

INFOTECH

AI기반 DGA 멀웨어 C&C서버 접속 탐지 시스템

2020 케이실드주니어 보안사고 분석대응 4기 정보보안 프로젝트

**AI 모델 기반
DGA 멀웨어 C&C서버
접속 탐지 시스템
개발 프로젝트**



목차

05	1. 개요
06	1.1 프로젝트 명 및 기간
06	1.2 프로젝트 배경
08	1.3 프로젝트 목적
08	1.4 기대효과
09	2. 프로젝트 조직
10	2.1 프로젝트 구성원
10	2.2 책임 및 역할
11	3. 프로젝트 수행 일정
12	3.1 프로젝트 추진 일정
12	3.2 단계별 세부 일정
13	4. 프로젝트 설계
14	4.1 프로세스 흐름도
14	4.2 프로세스 분석
14	4.2.1 자료 흐름도 (DFD)
15	4.2.2 기능 명세서
16	4.3 사용자 인터페이스 분석
16	4.3.1 입·출력 인터페이스 종류와 기능
16	4.4 시스템 구성도
17	4.5 자료 설계
17	4.5.1 자료 형태 상세 구분
18	4.6 시스템 설계
18	4.6.1 구조 차트
19	4.6.2 모듈 설계
19	4.6.2.1 DNS 패킷 캡처링 모듈
20	4.6.2.2 DGA 도메인 판별 모듈(AI)
23	4.6.2.3 판별 결과 전달 모듈
23	4.6.2.4 관리자 대시보드
25	4.6.2.5 데이터베이스

25	4.7 핵심 알고리즘 (AI 피쳐)
25	4.7.1 TLD_Index
27	4.7.2 N-gram Score
29	4.7.3 Length
30	4.7.4 Numeric_ratio
30	4.7.5 Vowel_ratio
31	4.7.6 Consonant_ratio
31	4.7.7 Consecutive_consonant
32	4.7.8 Consecutive_Vowel
32	4.7.9 period
33	4.7.10 Entropy
34	4.7.11 Max_Consecutive_Consonant
34	4.7.12 Max_Vowel_Consonant
35	4.7.13 Meaning_count
36	5. 시스템 구현
37	5.1 개발 언어 및 라이브러리
38	5.2 구현 결과
38	5.2.1 DNS 패킷 캡처링 모듈
39	5.2.2 데이터베이스
40	5.2.3 판별 결과 전달 모듈
41	5.2.4 관리자 대시보드
47	5.2.5 DGA 도메인 판별 모듈(AI)
50	5.3 시스템 시험
50	5.3.1 시험 환경
50	5.3.1.1 시험 환경 사양
51	5.3.2 가상 네트워크 구성
53	5.3.3 기능 시험
53	5.3.3.1 시험 요구사항 정의
54	5.3.3.2 시험 방법
55	5.3.3.3 시험 결과
56	6. 결론
57	6.1 향후계획

01

개요

1.1 프로젝트 명 및 기간

1.2 프로젝트 배경

1.3 프로젝트 목적

1.4 기대효과

01 개요

| 1.1 프로젝트 명 및 기간

프로젝트 명: AI기반 DGA 멀웨어 C&C서버 접속 탐지 시스템

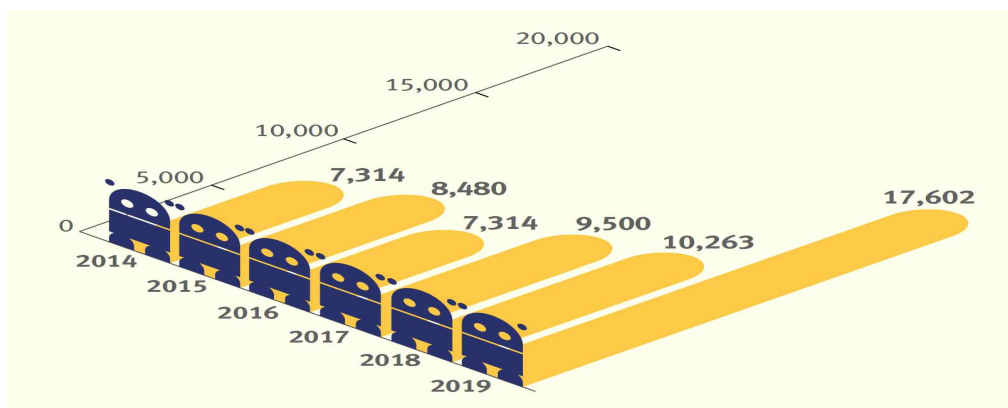
프로젝트 기간: 2020.05.01 ~ 2020.07.04 (약 2개월)

프로젝트 수행 팀: 인포텍트(INFOTECT)

프로젝트 대상: 기업이나 기관 네트워크

| 1.2 프로젝트 배경

Bot은 취약점을 점검하거나 합법적인 정보 수집 혹은 단순하고 반복적인 업무를 수행하기 위해 개발되고, 활용되고 있지만, 공격에도 사용되고 있다. BotNet은 주로 초보적인 공격 수단으로 감염되기 때문에, 대규모 감염을 일으키는 사이버 공격 수단으로 빠르게 확산되어 공격의 빈도와 규모가 더욱 커지고 있다. [그림1-1]에 따르면 18년에 탐지한 BotNet C&C서버 10,263개보다 19년에 71.5% 증가한 17,602개를 탐지했으며, 16년 이후부터 BotNet C&C서버 탐지 개수가 꾸준히 증가하고 있는 것을 볼 수 있다.



[그림1-1] Spamhaus 2019년 BotNet 보고서

BotNet에 의해 악성 공격들이 증가하고, 다양해지면서 봇의 감염코드를 이용한 시그니처 기법과 이상 행위 등 탐지 방법들을 통해 차단하려는 시도들이 있다. 하지만 최근 BotNet들은 이를 우회하는 기능들을 탑재하고 있으며, C&C서버의 IP가 발각될 위험이 있기 때문에 DNS를 이용해 자동으로 명령/제어 서버에 접속하여 공격을 시도한다. PaloAlto의 위협 조사팀 Unit42가 20년 초에 발표한 "Stop Attackers from using DNS against you" 보고서에 따르면 악성코드의 80%는 고정 IP 대신 DNS를 사용한 C&C서버와 통신하며, 그 중 18%는 무작위의 도메인을 생성하는 DGA를 사용하여 C&C서버로의 접속이 차단되는 것을 우회한다고 밝혔다.

