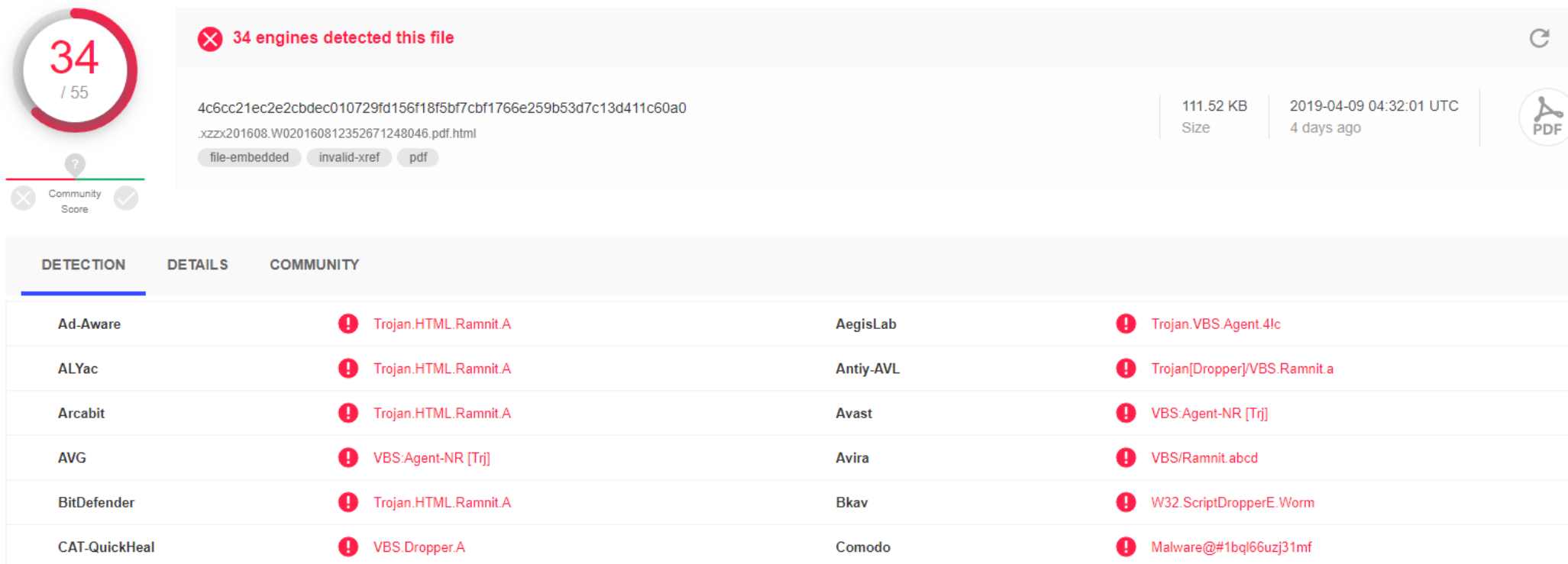


 Symantec™

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

2. 수행 내용

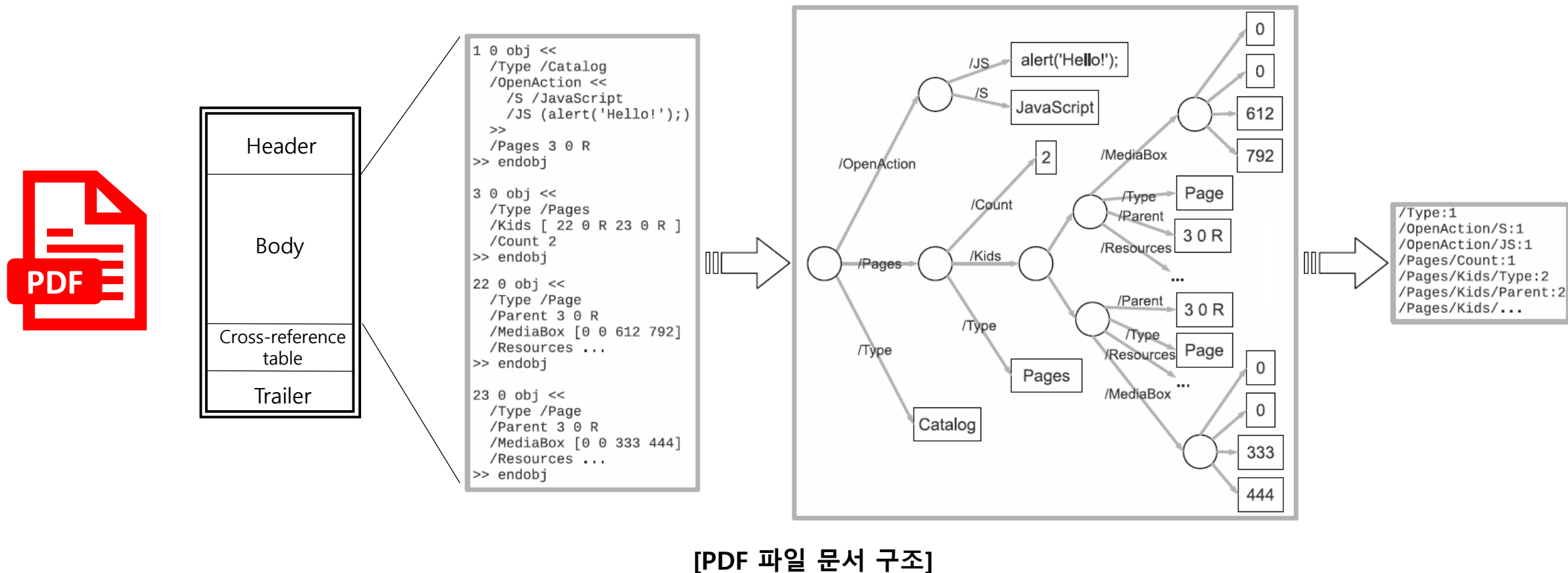
데이터 라벨링



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

2. 수행 내용

특징 추출 - 기존 연구 방법



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*

2. 수행 내용

특징 추출 - 기존 연구 방법

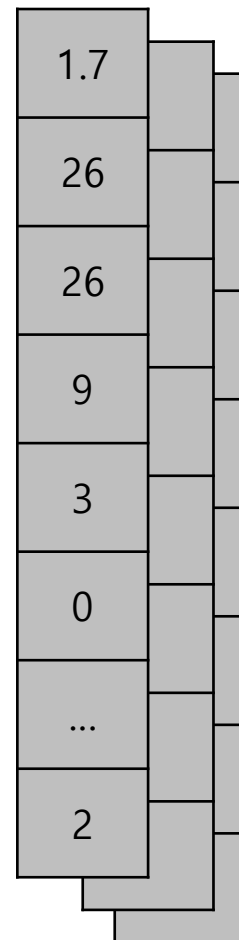
Count Tag Names

```
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 22
trailer 22
startxref 22
/Page 22
/Encrypt 00
/ObjStm 3
/JS 00
/JavaScript 00
/AA 00
