

# 캡스톤 디자인 I

## 종합설계 프로젝트

프로젝트 명	DREAM(Detecting in Real-time mAlicious document using Machine Learning)
팀 명	<i>Do it!</i>
문서 제목	수행 계획서

Version	1.5
Date	2019-3-13

팀원	문 다민(팀장)
	김 기환
	김 현석
	정 혜리
	방 유한(외국인)

	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time Malicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

#### CONFIDENTIALITY/SECURITY WARNING

이 문서에 포함되어 있는 정보는 국민대학교 소프트웨어융합대학 소프트웨어학부 개설 교과목 캡스톤 디자인Ⅰ 수강 학생 중 프로젝트 **DREAM**(Detecting in Real-time Malicious document using Machine Learning)를 수행하는 팀 **Do it!** 팀원들의 자산입니다. 국민대학교 소프트웨어학부 및 팀 **Do it!**의 팀원들의 서면 허락없이 사용되거나, 재가공 될 수 없습니다.

## 문서 정보 / 수정 내역

<b>Filename</b>	계획서
<b>원안작성자</b>	문다민 김현석 김기환 정혜리
<b>수정작성자</b>	문다민 김현석 김기환 정혜리 방유한

수정날짜	대표수정자	Revision	추가/수정 항목	내 용
2019-03-03	문다민	1.0	최초 작성	프로젝트 개요 작성 역할 분담 초안 작성 개발 목표 초안 작성
2019-03-03	정혜리	1.1	내용 추가	개발 배경 및 필요성 작성
2019-03-05	김현석	1.2	내용 추가	개발 목표 작성
2019-03-08	김기환	1.3	내용 추가	개발 결과 작성 시스템 구조도 작성
2019-03-09	문다민	1.4	내용 추가	연구/개발 방법 작성
2019-03-11	방유한	1.4.1	내용 수정	연구 방법 수정
2019-03-13	문다민	1.5	최종	최종 수정 및 작성



계획서		
프로젝트 명	DREAM(Detecting in Real-time mAlicious document using Machine Learning)	
팀 명	Do it!	
Confidential Restricted	Version 1.5	2019-MAR-13

## 목 차

1	개요	5
1.1	프로젝트 개요	5
1.2	추진 배경 및 필요성	6
1.2.1	추진 배경	6
1.2.2	현재 기술 시장 현황	8
1.2.3	현재 기술 시장의 문제점 및 개선 방향	11
1.2.3.1	기술 시장 문제 1	11
1.2.3.2	기술 시장 문제 2	12
1.2.3.3	기술 시장 문제 3	12
2	개발 목표 및 내용	13
2.1	목표	13
2.2	연구/개발 내용	13
2.2.1	시연 시나리오	13
2.2.2	연구/개발 방법	14
2.2.2.1	데이터 라벨링 (Data Labeling)	14
2.2.2.2	특징 추출	15
2.2.2.3	기계 학습	16
2.2.2.4	웹	17
2.3	개발 결과	18
2.3.1	시스템 기능 요구사항	18
2.3.2	시스템 비기능(품질) 요구사항	18
2.3.3	시스템 구조	19
2.3.4	결과물 목록 및 상세 사양	20
2.4	기대효과 및 활용방안	20
2.4.1	기대효과	20
2.4.2	활용방안	20
3	배경 기술	21
3.1	기술적 요구사항	21
3.2	현실적 제한 요소 및 그 해결 방안	22
3.2.1	하드웨어	22
3.2.2	소프트웨어	22
3.2.3	기타	22
4	프로젝트 팀 구성 및 역할 분담	23
5	프로젝트 비용	23
6	개발 일정 및 자원 관리	24

	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time malicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

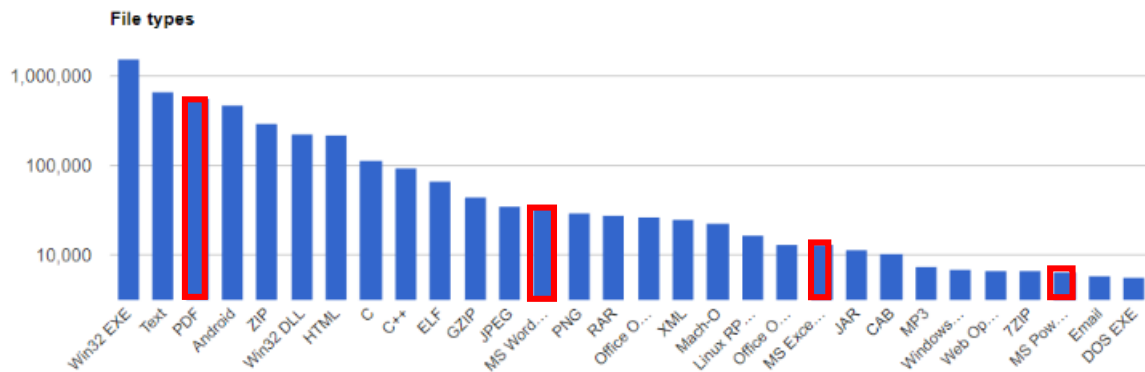
6.1	개발 일정 .....	24
6.2	일정별 주요 산출물.....	25
6.3	인력자원 투입 계획 .....	26
6.4	비 인력자원 투입 계획 .....	26
7	참고 문헌.....	27

	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time E mAlicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

# 1 개요

## 1.1 프로젝트 개요

정보화 시대에 들어서며 전 세계적으로 악성코드의 수는 급격히 늘어나고 있다. 최근에는 악성 코드 중 문서형 악성코드의 수가 증가하고 있고, 특히 이 문서형 악성코드의 유포 방법이 화두가 되고 있다. 대표적으로 2018년 1월부터 등장한 갠드크랩(GandCrab) 랜섬웨어는 사람들의 이목을 끌 수 있는 문서 파일로 위장하여 유포된다. 그리고 2018년 3월에 전 세계적으로 등장한 시그마(Sigma) 랜섬웨어는 이력서로 위장하여 유포된다. 이처럼 문서형 악성코드는 점점 지능적이고 정교하게 발전하고 있다.



<그림 1> 일주일 동안 바이러스토탈에 유입되는 파일의 종류와 수

(출처 = <https://www.virustotal.com/en/statistics/>)

바이러스토탈(VirusTotal)은 의심스러운 파일 및 URL을 분석하고 모든 종류의 악성 코드를 탐지하는 서비스이다. <그림 1> 은 일주일 동안 유입된 파일의 유형별 수에 대한 바이러스토탈의 그래프이다. 그래프에 따르면 PDF는 55만 개로 3번째를 차지하고 있으며 이 외에도 MS Word, MS Excel 등 우리가 자주 사용하는 문서형 파일이 속해있다.