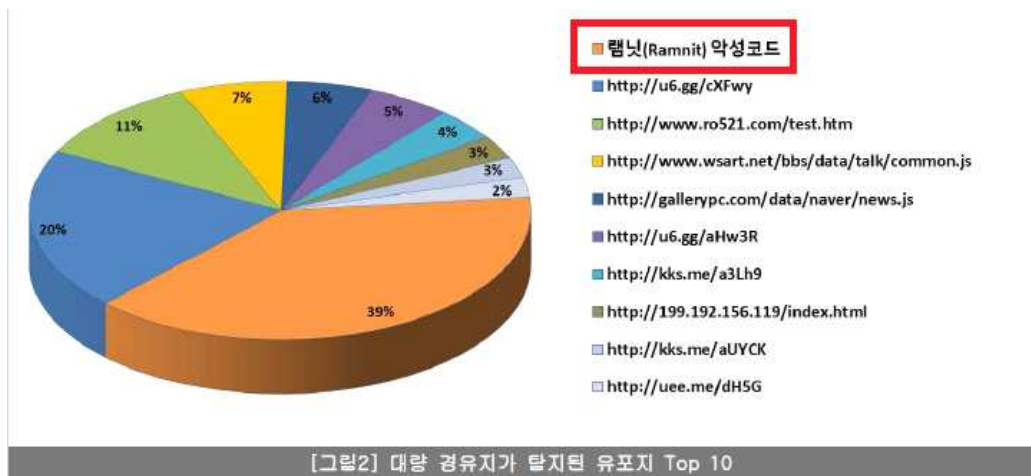
VII. 악성 행위 탐지를 우회하는 공격 기법의 진화

2 '19년 보안위협 전망

DGA(Domain Generation Algorithm)를 이용하여 C&C 차단을 회피하는 악성코드 증가

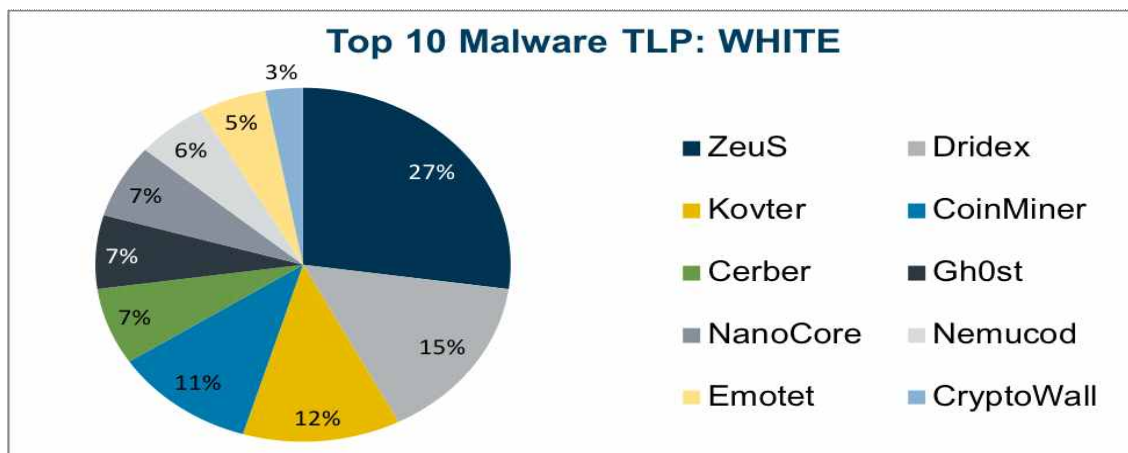
[그림1-2] KISA 2019년 7대 사이버 공격 전망

또한 KISA에서 발표한 2019년 7대 사이버 공격 전망 보고서를 통해 DGA를 사용하여 C&C 서버 차단을 회피하는 악성코드들이 증가될 것이라고 전망하고 있다.



[그림1-3] KISA 2019년 하반기 악성코드 은닉사이트 탐지 동향 보고서

실제로 KISA의 2019년 하반기 악성코드 은닉사이트 탐지 동향 보고서를 보면 DGA를 사용하는 램닛(Ramnit) 악성코드가 39%로 가장 많은 비율을 차지하고 있는 것을 알 수 있다.



[그림1-4] CIS Top 10 Malware January 2020

미국의 비영리조직 CIS(Center for Internet Security)에서 20년 1월에 발표한 Top 10 악성코드 보고서에서도 DGA를 사용하는 Zeus와 Emotet 악성코드가 39%를 차지하고 있었고, 이를 통해 DGA를 사용하는 악성코드가 증가뿐만 아니라 종류도 증가하고 있음을 확인할 수 있다.

일부 관제업체에서는 DGA와 같은 공격기법은 백신 또는 네트워크 보안장비로도 탐지가 어렵다고 한다. 때문에 최근에는 기존의 탐지 방식보다 AI를 이용하여 탐지하는 연구들이 이루어지고 있다.

본 프로젝트에서는 머신러닝 기술을 도입하여 DGA 도메인과 정상도메인을 판별하는 AI모델을 개발하고 이를 활용하여 'DGA 멀웨어 접속 탐지 시스템'을 개발하고자 한다.

| 1.3 프로젝트 목적

DGA 도메인과 정상 도메인을 판별하는 AI 모델이 탑재된 네트워크 보안 시스템을 개발하여 네트워크 내에서 발생하는 DNS 트래픽을 감시하고 C&C서버로의 접속 시도 행위를 탐지한다.

| 1.4 기대효과

1. 메일을 통한 탐지사실 알림기능과 대시보드 페이지를 통한 관리자의 업무효율성 증진
2. DGA 멀웨어에 감염된 HOST 색출이 가능하므로 빠른 초동조치가 가능
3. 수집된 정보를 중앙서버로 수집하여 사이버 위협 정보, 악성코드 특징분석 등에 활용 가능

02

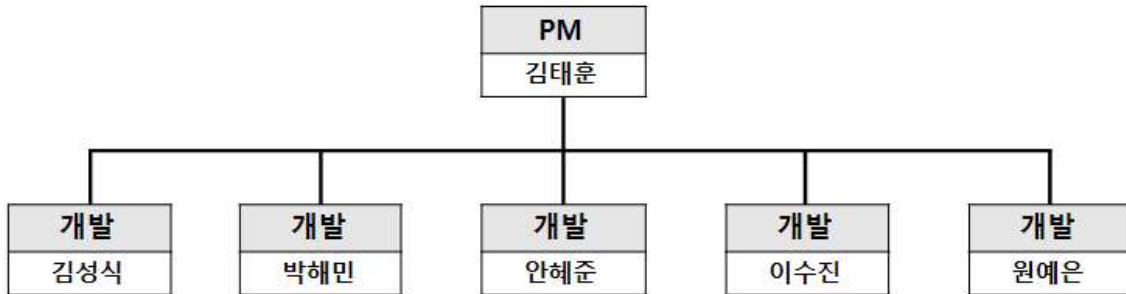
프로젝트 조직

2.1 프로젝트 구성원

2.2 책임 및 역할

02 프로젝트 조직

2.1 프로젝트 구성원



[표2-1] 프로젝트 조직도

2.2 책임 및 역할

구분	세부구분	구성원
PM, 시험 환경 구축	프로젝트 총괄 책임, 시험용 가상 네트워크 구축, 시험용 악성코드 제작	김태훈
개발	DNS 패킷 캡처링 모듈, 판별 결과 전달 모듈 제작	김성식
개발	DGA 도메인 판별 모듈(AI) 제작	박해민 이수진
개발	관리자 대시보드 개발	안혜준
개발	데이터베이스 구축	원예은