/OpenAction 00
/AcroForm 00
/JBIG2Decode 00
/RichMedia 00
/Launch 00
/EmbeddedFile 00
/XFA 00
/URI 00
/Colors > 2^24 0
```

Add Tag Names



'size'
'obj'
'stream'
'/JS'
'/OpenAction'
...
'trailer'

[Tag Names]

Feature Hashing

Hash
Function
 $H(m)$

index
(mod n)

Feature
Vector(FV)

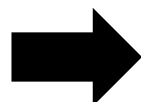
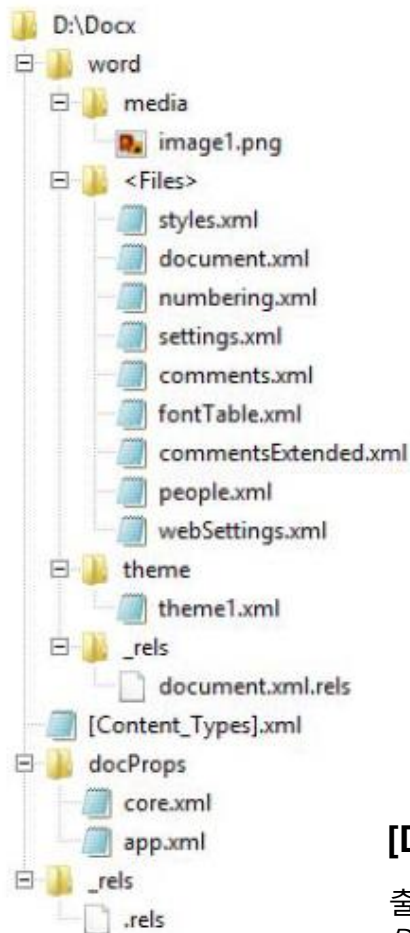
1.7
26
0
12
9
0
...
2

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

$\text{FEATURE_VECTOR}[h(\text{tag name}) \bmod \text{SIZE_OF_FV}] += \text{Num_of_feature}$