한 기사에 따르면 작년 10월, '국가 핵심 인력 등록 관리제 등 검토 요청.hwp'라는 파일명으로 문서형 악성코드가 유포되어 국내에서 실제 감염 피해가 발생한 바 있다. 바이러스토탈에 이 문서형 악성코드가 처음 업로드 된 날짜는 10월 23일이었지만 10월 24일에 57개의 안티바이러스 중 3개 안티바이러스만이 탐지하였으며 이어 25일에는 6개 안티바이러스, 26일에는 13~16개 안티바이러스, 11월 2일에는 56개의 안티바이러스 중 절반가량인 25개 안티바이러스에서만 탐지되었다. 즉 상당수의 안티바이러스들이 이 문서형 악성코드를 탐지하지 못하였다.

	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time MAlicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

보안뉴스 · **사이버보안**이 발간하는  
2019 국내외 보안시장 전망보고서

**보안뉴스**

로그인 | 회원가입 | 기사제보 | 사이트맵

f t N P

통합검색


≡

#전체기사 #시큐리티월드 #사건사고 #4차산업혁명 #세계보안엑스포(SECON)

동영상

카드뉴스

콘텐츠



Home > 전체기사

## 국내 유명 변호사 사칭한 악성코드, 상당수 백신 탐지 못해

좋아요 53개 | 입력: 2018-11-04 08:40

#백신 #악성코드 #유명 변호사 #사이버공격

10월 23일 발견 악성코드, 24일 57개 백신중 3개만 탐지...11월 2일 절반 탐지  
아직 상당수 백신들 탐지 못해...고도화되는 공격에 백신의 탐지력 더욱 높아야

[보안뉴스 김경애 기자] 최근 국내 유명 변호사 이름을 사칭한 북한 추정 사이버공격이 이슈가 된 가운데 아직까  
지 상당수의 백신에서 이를 탐지하지 못하고 있는 것으로 드러났다.

가장 많이 본 기사 [주간]

- 3D 프린팅, 인공지능 등 10대 미래유망기술...
- [SECON 2019] 첨단 보안 솔루션으로 ...
- 2019년 상체인식 대표기업들의 해외 공략 포...
- 구미 송강 데일리로 위장해 사용자 정보 노리는 ...
- SECON & eGISEC 2019에서 배우는...
- 2018년에 발견된 취약점은 전부 몇 개일까?
- '2차 북미정상회담' 이슈 악용한 사이버공격 ...
- 카스퍼스키 전 근무자, 국가 반역죄 최종 선고...
- [주말판] 스마트시티 프로젝트, 다른 곳은 어...
- [스페셜 인터뷰] 민원기 과기정통부 제2차관에...
- 2019년 5대 신산업 '사이버보안' ... G...

## <그림 2> 상당수의 안티바이러스가 문서형 악성코드를 탐지 못하는 사례

(출처= <https://www.boannews.com/media/view.asp?idx=75093>)

전체 악성코드에서 문서형 악성코드의 분포가 많지만, 전문적으로 탐지하는 안티바이러스는 많지 않다. 이는 문서형 악성코드가 쉽게 유포될 수 있어 사용자들의 PC에 감염될 수 있다. 따라서 사회적 문제가 발생할 수 있다.

본 프로젝트에서는 문서형 악성코드를 탐지할 수 있는 엔진을 제작하여 문서형 악성코드의 유포와 그로 인한 피해가 생기는 것을 막고자 한다.

## 1.2 추진 배경 및 필요성

### 1.2.1 추진 배경

최근 해커들의 주요 공격 중 하나는 사회 공학(Social Engineering)적 공격이다. 사회 공학적 공격은 시스템이 아닌 사람의 취약점을 공략하는 공격이다. 대상자의 성향, 동향, 추세 등을 파악하여 정보를 수집하고 그 정보를 바탕으로 정부 기관이나 회사 또는 지인으로 속여 대상자의 흥미를 유발할 수 있는 키워드로 내용을 작성한다. 해커들은 메일, SMS, 웹 게시물 등에 이러한 내용에 악성코드가 삽입된 문서형 악성코드를 첨부하여 대상자 또는 불특정 다수가 의심 없이 첨부파일을 실행하게 유도하는 것이 특징이다.



## 통일부 기자단에 악성코드 메일 배포돼...“北 소행 의심”

입력: 2019-01-07 11:30 | 수정: 2019-01-07 15:12

통일부 “관계기관에 상황 전파...새해 정부 사칭 해킹 많아”



▲ 통일부 [연합뉴스TV 제공] 연합뉴스

통일부 기자단에 북한의 소행으로 의심되는 악성코드가 담긴 메일이 7일 배포돼 정부가 사실관계 확인에 나섰다.

### <그림 3> 이메일을 활용한 문서형 악성코드 유포 사례

(출처 = <http://www.seoul.co.kr/news/newsView.php?id=20190107800022>)

한 가지 사례로 2019년 1월, 통일부를 출입했던 언론사 취재기자들에게 통일부로 사칭하여 일괄적으로 'TF 참고.zip'라는 제목의 메일이 배포됐다. 메일의 내용은 'TF 참고되시길 ~, 언론사별 브랜드 관련해서 관리 잘해주시고~. 비번은 "tf"'라며 첨부된 문서형 악성코드의 실행을 유도하는 문구가 포함되어 있었다. 압축파일 안에는 pdf 파일과 hwp 파일이 있었으며 개인정보를 수집하고 해킹을 시도하려는 문서형 악성코드가 발견됐다.

	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time MAlicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

## 1.2.2 현재 기술 시장 현황

### 1) 안랩(AhnLab) MDS



<그림 4> 안랩 MDS 장비

(출처 = <https://www.ahnlab.com/kr/site/product/productView.do?prodSeq=68>)

안랩(AhnLab)은 한국 정보 보안 업체 중 하나로, 안티바이러스인 V3로 잘 알려져 있다. 안랩은 V3 외에도 다른 소프트웨어 및 하드웨어 보안 솔루션, 모바일 보안, 정보보안 컨설팅, 기타 산업용 제품 보안 등 다양한 분야에서 보안 사업을 하고 있다.

안랩의 보안 솔루션 MDS는 네트워크 샌드박스 및 전용 에이전트를 통해 다양한 경로를 통해 유입되는 위협을 신속하게 수집하며, 시그니처 기반, 평판 기반, 비 시그니처(signature-less) 기반, 동적 행위 분석 등 멀티 엔진을 기반으로 기존 방식의 위협(Known), 알려지지 않은(Unknown) 신·변종 위협까지 정확하고 효율적으로 탐지 및 대응한다.



	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time Malicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

## 2) 시만텍(Symantec)

시만텍은 미국의 보안 소프트웨어 회사이다. 시만텍이 개발하고 배포하는 제품인 노턴 안티바이러스는 악성 코드 방지 및 제거 기능을 제공한다. 또한 스팸 메일 필터링과 피싱 보호 기능이 있으며 2007 년에 바이러스 검사 소프트웨어 시장에서 61%를 차지했다.