[표2-2] 책임 및 역할

03

프로젝트 수행 일정

3.1 프로젝트 추진 일정

3.2 단계별 세부 일정

03 프로젝트 수행 일정

3.1 프로젝트 추진 일정

프로젝트 일정		2020.05.01 ~ 2020.07.04. (약 2개월)								
단계		5월 1주차	5월 2주차	5월 3주차	5월 4주차	6월 1주차	6월 2주차	6월 3주차	6월 4주차	7월 1주차
기획	주제 & 아이디어									
	기획안 & 설계									
개발 및 시험환경 구축	시험 환경 구축									
	DGA 판별 모듈 개발(AI)									
	데이터베이스 구축									
	관리자 대시보드 개발									
	패킷캡쳐, 결과 전달 모듈 개발									
	보완 & 기능 확장									
시험	단위시험									
	통합기능시험									
완성	최종 검토 및 완성									

[표3-1] 프로젝트 추진 일정

3.2 단계별 세부 일정

일정	단계	작업
2020.05.01 ~ 2020.05.24	프로젝트 기획	프로젝트 주제 선정, 아이디어 도출, 기획안 PPT 작성, 프로젝트 범위 확정, 프로젝트 일정 확정, 프로젝트 진행 방향 확정,
2020.05.22 ~ 2020.06.13	개발 및 시험환경 구축	영역 별 개발 환경 조성, 네트워크 설계 및 구축, DGA 도메인 판별 모듈 개발, DB 설계 및 구축, 관리자 대시보드 UI 디자인/개발, DNS 패킷 캡처링 모듈 개발, 판별 결과 전달 모듈 개발, 단위 테스트
2020.06.12 ~ 2020.06.15	모듈 통합 및 연동	각 모듈을 통합하여 시험용 네트워크와 연동
2020.06.15 ~ 2020.07.02	시험 / 이행	1차 기능 시험 미비점 보완 확장 기능 구상 및 적용 최종 기능 시험 최종 미비점 보완
2020.06.25 ~ 2020.07.04	최종 검토 및 완성	최종 검토 및 완성

[표3-2] 프로젝트 추진 일정

04

프로젝트 설계

4.1 프로세스 흐름도

4.2 프로세스 분석

4.3 사용자 인터페이스 분석

4.4 시스템 구성도

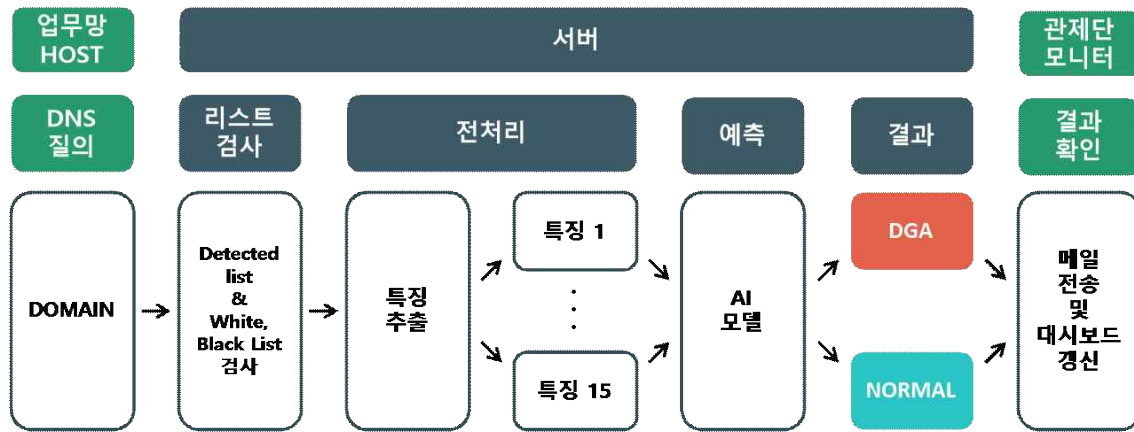
4.5 자료 설계

4.6 시스템 설계

4.7 핵심 알고리즘 (AI 피처)

04 프로젝트 설계

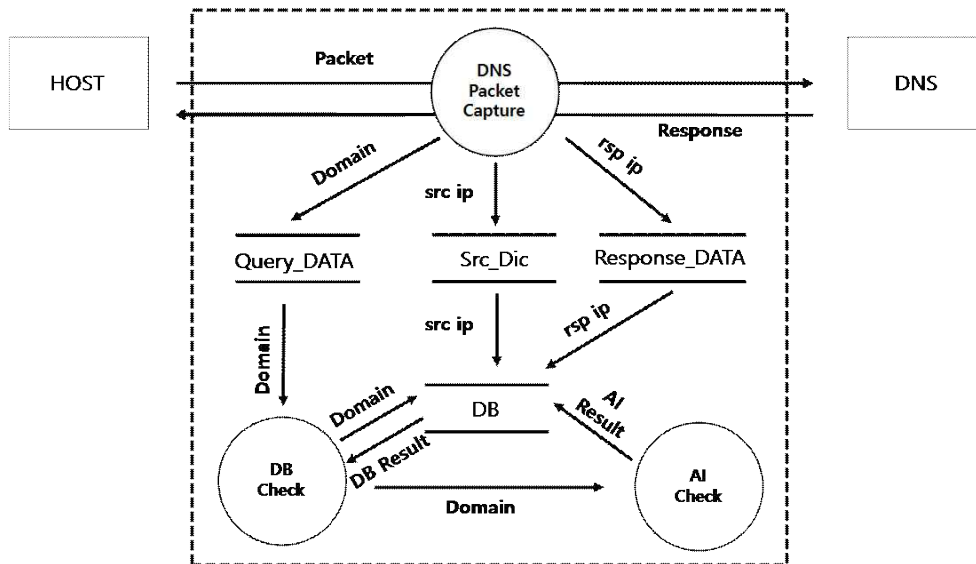
4.1 프로세스 흐름도



[그림4-1] 프로세스 흐름도

4.2 프로세스 분석

4.2.1 자료 흐름도 (DFD)



[그림4-2] 자료 흐름도 (DFD)