2. 수행 내용

특징 추출 - 기존 연구 방법



Feature

Path List

```
[Content_Types].xml
_rels/.rels
word/_rels/document.xml.rels
word/document.xml
word/media/image1.jpeg
word/theme/theme1.xml
word/settings.xml
word/webSettings.xml
word/stylesWithEffects.xml
word/webSettings.xml
word/styles.xml
word/fontTable.xml
docProps/app.xml
```

Tokenize Path

```
word
word/media
word/media/image1.jpeg
```

[DOCX 파일 문서 구조]

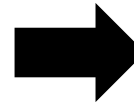
출처: Nir Nissim, Aviad Cohen, *ALDOX: Detection of Unknown Malicious Microsoft Office Documents Using Designated Active Learning Methods Based on New Structural Feature Extraction Methodology* (n.p.: IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, n.d.), p4.

2. 수행 내용

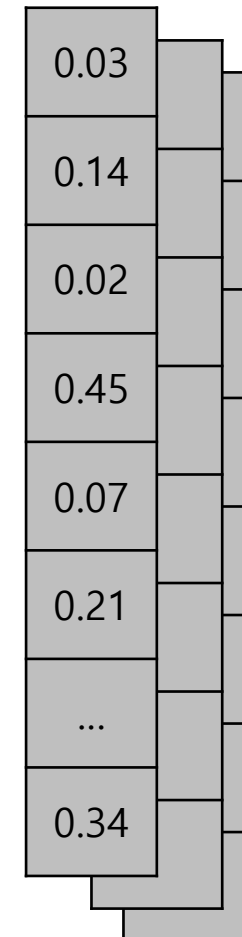
특징 추출 - 기존 연구 방법

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)

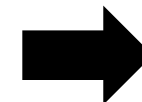
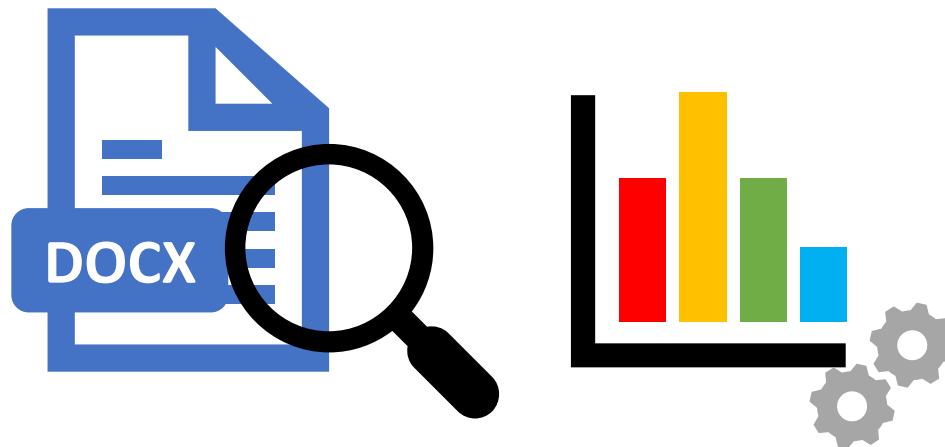


2. 수행 내용

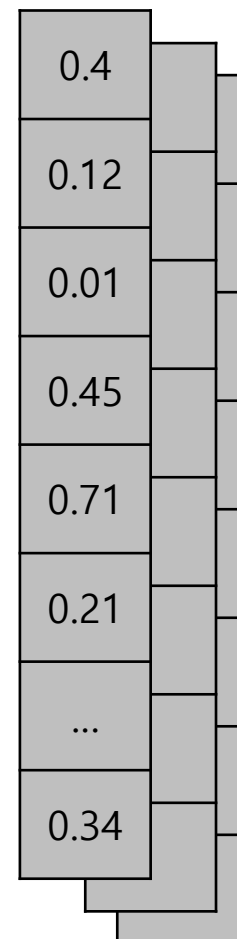
특징 추출 - 제안 방법

DF +

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



Feature
Vector(FV)