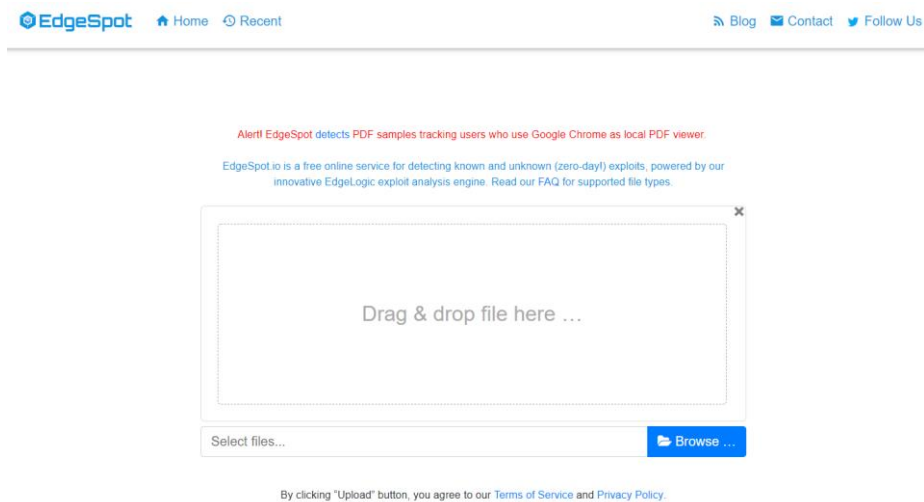
또한 시만텍은 시그니처의 한계를 제시하며 Symantec Endpoint Protection 14 를 개발하고 있다. Symantec Endpoint Protection 14 는 시그니처에 의존하지 않고 기계 학습(Machine Learning) 및 행위 분석을 통해 보안 효과를 극대화하고 오탐을 최소화한다.



<그림 5> 시만텍 로고

(출처 = <https://www.symantec.com/>)

## 3) EdgeSpot



<그림 6> EdgeSpot 화면

(출처 = <https://edgespot.io/>)

Edge Spot은 알려지거나 알려지지 않은 (Zero Day) 공격에 대해 탐지 기능을 제공하는 무료 온라인 웹 서비스이다. PDF, Microsoft Office 파일 등 문서형 파일을 업로드하면 탐지 결과를 총 4가지(Malicious, Suspicious, Information, No threat found)로 분류하여 사용자에게 보여준다. Edge Spot은 정적 분석 및 동적 분석, 기계 학습과 같은 기술을 사용하여 분석한다.

	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time mAlicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

#### 4) 지란지교시큐리티 SaniTOX



**<그림 7> 지란지교시큐리티 SaniTOX**

(출처 = <https://www.jiransecurity.com/products/sanitox>)

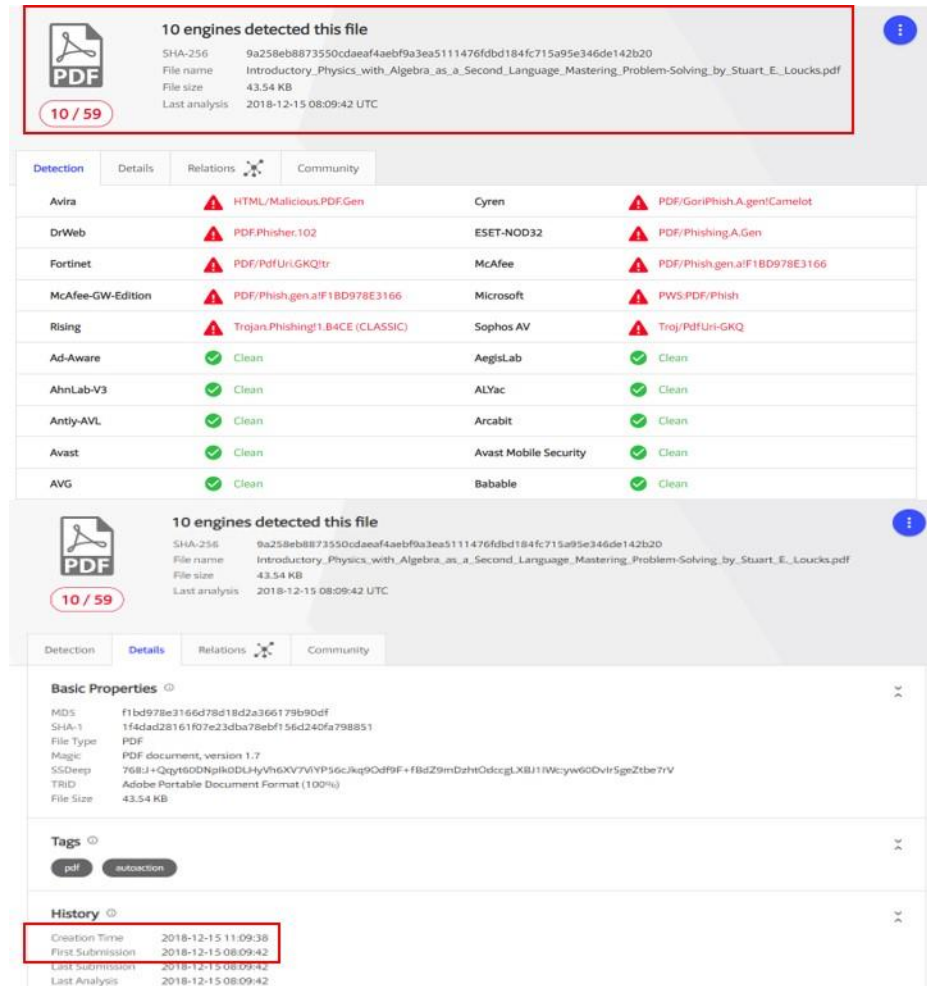
지란지교시큐리티는 문서보안, 암호화폐 보안과 모바일보안 등 소프트웨어 보안 전문 기업이다. 최근에 지란지교시큐리티가 자체 개발한 CDR (Content Disarm & Reconstruction) 기술을 이용하여 콘텐츠 악성코드를 무해화하는 새니톡스(SaniTox) 솔루션을 공개하였다. CDR은 파일 내 잠재적 보안 위협 요소를 탐지하여 제거한 뒤에 안전한 파일로 재조합하여 악성코드 감염 위험을 사전에 방지할 수 있는 기술이다.