4.2.2 기능 명세서

구분	기능	설명
DNS 패킷 캡처링 모듈	DNS 패킷 캡처링	1. 특정 네트워크 구간의 DNS 패킷 캡처 2. DNS 패킷에서 질의 도메인, 출발지IP, 응답IP 추출
	DB조회	1. 화이트리스트 조회 2. 탐지된 도메인 리스트 조회 3. 블랙리스트 조회 4. 금일 탐지 리스트 조회
	다중 처리	멀티 프로세싱을 구현하여 캡처링 및 추출을 동시에 처리
판별 결과 전달 모듈	데이터 전달	1. 금일 탐지 리스트에 데이터 추가 2. 탐지된 도메인 리스트에 데이터 추가
	알림	관리자 메일로 탐지 사실 전달
	표준 출력 메시지	1. 금일 탐지 리스트에 신규 데이터로 추가되는 경우, 파란색 중복 데이터 존재할 경우 보라색으로 라벨링 2. 탐지된 도메인 리스트에 신규 데이터로 추가되는 경우 또는 중복 데이터 존재할 경우 빨간색으로 라벨링
데이터베이스	데이터 저장	1. 탐지된 도메인 리스트 저장 2. 화이트리스트 저장 3. 블랙리스트 저장 4. 금일 탐지 리스트 저장 5. 월간 탐지 리스트 저장 6. 관리자 계정 정보 관리 저장
	데이터 삭제	매일 자정이 되면 금일 탐지 리스트 초기화
관리자 대시보드	관리자 계정 관리	1. 로그인 계정정보 검증 2. 계정 조회 3. 신규 계정 등록 4. 기존 계정 삭제
	도메인 리스트 관리	1. 탐지된 도메인 리스트 조회/삭제 2. 블랙리스트 조회 및 도메인 추가/삭제 3. 화이트리스트 조회 및 도메인 추가/삭제
	탐지 현황 대시보드	1. 금일 탐지율을 원형그래프 형태로 표기 2. 최근 AI판별을 거친 10개의 데이터 중 50% 이상 DGA로 탐지되었을 경우 대시보드의 원형그래프 배경색을 붉게 변경 3. 월간 탐지 현황 표기
	탐지 현황 보고서	일간 보고서 제공
DGA 도메인 판별 모듈 (AI)	데이터셋 관리	1. 중복데이터 삭제 2. 학습에 필요한 데이터셋을 하나의 파일로 통합
	머신러닝 모델	1. 모델 학습 2. 모델 저장/불러오기
	피쳐 추출	1. 15개의 피쳐값 추출 2. 피쳐 스케일링 과정을 통해 정규화 3. 피쳐값 저장 4. 피쳐값 시각화 그래프 출력
	판별 모듈(AI모델)	실시간 판별

[표4-1] 기능 명세서

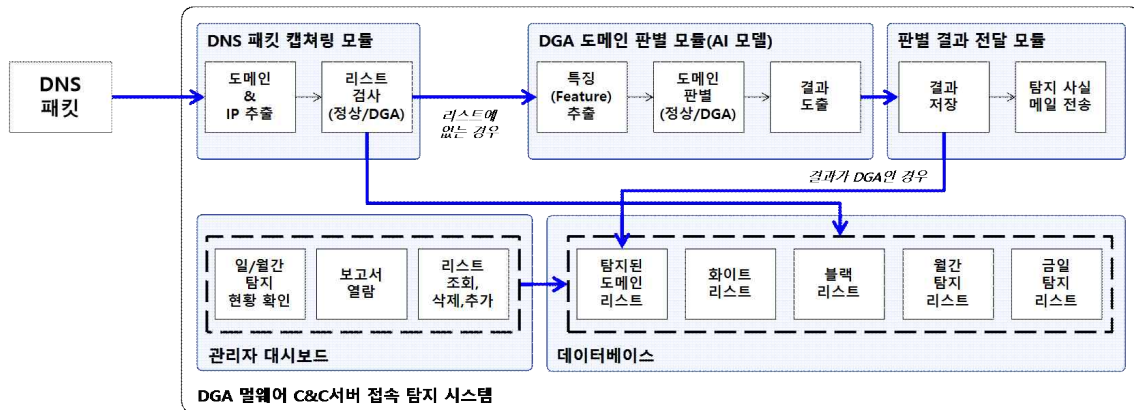
4.3 사용자 인터페이스 분석

4.3.1 입.출력 인터페이스 종류와 기능

종류	기능 설명	I/O 구분
관리자 로그인	사용자가 입력한 ID/PW를 검증하고, 유효한 경우 인증	INPUT
관리자 계정 추가	새로운 사용자가 사용할 ID/PW를 입력받아 계정추가	INPUT
블랙리스트 조회	블랙리스트에 존재하는 도메인 리스트 조회	OUTPUT
블랙리스트 삭제	블랙리스트에 존재하는 도메인 선택하여 삭제	INPUT
블랙리스트 추가	블랙리스트에 새로운 도메인 추가	INPUT
화이트리스트 조회	화이트리스트에 존재하는 도메인 리스트 조회	OUTPUT
화이트리스트 삭제	화이트리스트에 존재하는 도메인 선택하여 삭제	INPUT
화이트리스트 추가	화이트리스트에 새로운 도메인 추가	INPUT
탐지된 도메인 리스트 조회	탐지된 도메인 리스트에 존재하는 도메인 리스트 조회	OUTPUT
탐지된 도메인 리스트 삭제	탐지된 도메인 리스트에 존재하는 도메인 선택하여 삭제	INPUT
일간 보고서	금일 탐지된 DGA 도메인 리스트를 보고서 형태로 출력	OUTPUT
대시보드	수치/차트 형태의 실시간 탐지현황 정보 제공	OUTPUT

[표4-2] 관리자 대시보드 입.출력 인터페이스 구분

4.4 시스템 구성도



[그림4-3] DGA 멀웨어 C&C서버접속 탐지 시스템 구성도

4.5 자료 설계

DB명	테이블명	설명
KS_INFO	CAL	화이트리스트 테이블
	CDL	블랙리스트 테이블
	CTL	탐지된 도메인 리스트 테이블
	TODAY_CTL	금일 탐지 리스트 테이블
	MONTH_C	월간 탐지 리스트 테이블
KS_USER	CUSER	관리자 계정정보 테이블

[표4-3] 데이터베이스 구조

4.5.1 자료 형태 상세 구분

테이블 명		CAL		종류	화이트리스트 테이블		
NO	칼럼ID	칼럼설명	TYPE	길이	NULL여부	PK	비고
1	a_no	색인 번호	int		not null	V	AUTO_INCREMETN
2	a_domain	도메인 데이터	varchar	150	not null		
3	a_time	데이터 추가시간	timestamp		not null		current_timestamp()

[표4-4] 화이트리스트 테이블 구조

테이블 명		CDL		종류	블랙리스트 테이블		
NO	칼럼ID	칼럼설명	TYPE	길이	NULL여부	PK	비고
1	d_no	색인 번호	int		not null	V	AUTO_INCREMETN
2	d_domain	도메인 데이터	varchar	150	not null		
3	d_ip	도메인 매칭 IP	varchar	50	not null		
4	d_time	데이터 추가시간	timestamp		not null		current_timestamp()

[표4-5] 블랙리스트 테이블 구조

테이블 명		CTL		종류	탐지된 도메인 리스트 테이블		
NO	칼럼ID	칼럼설명	TYPE	길이	NULL여부	PK	비고
1	t_no	색인 번호	int		not null	V	AUTO_INCREMETN
2	t_domain	도메인 데이터	varchar	150	not null		
3	t_sip	패킷 출발지 IP	varchar	50	not null		
4	t_dip	도메인 매칭 IP	varchar	50	not null		
5	t_time	데이터 추가시간	timestamp		not null		current_timestamp()

[표4-6] 탐지된 도메인 리스트 테이블 구조

테이블 명		TODAY_CTL		종류	금일 탐지 리스트 테이블		
NO	칼럼ID	칼럼설명	TYPE	길이	NULL여부	PK	비고
1	to_no	색인 번호	int		not null	V	AUTO_INCREMETN
2	to_domain	도메인 데이터	varchar	150	not null		
3	to_dga	도메인 분류용 라벨	varchar	50	not null		DGA 또는 Normal
4	to_sip	패킷 출발지 IP	varchar	50	not null		
5	to_time	데이터 추가시간	timestamp		not null		current_timestamp()