2. 수행 내용

학습 모델 평가 지표



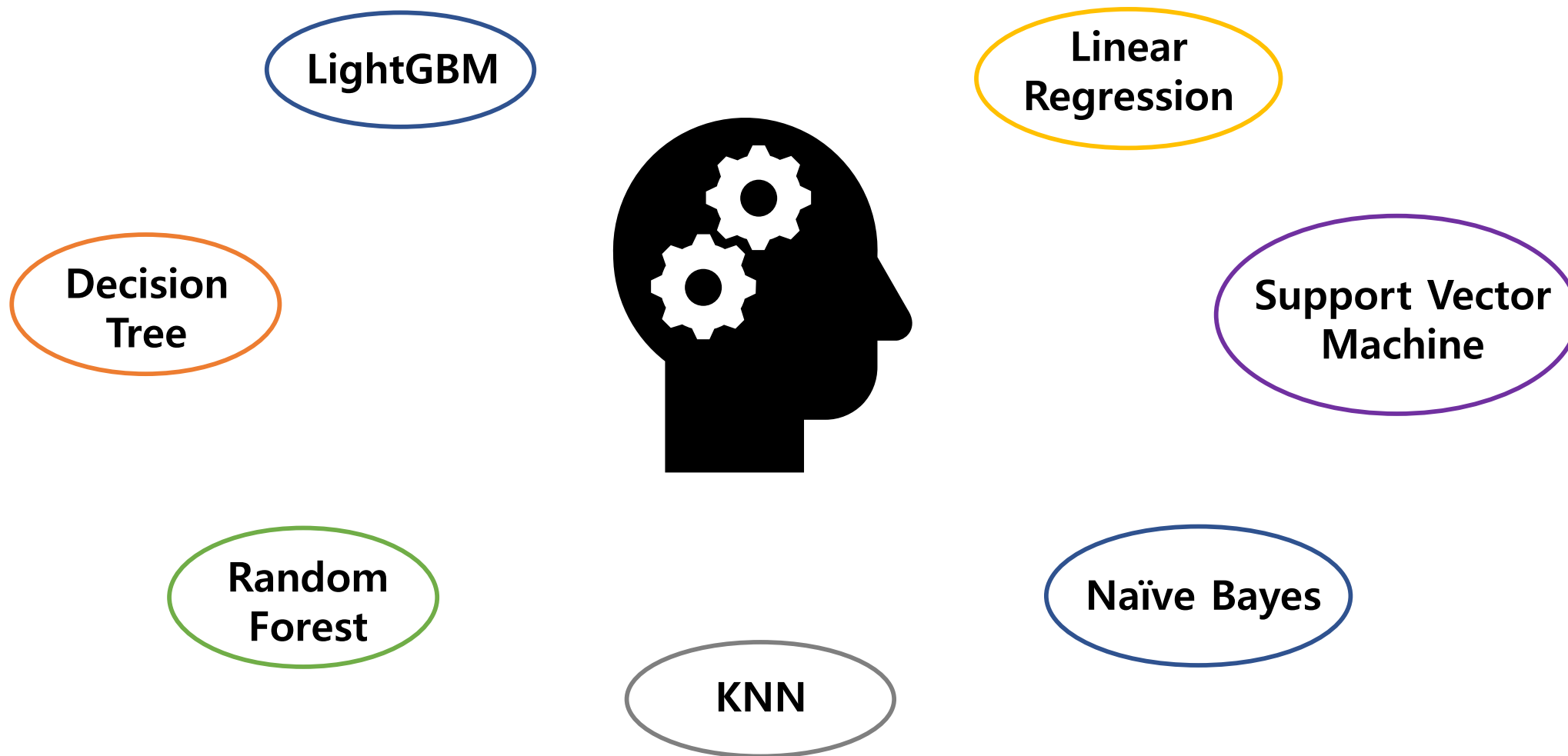
Accuracy



F1 score

2. 수행 내용

사용한 학습 모델



2. 수행 내용

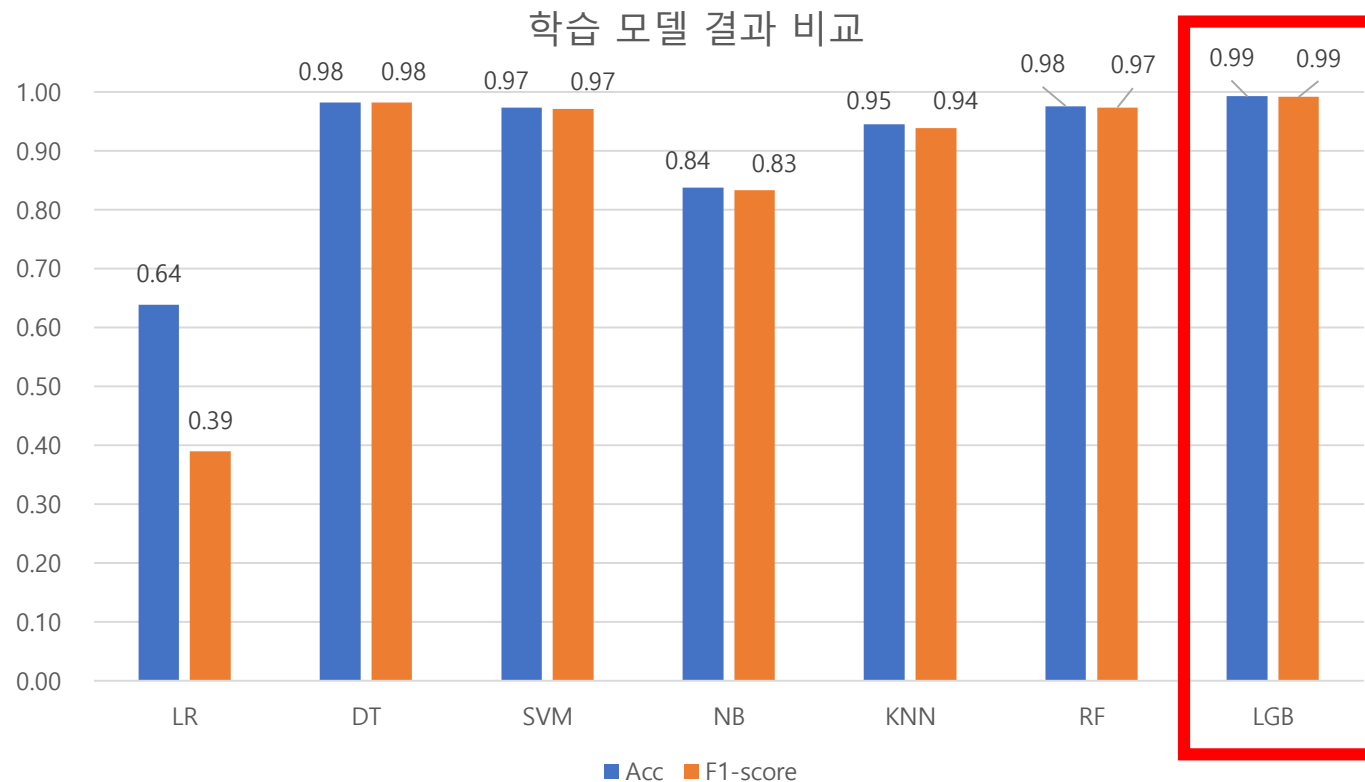