새니톡스 솔루션은 2 가지의 형태로 제공이 된다. 첫 번째, 새니톡스 어플라이언스는 별도의 소프트웨어 설치나 설정이 필요 없는 일체형 장비로 쉽게 도입이 가능하다. 그리고 Contnet Prevention Engine(Anti-Virus + CDR) 기반의 알려진 위협에 대한 1 차 필터링과 문서 기반의 표적형 악성코드에 대한 2 차 예방적 보안을 통해 전방위 위협에 대응한다. 두 번째, 새니톡스 SDK는 소프트웨어 개발사 및 서비스 제공 기업이 새니톡스 CDR 엔진을 자체 소프트웨어, 하드웨어 혹은 서비스에 통합할 수 있도록 API 를 제공한다. 다양한 콘텐츠 유입 채널이 있는 제품을 통해 콘텐츠 악성코드 무해화 기능을 제공 할 수 있다.

	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time Malicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

## 1.2.3 현재 기술 시장의 문제점 및 개선 방향

### 1.2.3.1. 기술 시장 문제 1



**10 engines detected this file**

SHA-256: 9a258eb8873550cdaef4aebf9a3ea5111476fbd184fc715a95e346de142b20  
File name: Introductory\_Physics\_with\_Algebra\_as\_a\_Second\_Language\_Mastering\_Problem-Solving\_by\_Stuart\_E\_Loucks.pdf  
File size: 43.54 KB  
Last analysis: 2018-12-15 08:09:42 UTC

Detection	Details	Relations	Community
Avira	HTML/Malicious.PDF.Gen	Cyren	PDF/GonPhish.Agent/Camelot
DrWeb	PDF.Phisher.102	ESET-NOD32	PDF/Phishing.A.Gen
Fortinet	PDF/PdfUri.GKQ/tr	McAfee	PDF/Phish.gen.a/F1BD978E3166
McAfee-GW-Edition	PDF/Phish.gen.a/F1BD978E3166	Microsoft	PWS.PDF/Phish
Rising	Trojan.Phishing!1.B4CE (CLASSIC)	Sophos AV	Troj/PdfUri-GKQ
Ad-Aware	Clean	AegisLab	Clean
AhnLab-V3	Clean	ALYac	Clean
Antiy-AVL	Clean	Arcabit	Clean
Avast	Clean	Avast Mobile Security	Clean
AVG	Clean	Babable	Clean

**Basic Properties**

MD5: f1bd978e3166d78d18d2a366179b90df  
SHA-1: 1f4dad2816107e23dba78ebf156d240fa798851  
File Type: PDF  
Magic: PDF document, version 1.7  
SSDeep: 768J+Qqy60DNpik0DLHyVh6XV7VVP56cJkq9OdR9F+fBdZ9mDzhtOdcg1XBJ1fWcyw60DvirSge2tbe7V  
TRID: Adobe Portable Document Format (100%)  
File Size: 43.54 KB

**Tags**

pdf, automation

**History**

Creation Time	2018-12-15 11:09:38
First Submission	2018-12-15 08:09:42
Last Submission	2018-12-15 08:09:42
Last Analysis	2018-12-15 08:09:42

<그림 9> 바이러스토탈의 문서형 악성코드 결과 예시

<그림 9>는 2018년 12월에 등장했던 문서형 악성코드를 바이러스토탈에 업로드한 화면이다. 바이러스토탈 결과 59곳의 안티바이러스 중 오직 10곳의 안티바이러스가 악성이라고 탐지했다. 최신 문서형 악성코드를 탐지하는 안티바이러스가 적다는 것을 확인 할 수 있다.

	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time Malicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

### 1.2.3.2. 기술 시장 문제 2


A 사의 솔루션은 문서 파일 내 실행 가능한 액티브 콘텐츠(매크로, 자바스크립트 등)을 원천 제거하여 문서 파일이 어떠한 동적 행위를 할 수 없는 파일로 만들기 때문에 악성 행위를 일절 차단할 수 있는 장점이 있다. 하지만 문서가 정상 파일 일지라도 문서 내에 존재하는 액티브 콘텐츠를 일절 제거하기 때문에 사용자들은 정상적으로 문서를 사용할 수 없게 된다.

본 프로젝트에서는 파일의 구조를 확인 후 정적 분석 과정으로 특징을 추출 후 기계 학습 기법으로 학습한 모델로 악성 코드를 탐지하기 때문에 위 솔루션에서의 정상 파일까지 변환되는 문제가 해결된다.

### 1.2.3.3. 기술 시장 문제 3

B사는 파일 탐지 주요 기술로 동적 행위 분석을 사용하는데 이때 많은 시간과 비용이 발생한다. 하루에 분석 할 수 있는 데이터의 양의 한계가 있으며 대용량의 데이터를 처리하는데 어려움이 있다. 또한 많은 시간과 비용이 발생하기 때문에 일반 사용자나 가정에서는 사용하기 어려운 단점이 있다. 본 프로젝트는 정적 분석 기반의 기계 학습 기법을 사용하여 데이터를 탐지하기 때문에 많은 양의 데이터를 대응 하는데 유리하다.

따라서 본 프로젝트는 PDF나 MS Office 등 의 문서형 파일이 악성인지 아닌지 탐지하여 문서형 악성코드가 유포되는 것을 방지하고자 한다. 그리고 프로젝트를 오픈소스화하여 사용자들이 필요로 하는 방향으로 코드를 자유롭게 수정 할 수 있도록 한다.

	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time Malicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

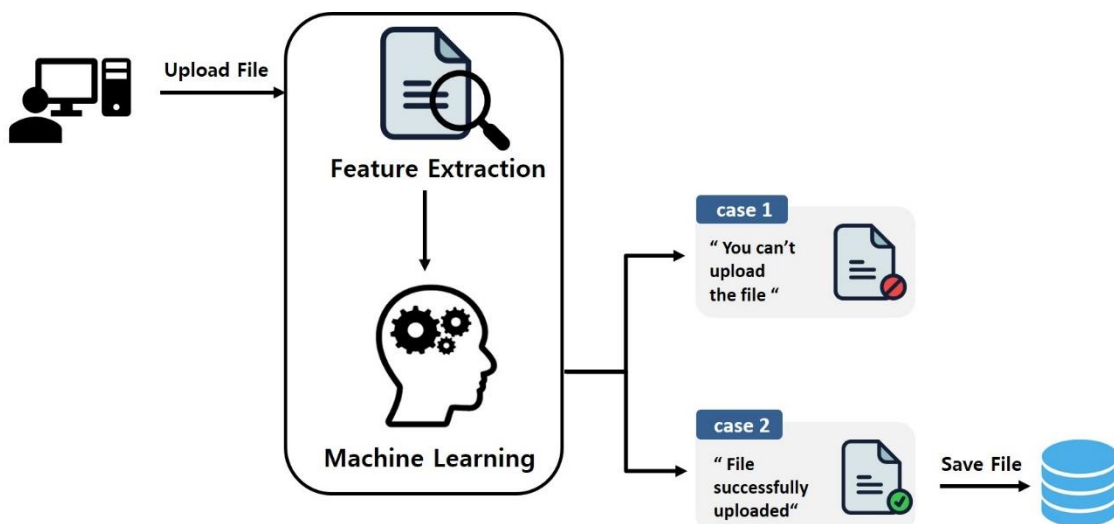
## 2 개발 목표 및 내용

### 2.1 목표

본 프로젝트는 문서형 악성코드를 탐지하는 엔진을 개발하여 문서형 악성코드가 유포되는 것을 방지하고 확장성이 좋은 구조로 개발하여 웹 서비스 또는 메일 서비스 등 여러 시스템에 쉽게 확장하는 것을 목표로 한다. 또한 오픈소스 소프트웨어로 개발하여 사용자가 수정 및 보완할 수 있도록 하는 것을 목표로 한다.