[표4-7] 금일 탐지 리스트 테이블 구조

테이블 명		MONTH_C		종류	월간 탐지 리스트 테이블		
NO	칼럼ID	칼럼설명	TYPE	길이	NULL여부	PK	비고
1	m_no	색인 번호	int		not null	V	AUTO_INCREMENT
2	m_year	년도	int		not null		
3	m_month	월	int		not null		
4	m_count	월간 누적 탐지 갯수	int		not null		

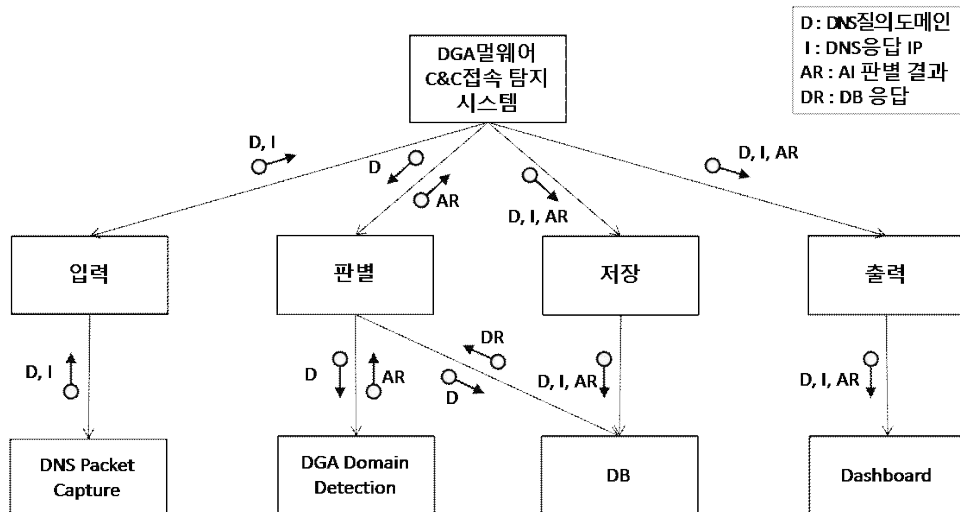
[표4-8] 월간 탐지 리스트 테이블 구조

테이블 명		CUSER		종류	화이트리스트 테이블			
NO	칼럼ID	칼럼설명	TYPE	길이	NULL여부	PK	UNIQUE	비고
1	c_no	색인 번호	int		not null		V	
2	c_id	ID	varchar	50	not null	V		
3	c_pw	PW	blob		not null			

[표4-9] 관리자 계정정보 테이블 구조

4.6 시스템 설계

4.6.1 구조 차트



[그림4-4] 시스템 계층 구조도

4.6.2 모듈 설계

4.6.2.1 DNS 패킷 캡처링 모듈

구분	내부 구성	형태
DNS 패킷 캡처링 모듈	sniffing()	Fuction
	Showpacket(packet)	Fuction
	search_qr(qr_queue, testee_queue)	Fuction
	db_search_today(domain)	Fuction
	db_search_ctl(domain)	Fuction
	db_search_cal(domain)	Fuction
	db_search_cdl(domain)	Fuction
	ML_start()	Fuction
	qr_queue	Queue
	testee_queue	Queue
	rr_data	Dict
	src_dic	Dict

[표4-10] DNS 패킷 캡처링 모듈 구성

모듈이름	Sniffing()
모듈형	Function
기능설명	DNS 패킷을 캡처링한다.
모듈이름	Showpacket(packet)
모듈형	Function
호출모듈	Sniffing()
기능설명	Sniffing() 함수를 통해 수집한 패킷에서 필요한 정보만 추출한다.
모듈이름	search_qr(qr_queue, testee_queue)
모듈형	Function
기능설명	Showpacket() 함수를 통해 추출된 도메인이 DB에 존재하는지 검사한다.
모듈이름	db_search_today(domain)
모듈형	Function
호출모듈	search_qr(qr_queue, testee_queue)
기능설명	Showpacket() 함수를 통해 추출된 도메인이 금일 탐지 리스트에 존재하는지 검사한다.
모듈이름	db_search_ctl(domain)
모듈형	Function
호출모듈	search_qr(qr_queue, testee_queue)
기능 설명	Showpacket() 함수를 통해 추출된 도메인이 탐지된 도메인 리스트에 존재하는지 검사한다.
모듈이름	db_search_cal(domain)
모듈형	Function
호출모듈	search_qr(qr_queue, testee_queue)
기능설명	Showpacket() 함수를 통해 추출된 도메인이 화이트리스트에 존재하는지 검사한다.
모듈이름	db_search_cdl(domain)
모듈형	Function
호출모듈	search_qr(qr_queue, testee_queue)
기능설명	Showpacket() 함수를 통해 추출된 도메인이 블랙리스트에 존재하는지 검사한다.
모듈이름	ML_start()
모듈형	Function

기능설명	도메인을 판별하는 AI를 동작시킨다.
모듈이름	qr_queue
모듈형	Queue
기능설명	DNS Query의 도메인을 저장하는 Queue이다.
모듈이름	testee_queue
모듈형	Queue
기능설명	AI가 판별해야 하는 리스트가 저장되는 Queue이다.
모듈이름	rr_data
모듈형	dict
기능설명	DNS Response 패킷에 들어있는 도메인과 IP를 dict의 형태로 관리한다.
모듈이름	src_dic
모듈형	dict
기능설명	src ip의 정보를 저장하고 관리하는 dict이다.

[표4-11] DNS 패킷 캡처링 모듈의 내부 모듈 상세 설명

4.6.2.2 DGA 도메인 판별 모듈(AI)

구분	내부 구성	형태
DGA 도메인 판별 모듈(AI)	concat(df1, df2, ...)	Function
	drop_duplicates("Domain", keep="first")	Function
	sort_values(["Class"])	Function
	read_csv('.csv')	Function
	to_csv(".csv", index=False)	Function
	pickle.dump()	Function
	pickle.load()	Function
	joblib.dump()	Function
	joblib.load()	Function
	fit(Train_data, Train_label)	Function
	predict(Test_data)	Function
	fit_transform(Train_data)	Function
	transform(Test_data)	Function
	confusion_matrix()	Function
	precision_score()	Function
	metrics.accuracy_score()	Function
	recall_score()	Function
	classification_report()	Function
	scatterplot()	Function
	clf_from_joblib	pkl
	std_scaler	pkl
	3gram	pkl
	4gram	pkl
	5gram	pkl
	googlebooks-eng-10000	txt
	TLD_list	txt