학습 모델 - PDF

학습 데이터

악성 PDF 65,000개, 정상 PDF 110,000개

검증 데이터

악성 PDF 10,000개, 정상 PDF 10,000개



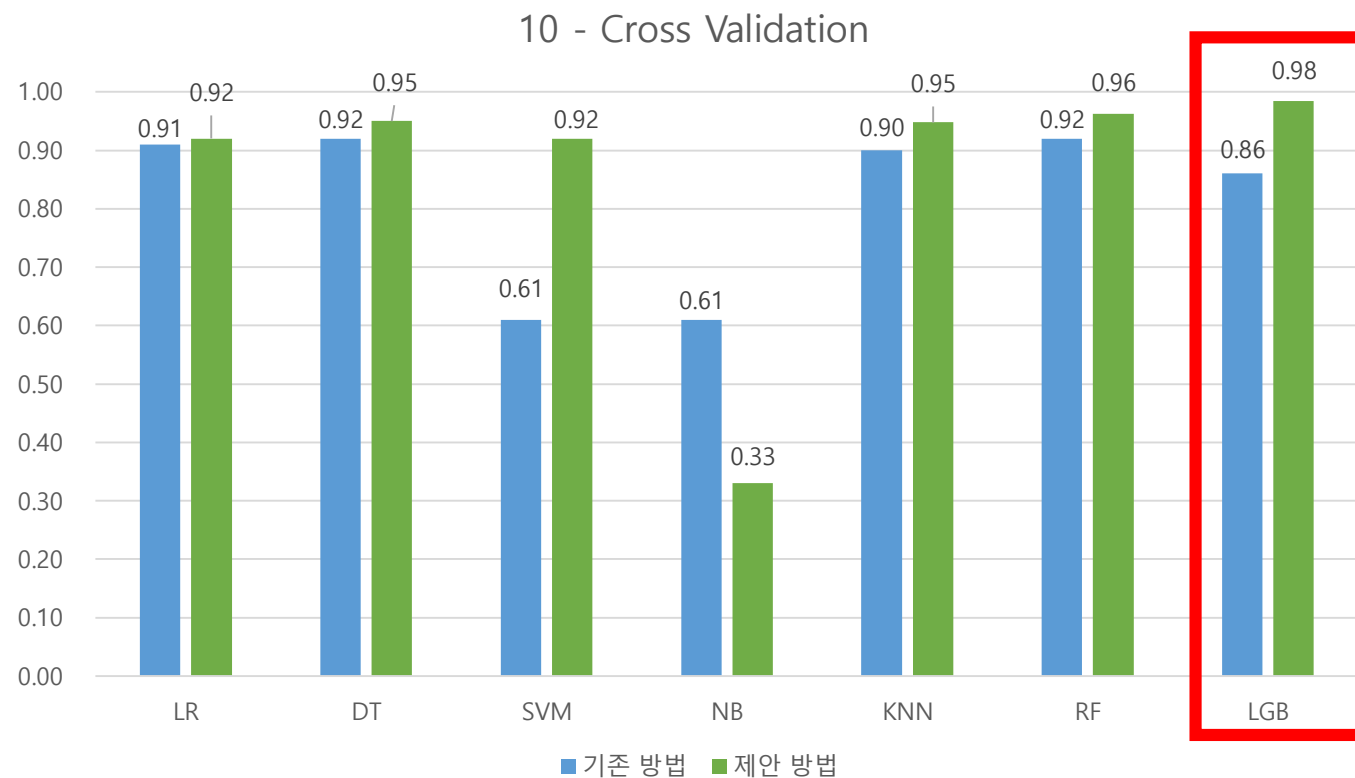