### 2.2 연구/개발 내용

#### 2.2.1 시연 시나리오



11

<그림 10> 프로젝트 예상 시연 시나리오

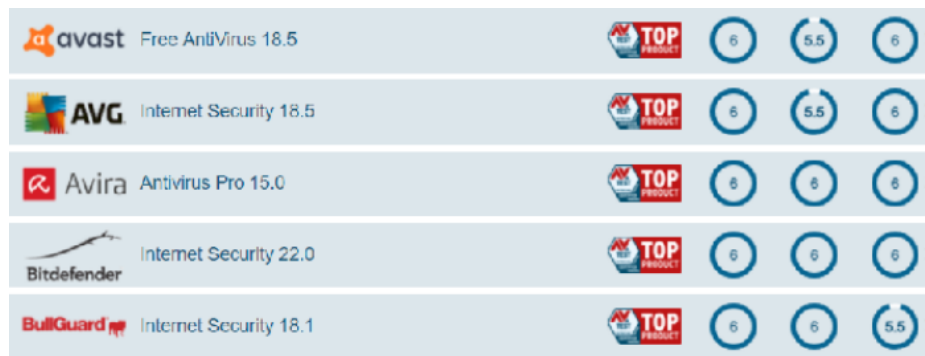
사용자가 문서 파일을 업로드하면 파일의 유형(pdf, docx 등)에 따라 피처를 추출한다. 업로드한 파일에서 추출한 특징은 학습된 모델의 입력으로 하여 그 결과에 따라 사용자에게 메시지를 띄워준다. 만약 업로드한 문서 파일이 악성으로 판단되면 “바이러스에 감염된 파일로 의심되어 업로드 할 수 없습니다.”라는 메시지를 띄워준다. 반대로 정상 파일로 판단되면 “정상적으로 업로드 되었습니다.”라는 메시지를 띄운다.

	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time mAlicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

## 2.2.2 연구/개발 방법

### 2.2.2.1. 데이터 라벨링 (Data Labeling)

기계 학습에 있어서 중요한 것 중 하나는 데이터 라벨링이다. 본 프로젝트에서는 지도 학습을 사용하는데, 지도학습에서 학습 데이터에 대한 라벨링을 올바르게 않는다면 그에 대한 결과는 전혀 다른 결과를 초래한다. 그러므로 올바른 라벨링을 하는 것이 중요하다.



<그림 11> AV-TEST에서 각각의 안티바이러스에 대한 평가 지표 점수

(출처 = <https://www.av-test.org/en/antivirus/home-windows/>)

본 프로젝트는 파일에 대한 바이러스토탈 결과에서, 글로벌 안티바이러스 테스트 기관인 'AV-TEST' 와 'Virus Bulletin'의 성능 지표를 종합하여, 상위 5 개의 안티바이러스가 악성이라고 판단하면 악성으로, 바이러스토탈에 등록된 안티바이러스 모두가 정상이라고 판단하면 정상으로 라벨링 하고자 한다.

	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time mAlicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

## 2.2.2.2. 특징 추출

문서의 유형별로 악성과 정상을 구분할 수 있는 특징을 추출한다. PDF나 MS Office 문서 등 문서의 유형별로 구조가 다르기 때문에 유형에 따라 특징을 추출하여 분석한다. 예를 들어 PDF는 문서의 구조적 특징을 분석할 수 있는 PDF파서(Parser)를 사용한다.


```

PDFiD 0.2.5 test.pdf
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

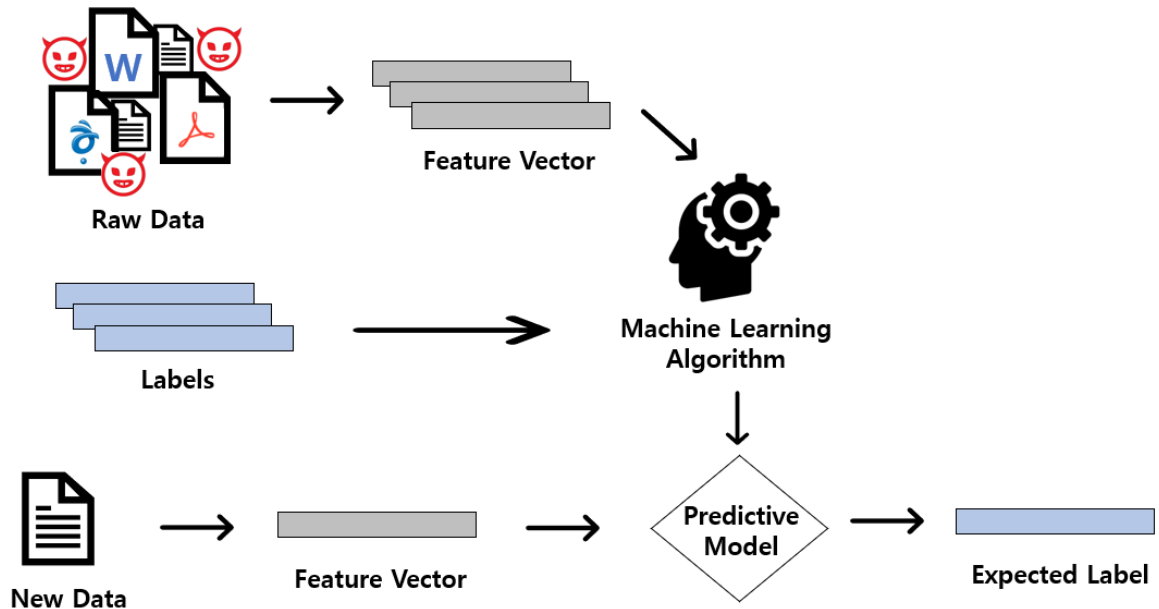
```

<그림 12> PDF파서 실행 화면 예시

<그림 12>는 PDF 파서(Parser)를 사용하여 “test.pdf” 파일을 분석한 결과이다. 왼쪽 열은 문서의 구성요소를 나타내며 오른쪽 열은 구성요소의 개수를 의미한다. 파서 코드를 수정하거나 추가하여 문서에서 추출되는 특징을 추가할 것이다. 구한 특징은 벡터화하여 기계 학습에 사용할 예정이다.