[표4-12] DGA 도메인 판별 모듈(AI) 구성

모듈이름	concat(df1, df2, ...)
모듈형	Function
기능설명	데이터셋 관리 과정 중에 DataFrame을 사용하여 데이터셋을 통합하는 함수이다.
모듈이름	drop_duplicates("Domain", keep="first")
모듈형	Function
기능설명	데이터셋 관리 과정 중에 데이터셋에서 중복되는 도메인이 있으면 중복되는 마지막 도메인을 삭제하는 함수이다.
모듈이름	sort_values(["Class"])
모듈형	Function
기능설명	데이터셋 관리 과정 중에 DGA 알고리즘별로 DGA 도메인들을 정렬하는 함수이다.
모듈이름	read_csv('.csv')
모듈형	Function
기능설명	데이터셋 관리와 AI 모델 학습 과정에서 필요한 데이터셋(csv)을 불러오는 함수이다.
모듈이름	to_csv(".csv",index=False)
모듈형	Function
기능 설명	통합, 중복 삭제, 정렬이 끝난 데이터셋을 학습, 검증용 데이터셋(csv)으로 저장하는 함수로 데이터셋의 인덱스는 저장하지 않는다.
모듈이름	pickle.dump()
모듈형	Function
기능설명	피쳐 계산에 필요한 파일과 피쳐 스케일러를 pkl로 저장하는 함수이다.
모듈이름	pickle.load()
모듈형	Function
기능설명	AI 모델이 실시간 판단할 때 pkl파일로 저장한 피쳐 스케일러와 피쳐 계산에 필요한 파일들을 불러오는 함수이다.
모듈이름	joblib.dump()
모듈형	Function
기능설명	학습한 AI 모델을 실시간 판단에서 학습 없이 바로 실행하기 위해 pkl로 저장하는 함수이다.
모듈이름	joblib.load()
모듈형	Function
기능설명	실시간 판단을 위해 미리 학습해 저장해놓은 AI 모델 파일을 불러오는 함수이다.
모듈이름	fit(Train_data, Train_label)
모듈형	Function
기능설명	학습용 데이터셋으로 모델을 학습시키는 함수이다.
모듈이름	predict(Test_data)
모듈형	Function
기능설명	모델 성능 평가와 실시간 판단 과정에서 도메인이 정상인지 DGA인지 판단하는 함수이다.
모듈이름	fit_transform(Train_data)
모듈형	Function
기능설명	학습용 데이터셋으로 피쳐 스케일러를 학습시키는 함수이다.
모듈이름	transform(Test_data)
모듈형	Function
기능설명	학습한 피쳐 스케일러를 사용해 피쳐값을 정규화하는 함수이다.
모듈이름	confusion_matrix()
모듈형	Function
기능설명	모델 성능 평가 과정에서 성능 평가 지표를 계산하는 함수이다.

모듈이름	precision_score()
모듈형	Function
기능설명	모델 성능 평가 과정에서 모델의 정밀도를 계산하는 함수이다.

모듈이름	metrics.accuracy_score()
모듈형	Function
기능설명	모델 성능 평가 과정에서 모델의 정확도를 계산하는 함수이다.

모듈이름	recall_score()
모듈형	Function
기능설명	모델 성능 평가 과정에서 재현율을 계산하는 함수이다.

모듈이름	classification_report()
모듈형	Function
기능설명	데이터 라벨별 정밀도, 재현율, F1-Score를 계산하는 함수이다.

모듈이름	scatterplot()
모듈형	Function
기능설명	피쳐들의 산점도 그래프를 나타내는 함수이다.

모듈이름	clf_from_joblib
모듈형	pkl
기능설명	pkl로 저장한 AI 모델 파일을 실행시킨다.

모듈이름	std_scaler
모듈형	pkl
기능설명	피쳐값 정규화를 위해 pkl로 저장한 스케일러를 실행시킨다.

모듈이름	3gram
모듈형	pkl
기능설명	3-gram Score 피쳐 계산을 위해 정상 도메인들의 3-gram 빈도수를 저장한 파일이다.

모듈이름	4gram
모듈형	pkl
기능설명	4-gram Score 피쳐 계산을 위해 정상 도메인들의 4-gram 빈도수를 저장한 파일이다.

모듈이름	5gram
모듈형	pkl
기능설명	5-gram Score 피쳐 계산을 위해 정상 도메인들의 5-gram 빈도수를 저장한 파일이다.

모듈이름	googlebooks-eng-10000
모듈형	txt
기능설명	Meaning_count 피쳐 계산에 필요한 영어 단어 사전 파일이다.

모듈이름	TLD_list
모듈형	txt
기능설명	TLD_index 피쳐 계산에 필요한 TLD들의 목록을 저장한 파일이다.

[표4-13] DGA 도메인 판별 모듈(AI) 상세 설명

4.6.2.3 판별 결과 전달 모듈

구분	내부 구성	형태
DNS 패킷 캡처링 모듈	db_insert_today(domain,src_dic[domain],tmp)	Fuction
	db_insert_ctl(domain,src_dic[domain]rr_data[domain])	Fuction
	mailer.py	script file

[표4-14] 판별 결과 전달 모듈 구성

모듈이름	db_insert_today(domain,src_dic[domain],tmp)
모듈형	Function
호출모듈	ML_start()
기능설명	AI가 확인한 결과 값을 TODAY_CTL 테이블에 저장한다.

모듈이름	db_insert_ctl(domain,src_dic[domain]rr_data[domain])
모듈형	Function
호출모듈	ML_start()
기능설명	AI가 확인한 결과 값을 CTL 테이블에 저장한다.

모듈이름	mailer.py
모듈형	script file
기능설명	AI판별 결과가 DGA 도메인일 경우 관리자 메일로 탐지사실을 알린다.

[표4-15] 판별 결과 전달 모듈 상세 설명

4.6.2.4 관리자 대시보드

구분	내부 구성	형태
관리자 대시보드	admin_login	script file
	login_check	script file
	logout	script file
	admin_management	script file
	cal	script file
	ctl	script file
	cdl	script file
	daily_report	script file
	home	script file
	index	script file
	Google Chart API	API

[표4-16] 관리자 대시보드 구성

모듈이름	admin_login
모듈형	script file
호출모듈	login_check
기능설명	대시보드에 접속 시 처음 호출되는 페이지이며 login_check 모듈을 호출한다.

모듈이름	login_check
모듈형	script file
호출모듈	index
기능설명	사용자가 입력한 ID/PW를 검증하여 등록된 계정일 경우 index모듈을 호출한다.

모듈이름	logout
모듈형	script file
기능설명	로그아웃을 위한 모듈이다.

모듈이름	admin_management
모듈형	script file
기능설명	최고 관리자가 서버 관리자 계정 목록을 확인하기 위한 모듈이며 서버 관리자 추가/삭제가 가능하다.

모듈이름	cal
모듈형	script file
기능설명	화이트리스트에 저장된 데이터를 출력하는 모듈이며 도메인 추가/삭제가 가능하다.

모듈이름	ctl
모듈형	script file
기능설명	탐지된 도메인 리스트에 저장된 데이터를 출력하는 모듈이며 목록에 있는 데이터를 블랙리스트로 이동시키거나 삭제가 가능하다.

모듈이름	cdl
모듈형	script file
기능설명	블랙리스트에 저장된 데이터를 출력하는 모듈이며 도메인 추가/삭제가 가능하다.

모듈이름	daily_report
모듈형	script file
기능설명	일간 보고서를 제공하는 모듈로써 보고서 형태로 출력이 가능하다.

모듈이름	home
모듈형	script file
기능설명	대시보드의 메인화면이며 수치 및 차트형태로 데이터를 출력하여 보여준다.