2. 수행 내용

학습 모델 - DOCX

학습 데이터

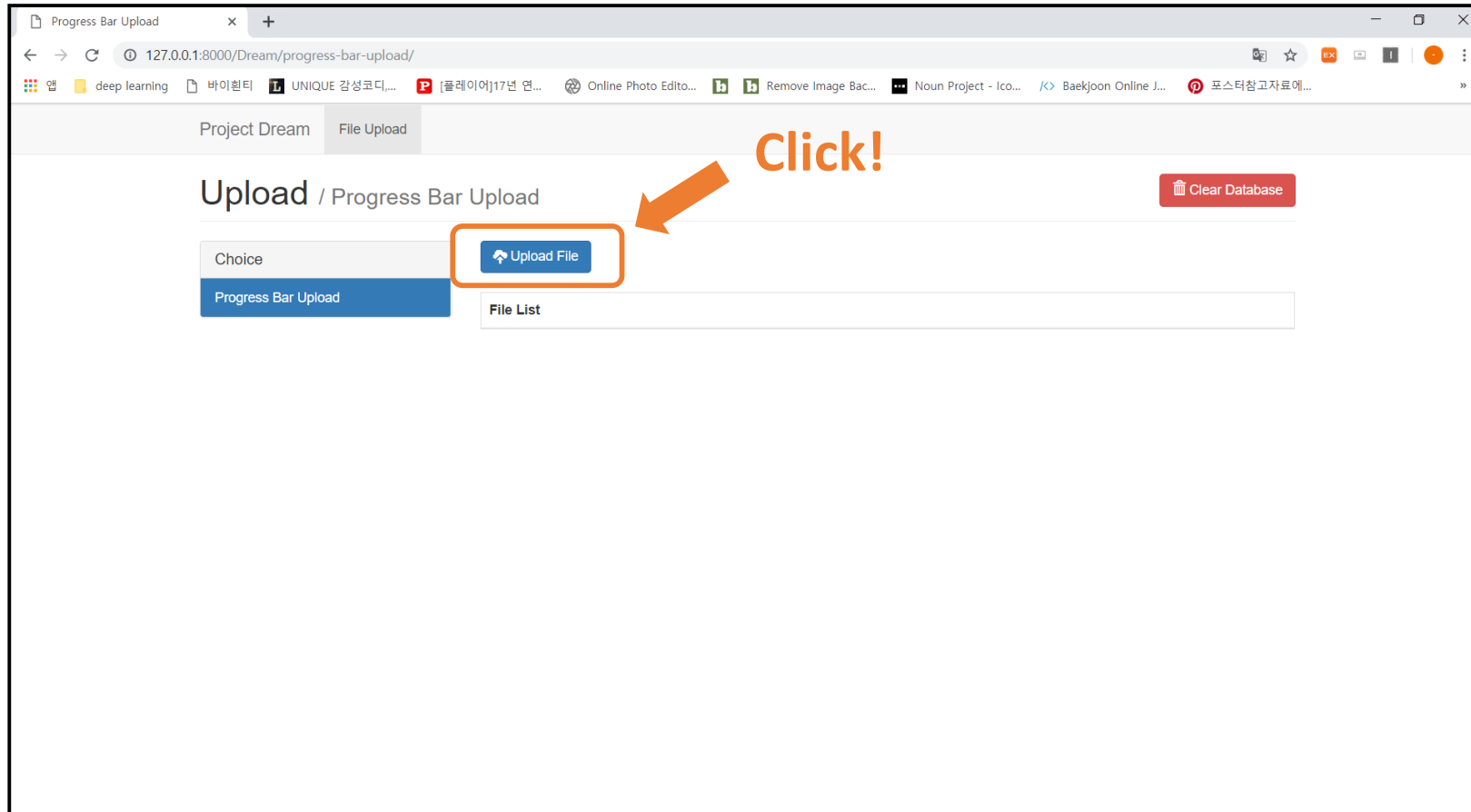
악성 DOCX 3,500개, 정상 DOCX 300개

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