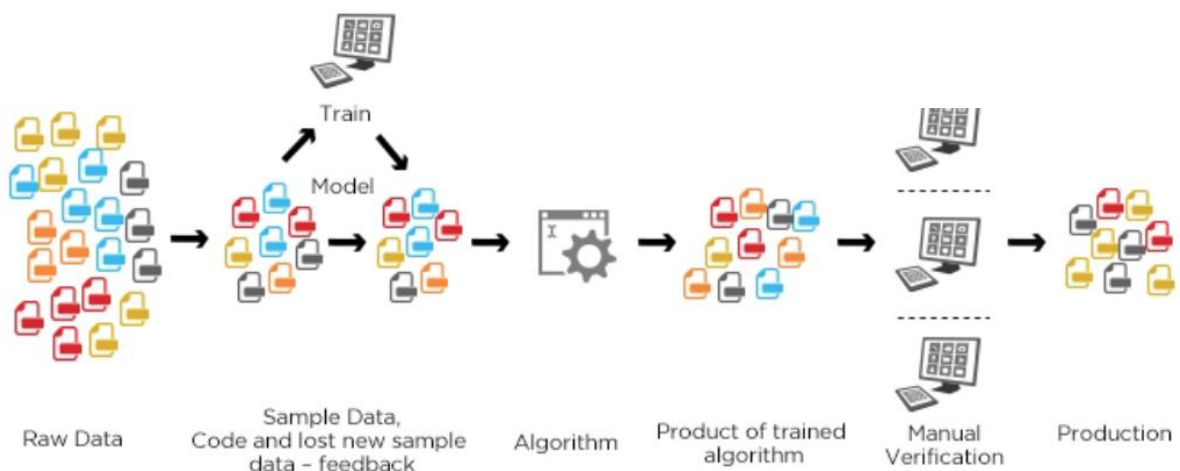
	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time MAlicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

### 2.2.2.3. 기계 학습



<그림 13> 학습 및 테스트 과정

기계 학습을 사용하여 문서 파일이 악성인지 정상인지 판별하고자 한다. 기계 학습의 학습 방법은 지도 학습, 비지도 학습, 반지도 학습으로 나뉘는데, 본 프로젝트에서는 지도 학습을 사용한다. 탐지 모델의 학습 및 테스트 과정은 <그림 13>의 과정으로 이루어진다.



<그림 14> 지도 학습의 개요



	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time MAlicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

(출처 = <http://blog.naver.com/PostView.nhn?blogId=jn-solution&logNo=221278915519&parentCategoryNo=&categoryNo=7&viewDate=&isShowPopularPosts=true&from=search>)

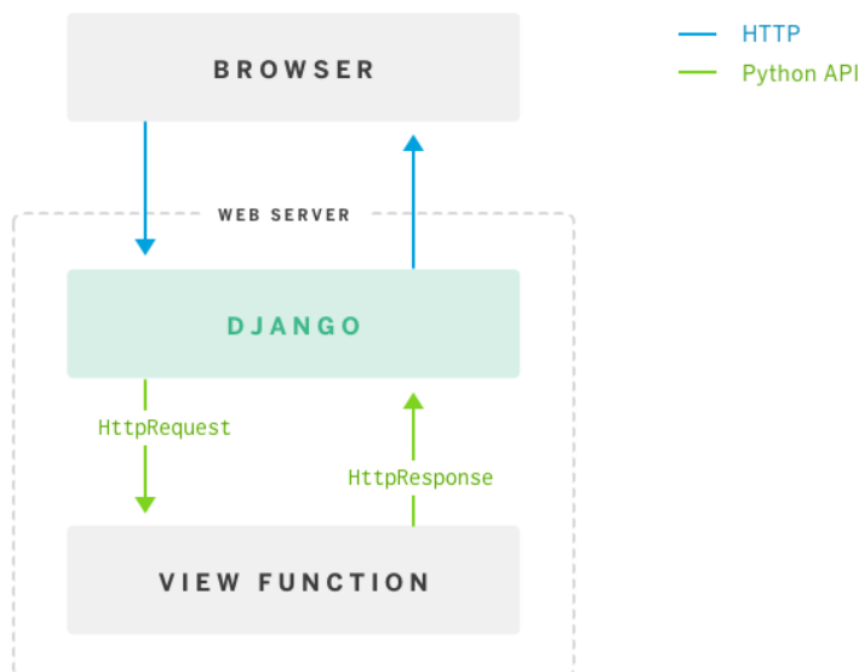
라벨링 된 정상 파일과 악성 파일의 특징을 추출하여 벡터화를 한 뒤, 기계 학습 모델의 입력 값으로 활용한다. 기계 학습 모델의 종류는 SVM, Random Forest, GBDT, DNN 등 다양한 모델을 사용하여 실험을 진행할 예정이다.

본 프로젝트는 다양한 모델로 실험을 진행해서 가장 좋은 성능을 보이는 모델을 사용하거나 여러 모델을 앙상블 한 모델을 사용할 예정이다.

#### 2.2.2.4. 웹

시나리오를 위한 기본적인 웹 서비스를 구현한다.


본 프로젝트에서 개발한 엔진을 웹 서비스에 확장한 상황을 가정한다. 웹 서비스는 파일 업로드가 가능한 게시판 형태로 구현한다. 파일이 업로드 되었을 때 정상 파일이라면 정상적으로 업로드 되었음을 클라이언트에 전달하여야 한다. 또한 게시판에 업로드 된 파일을 저장하기 위해 DB 연동이 되어야 하며 Django를 사용하여 웹 서버와 DB를 연동할 계획이다.



<그림 15> 웹 서비스 구조

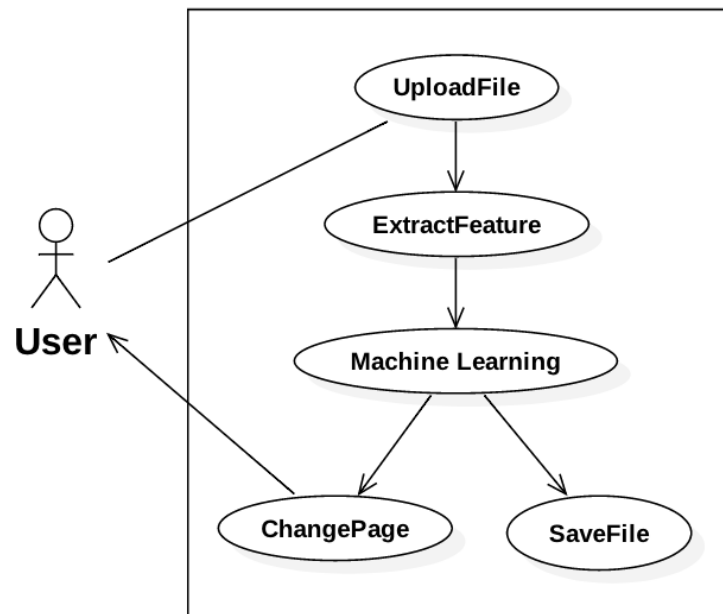
(출처 =

[https://blog.heroku.com/in\\_deep\\_with\\_django\\_channels\\_the\\_future\\_of\\_real\\_time\\_apps\\_in\\_django](https://blog.heroku.com/in_deep_with_django_channels_the_future_of_real_time_apps_in_django))