모듈이름	index
모듈형	script file
기능설명	대시보드의 틀이 되는 모듈로써 다른 모듈들을 호출한다.

모듈이름	Google Chart API
모듈형	API
기능설명	원형 차트와 꺾은선 그래프를 사용하기 위한 모듈이다.

[표4-17] 관리자 대시보드 상세 설명

4.6.2.5 데이터베이스

구분	내부 구성	형태
데이터베이스	TC_DELETE (TODAY_CTL DELETE)	EVENT_SCHEDULER
	TCN_SAVE (TODAY_CTL NUM SAVE)	EVENT_SCHEDULER

[표4-18] 데이터베이스 이벤트 스케줄러

모듈이름	TC_DELETE (TODAY_CTL DELETE)
모듈형	EVENT_SCHEDULER
기능설명	매일 자정에 금일 탐지 리스트를 초기화한다.

모듈이름	TCN_SAVE (TODAY_CTL NUM SAVE)
모듈형	EVENT_SCHEDULER
기능설명	금일 탐지 리스트를 1초 주기로 검사하여 DGA로 라벨링된 튜플이 존재할 경우 탐지된 시간 기준으로 몇 년 몇 월인지 체크하여 월간 탐지 리스트의 누적탐지갯수를 증가시킨다.

[표4-19] 데이터베이스 이벤트 스케줄러 상세 설명

4.7 핵심 알고리즘 (AI 피쳐)

4.7.1 TLD_Index

1 if 도메인 TLD in TLD_list :	① 도메인의 TLD가 TLD 순위 파일에 있으면
2 return TLD 순위	② 해당 TLD 순위 반환
3 else	③ 도메인의 TLD가 TLD 순위 파일에 없으면
4 return 0	④ 0 반환

[그림4-5] TLD_Index 피쳐

<https://data.netlab.360.com/dga/>에 따르면 DGA의 TLD는 정상 도메인이 주로 사용하는 com, org, net, info뿐만 아니라 자주 사용하지 않는 TLD들도 사용하는 것을 볼 수 있다.

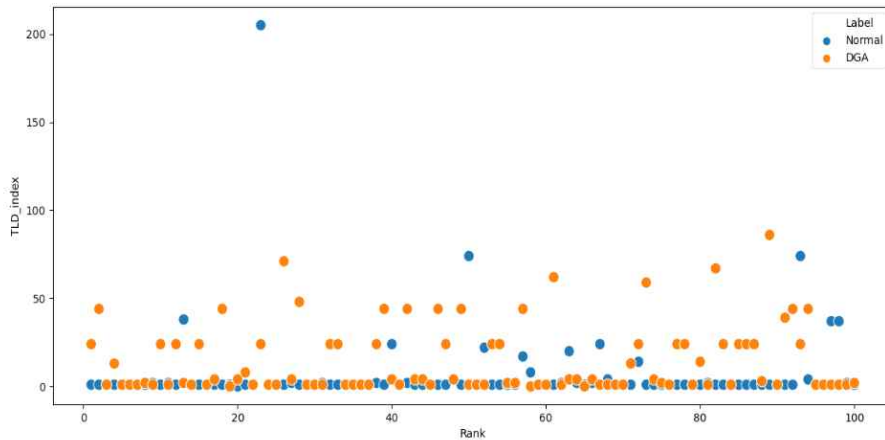
	A	241	.rocks
1	.com	242	.pr.gov.br
2	.org	243	.edu.my
3	.ru	244	.coop
4	.net	245	.gov.vn
5	.de	246	.gov.az
6	.com.br	247	.gov.hk
7	.ir	248	.stream
8	.co.uk	249	.net.pl
9	.pl	250	
10	.it		

[그림4-6] TLD 종류

TLD 순위는 <https://www.hayksaakian.com/most-popular-tlds/>에서 Alexa Top 1Million Ranking 도메인들의 TLD 개수를 선 표를 참고하였다. 그림 2는 Alexa Top 1 Million에서 사용중인 상위 10개



그림 4-8] 정산 도메인과 DGA 도메인 데이터셋의 TID 비교



[그림4-9] 정상/DGA 도메인 TLD 그래프

[그림4-9]는 정상, DGA 도메인 TLD 분포를 나타낸 점 그래프이다. 정상 도메인의 TLD보다 DGA 도메인의 TLD가 높게 나타나고 있음을 확인할 수 있다. 정상 도메인은 The Majestic Million의 1~100위 도메인이며 DGA 도메인은 무작위로 순위를 지정했다.

4.7.2 N-gram Score

```

1 for i in len(서브 도메인) :
2     if 서브 도메인[i:i+N] in N-gram_list
3         sum += N-gram_list(서브 도메인[i:i+N])
4 return sum / len(서브 도메인)

```

- ① 서브 도메인의 길이까지 반복
- ② N-gram_list 파일에 서브 도메인의 부분 문자열[i:i+N]이 있으면
- ③ N-gram_list 파일에 저장되어 있는 부분 문자열 빈도수를 더함
- ④ 전체 빈도수 합을 서브 도메인 길이로 나누어 반환

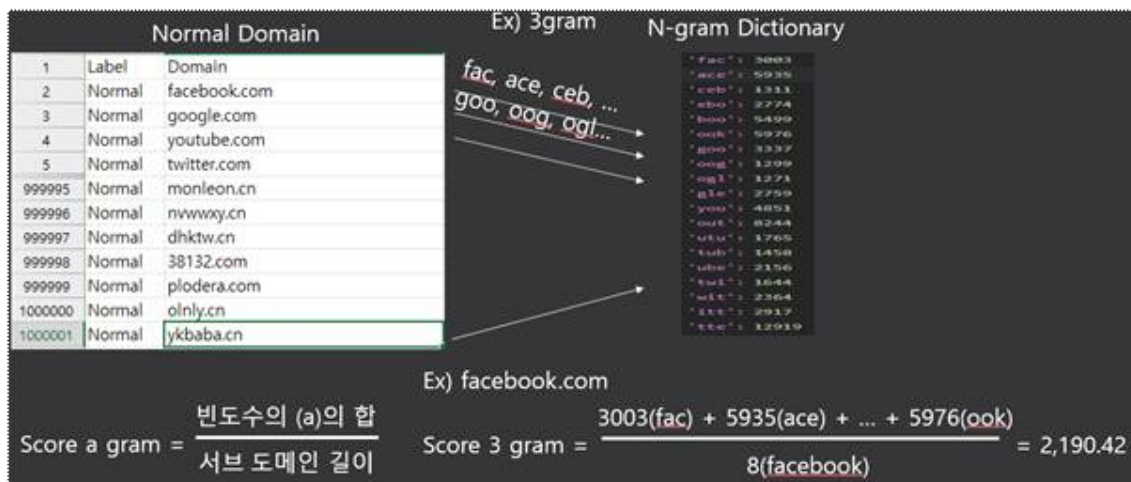
[그림4-10] N-gram Score 피쳐

N-gram score는 서브 도메인을 N-gram으로 자르고, 부분 문자열들의 빈도수를 각각 합하여 서브 도메인의 길이로 나눈 값이다.