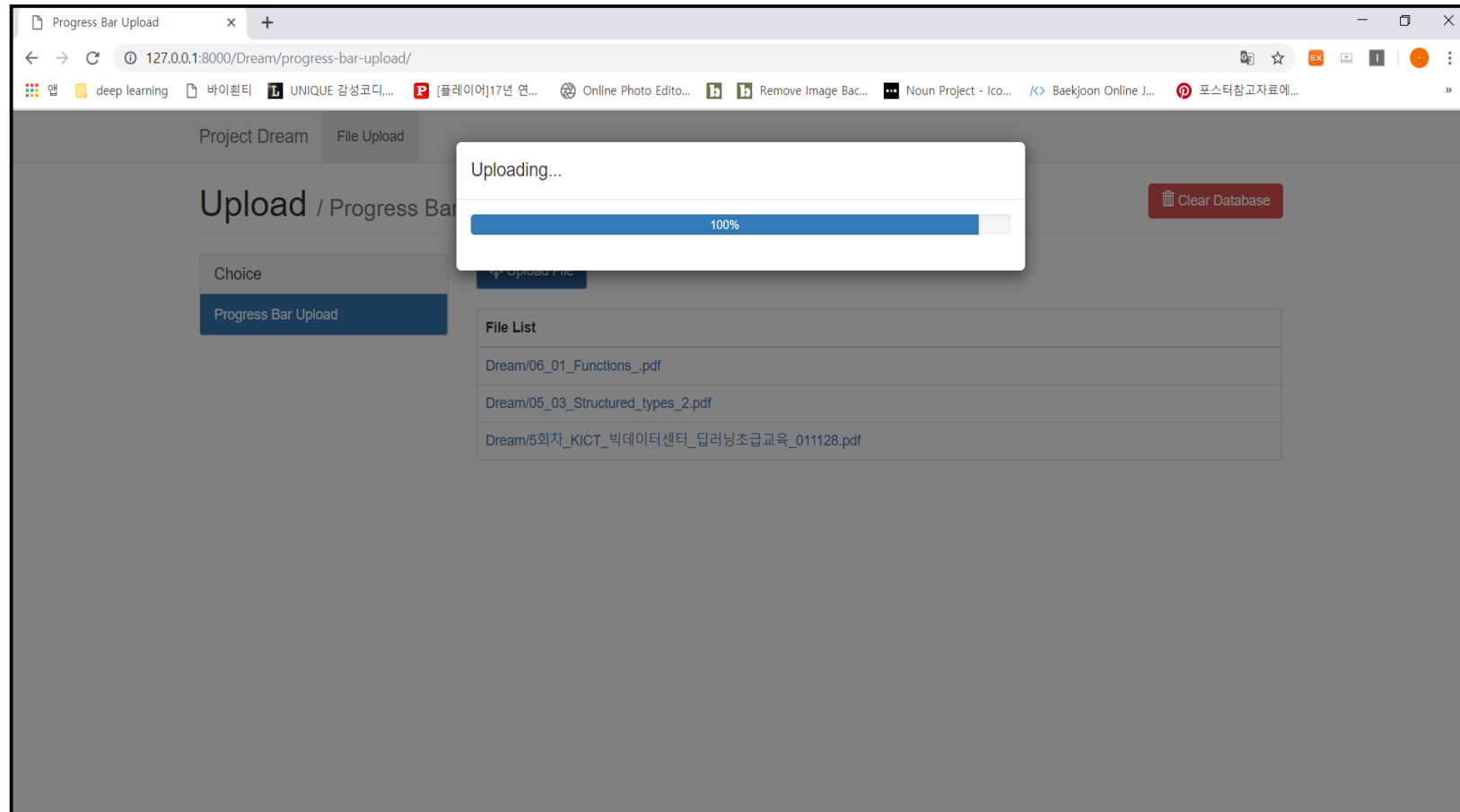
2. 수행 내용

웹 - 파일 업로드가 가능한 페이지 구현



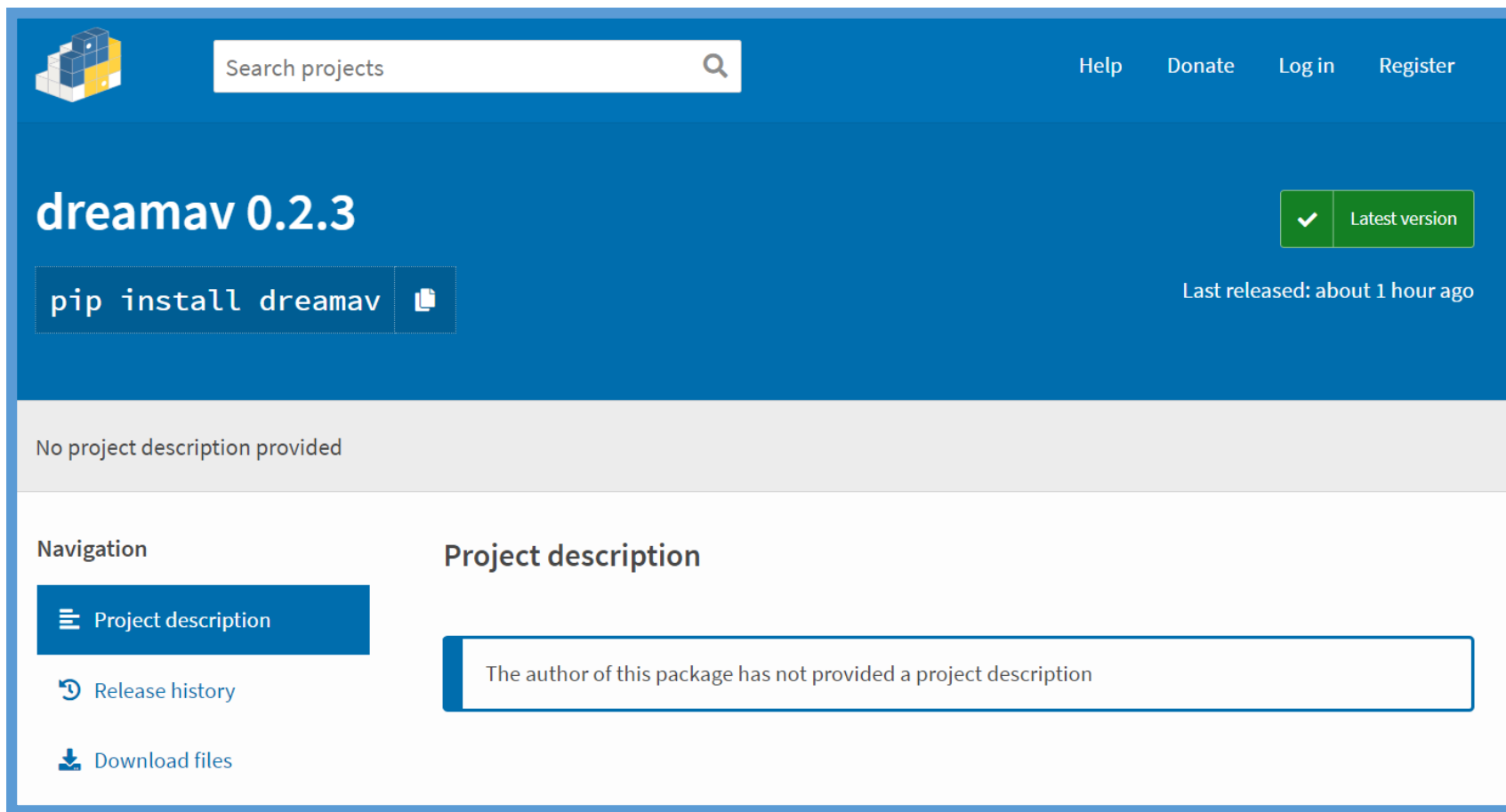
2. 수행 내용

웹 - 파일 업로드 방법



2. 수행 내용

PyPI 패키지 등록



DREAM 을 간편하게 사용할 수 있도록 PyPI를 통해 배포

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]



Q&A

문다민 김기환 김현석
정혜리 방유한

8조 Do it !