	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time mAlicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

## 2.3 개발 결과

### 2.3.1 시스템 기능 요구사항



<그림 16> 유즈 케이스(Use Case)

### 2.3.2 시스템 비기능(품질) 요구사항

#### 1) 성능

사용자가 파일을 업로드 했을 때, 사용자에게 최대한 빠르게 업로드 결과를 보여줘야 한다. 따라서 업로드 된 파일의 특징을 추출해서 악성인지 판단하는데 걸리는 시간을 최대 2초로 제한한다. 그리고 사용자가 믿을 수 있는 결과를 도출해야 한다. 따라서 본 프로젝트의 엔진은 98% 이상의 정확도를 보장해야 한다.

#### 2) 보안성

악의적으로 대용량의 파일을 업로드하여 서버에 부하를 가하는 경우를 대비하여 업로드 가능한 파일의 개수를 5개, 파일의 크기를 개당 5MB로 제한한다.

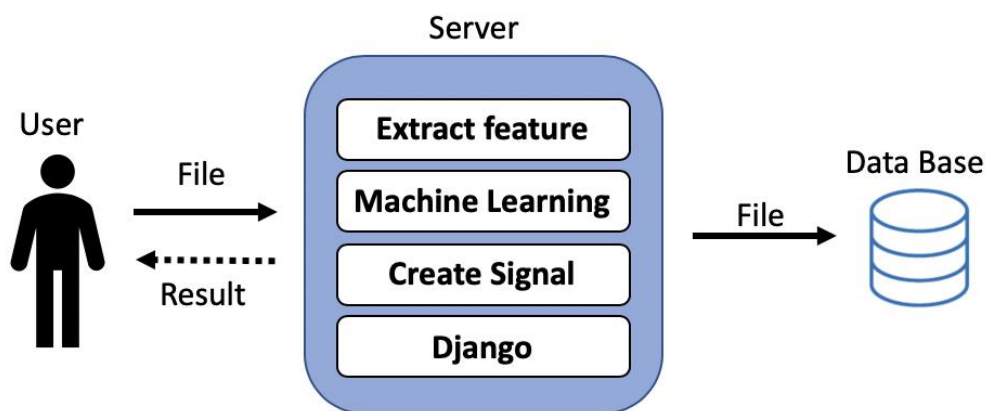
	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time mAlicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

### 3) 유지보수성

사용자들은 보유한 데이터가 없어 모델을 학습하는데 어려울 수 있다. 따라서 주기적으로 모델을 업데이트하여 사용자들에게 배포한다.


사용자 오픈소스 커뮤니티를 운영하여 사용자와 데이터를 공유한다. 만약 오탐 발생 시 사용자 커뮤니티에서 데이터 정보를 즉시 수집 받고 모델을 재 학습 후 재배포한다.

### 2.3.3 시스템 구조



<그림 17> 시스템 구조도

사용자가 파일을 업로드한다. 업로드 된 파일은 서버로 전송된다. 서버에서는 파일의 특징을 추출하여 특징 벡터를 생성한다. 특징 벡터로 학습된 모델 기반으로 악성 여부를 검사하고 검사 결과가 정상이라면 업로드 된 파일을 DB 에 저장한다.

	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time Malicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

## 2.3.4 결과물 목록 및 상세 사양

대분류	소분류	기능	형식	비고
탐지 엔진	특징 추출기	문서형 파일을 정적 분석하여 특징을 추출한다.	모듈	
	탐지기	문서형 파일이 악성인지 정상인지 검사한다.	모듈	

## 2.4 기대효과 및 활용방안

### 2.4.1 기대효과

본 프로젝트는 문서형 파일을 업로드 하기 전 파일이 악성인지 정상인지 판별하여 유포 및 감염을 막는다. 현재 문서형 악성코드는 사람들의 관심을 끌 만한 제목으로 유포되고 있으며 사용자들은 해당 파일들에 경계심을 갖지 않고 다운로드 하여 실행할 수 있다. 이로 인한 피해로 발생 할 수 있는 사회적 문제를 예방할 수 있을 것이라 기대한다.

### 2.4.2 활용방안

#### ■ 웹 서비스

웹 서비스 운영자는 웹 서버에 본 프로젝트의 엔진을 적용함으로써 문서형 악성코드가 무단으로 업로드 되는 것을 방지할 수 있다.

#### ■ PC

본 프로젝트의 엔진을 PC에서 실행하여 사용자의 컴퓨터에 문서형 악성코드가 침입하는 것을 방지 할 수 있다.

	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time malicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

## 3 배경 기술

### 3.1 기술적 요구사항

#### ■ 개발 언어

- Python 3.6.5

#### ■ 개발 환경

- CPU : Intel Core i7-5500U @ 2.40GHz 2.39 GHz
- Storage : 512GB HDD
- RAM : Samsung DDR3 4GB

#### ■ 라이브러리

- Django 2.1
- numpy 1.14.5
- Pandas 0.23.4
- Scikit-Learn 0.20.0
- LightGBM 2.2.4
- Pytorch 1.0.1
- Tensorflow 1.8.1

#### ■ 데이터

- Virussign
- VirusShare
- Contagio Blog

	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time malicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

## 3.2 현실적 제한 요소 및 그 해결 방안

### 3.2.1 하드웨어

웹 서버에 엔진이 정상적으로 확장될 수 있는 서버 사양이 되어야 한다.

기계 학습 모델을 학습할 때 많은 양의 데이터를 학습할 경우 오랜 시간이 소요된다. 이는 GPU를 사용함으로써 모델이 학습되는데 걸리는 시간을 단축할 수 있다.

### 3.2.2 소프트웨어

웹 서버와 엔진 간의 통신이 원활하게 되도록 해야 한다.

시간이 지남에 따라 악성 데이터의 특징이 변화하기 때문에 주기적으로 학습 모델을 재학습해야 한다. 모델을 학습할 때, 한 도메인의 데이터만 학습한다면 과적합(Overfitting)이 일어날 수 있다. 따라서 여러 도메인의 자료를 수집해서 학습하고, 데이터의 일반적인 특징을 추출해서 학습하도록 한다.

### 3.2.3 기타

파일을 ZIP, 7Z 등으로 압축하여 업로드 할 수 있다. 따라서 업로드 한 파일이 압축 파일 일 때 압축 파일 내부의 파일들을 모두 검사하도록 한다.

	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time MAlicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

## 4 프로젝트 팀 구성 및 역할 분담

이름	역할
김기환	- 데이터 전처리 - 데이터 수집
김현석	- 데이터 라벨링 - 웹 서버 구축
문다민	- 기계 학습 모델 설계 및 구축 - 엔진 설계 및 개발
방유한	- 유저 인터페이스 구현 - 로고 디자인
정혜리	- 데이터 전처리 - 문서 작업

## 5 프로젝트 비용

항목	예상치 (MD)
아이디어 구상	10
AWS 서버	10
웹 디자인	20
엔진 설계	35
엔진과 웹 서비스 연동	20
탐지 모델 제작	20
프로젝트 테스트 및 유지보수	25
프로젝트 평가 및 보고서 작성	30
합	170

	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time Malicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

## 6 개발 일정 및 자원 관리

### 6.1 개발 일정

항목	세부내용	1 월	2 월	3 월	4 월	5 월	6 월	비고
요구사항분석	요구 분석							
	SRS 작성							
관련분야연구	주요 기술 연구							
	관련 시스템 분석							
설계	시스템 설계							
구현	코딩 및 엔진 테스트							
테스트	시스템 테스트							



	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time mAlicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

## 6.2 일정별 주요 산출물

마일스톤	개요	시작일	종료일
계획서 발표	개발 환경 구축 <b>산출물 :</b> 1. 프로젝트 수행 계획서 2. 계획서 발표 자료	2019-03-03	2019-03-15
설계 완료	시스템 설계 완료(Django 와 엔진 연동) <b>산출물 :</b> 1. 시스템 설계 사양서	2019-03-15	2019-03-23
중간 보고	프로그램 기본 기능 구현 완료 <b>산출물 :</b> 1. 프로젝트 1 차 중간 보고서 2. 프로젝트 진도 점검표 3. 1 차 구현 소스 코드	2019-03-23	2019-04-19
구현 완료	서버, 엔진 구현 완료 <b>산출물:</b> 1. 문서형 악성코드 탐지 엔진	2019-04-01	2019-04-19
테스트	시스템 테스트 <b>산출물 :</b> 시연용 문서형 악성코드 탐지 시스템	2019-04-13	2019-05-31
최종 보고서	<b>산출물</b> 최종 보고서	2019-05-31	2019-06-07

	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time mAlicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

### 6.3 인력자원 투입 계획

이름	개발항목	시작일	종료일	총개발일(MD)
전원	아이디어 구상	2019-03-03	2019-03-08	10
문다민	머신 러닝 기반 학습 모델 개발 프로젝트 유닛테스트 및 유지보수	2019-04-04	2019-05-26	25
김기환	최신 기술 현황 분석 바이러스토탈 레포트 수집	2019-03-09	2019-05-26	15
김현석	데이터 라벨링 웹 서버 구축	2019-03-09	2019-05-26	20
정혜리	데이터 수집 데이터 전처리	2019-03-09	2019-05-26	25
방유한	유저 인터페이스 구현	2019-04-01	2019-05-26	20


### 6.4 비 인력자원 투입 계획

항목	Provider	시작일	종료일	Required Option
개발용 PC 4 대	ThinkPad	2019-03-03	2019-05-31	
AWS EC2	AWS	2019-03-03	2019-05-31	p2.xlarge

	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time Malicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

## 7 참고 문헌

번호	종류	제목	출처	발행 년도	저자	기타
1	웹 페이지	VirusTotal	<a href="https://www.virustotal.com/en/statistics/">https://www.virustotal.com/en/statistics/</a>			
2	기사	이력서 위장 '시그마' 랜섬웨어... 전 세계 유포	<a href="https://www.sedaily.com/NewsView/1RX4PK7LDU">https://www.sedaily.com/NewsView/1RX4PK7LDU</a>			
3	기사	MS 워드 매크로 악용 '갠드크랩' 랜섬웨어 기승	<a href="http://it.chosun.com/site/data/html_dir/2018/11/16/2018111601952.html">http://it.chosun.com/site/data/html_dir/2018/11/16/2018111601952.html</a>			
4	기사	정부 사칭 이메일 공격 계속 발견	<a href="http://www.boan.com/news/article.html?id=20181130150004">http://www.boan.com/news/article.html?id=20181130150004</a>			
5	논문	문서 구조 및 스트림 오브젝트 분석을 통한 문서형 악성코드 탐지	<a href="http://www.dbpia.co.kr/Journal/ArticleDetail/NODE07565787">http://www.dbpia.co.kr/Journal/ArticleDetail/NODE07565787</a>			
6	논문	Malicious PDF Detection using Metadata and Structural Features	<a href="http://delivery.acm.org/10.1145/2430000/2420987/p239-smutz.pdf?ip=203.246.112.134&amp;id=2420987&amp;acc=ACTIVE%20SERVICE&amp;key=0EC22F8658578FE1%2EB574870CA11B57BF%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&amp;__acm__=1552357255_267897824fceb113cb198d370cb5329">http://delivery.acm.org/10.1145/2430000/2420987/p239-smutz.pdf?ip=203.246.112.134&amp;id=2420987&amp;acc=ACTIVE%20SERVICE&amp;key=0EC22F8658578FE1%2EB574870CA11B57BF%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&amp;__acm__=1552357255_267897824fceb113cb198d370cb5329</a>			
7	기사	국내 유명 변호사 사칭한 악성코드, 상당수 백신 탐지 못해	<a href="https://www.boannews.com/media/view.asp?idx=74302">https://www.boannews.com/media/view.asp?idx=74302</a>			
8	기사	요즘 해커들 사이에서 가장 인기 높은 건, MS 오피스	<a href="https://www.boannews.com/media/view.asp?idx=76967&amp;page=1&amp;mkind=1&amp;kind=1">https://www.boannews.com/media/view.asp?idx=76967&amp;page=1&amp;mkind=1&amp;kind=1</a>			
9	논문	ALDOX: Detection of Unknown Malicious Microsoft Office Documents Using Designated Active Learning Methods Based on New	<a href="https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&amp;arnumber=7762928">https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&amp;arnumber=7762928</a>			

	<b>국민대학교</b> <b>소프트웨어학부</b> <b>캡스톤 디자인 I</b>	<b>계획서</b>		
		<b>프로젝트 명</b>	DREAM(Detecting in Real-time mAlicious document using Machine Learning)	
		<b>팀 명</b>	Do it!	
		Confidential Restricted	Version 1.5	2019-MAR-13

		Structural Feature Extraction Methodology				
10	논문	A Research of Anomaly Detection Method in MS Office Document	<a href="http://kiss.kstudy.com/thesis/thesis-view.asp?key=3498648">http://kiss.kstudy.com/thesis/thesis-view.asp?key=3498648</a>			