$$\text{N-gram Score} = \frac{\text{N의 빈도수 합}}{\text{서브도메인 길이}}$$

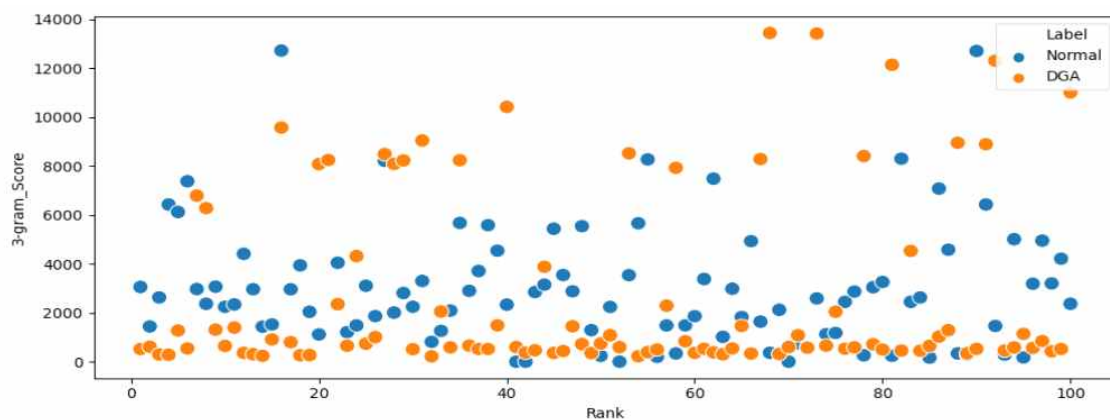
[그림4-11] N-gram Score 수식

부분 문자열들의 빈도수를 구하는 방법은 Majestic Top 1 Million에서 도메인의 서브 도메인들을 N-gram으로 자르고, 모든 부분 문자열들의 빈도수 저장한 N-gram_list 파일에서 불러온다.

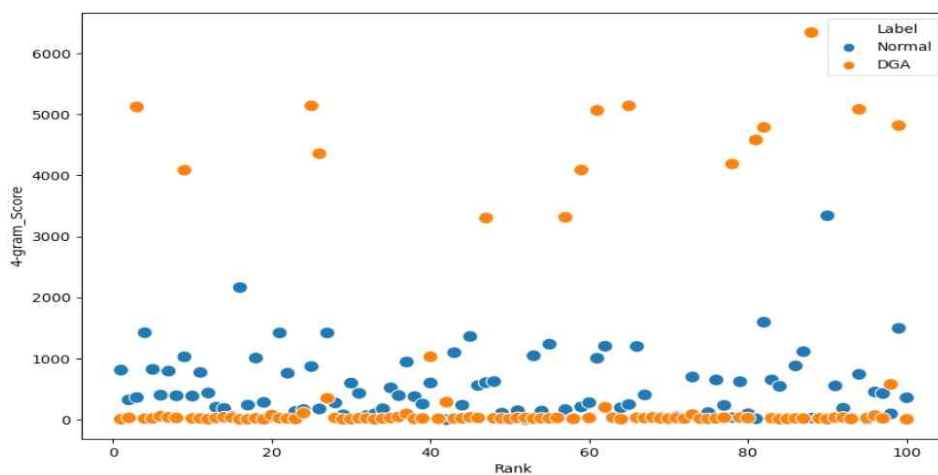


[그림4-12] 3-gram_list

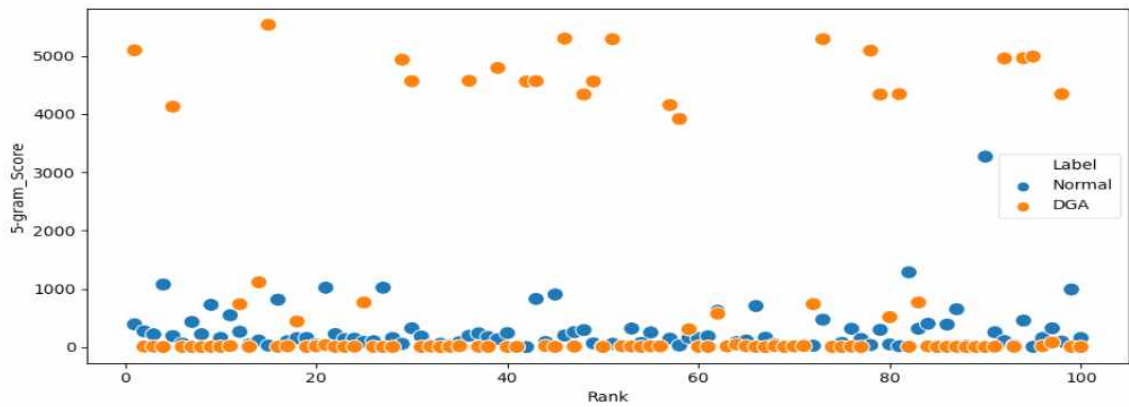
[그림4-12]는 정상 도메인에서 3-gram을 사용하여 3-gram_list를 만드는 방법과 3-gram Score의 계산 방법을 예시로 보여주고 있다.



[그림4-13] 정상/DGA 도메인 3-gram Score



[그림4-14] 정상/DGA 도메인 4-gram Score



[그림4-15] 정상/DGA 도메인 5-gram Score

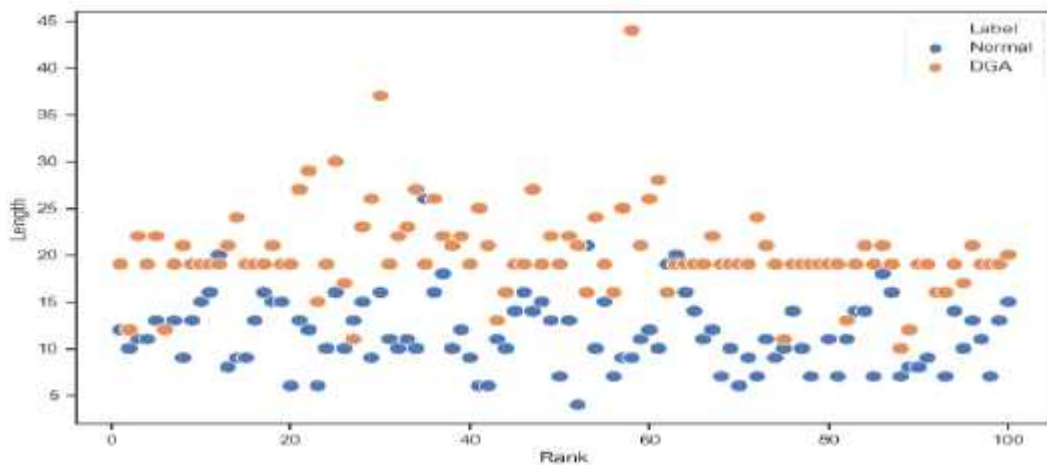
[그림4-13], [그림4-14], [그림4-15]는 정상, DGA 도메인들의 3~5 gram Score를 나타낸 점 그래프이다. 그래프를 보면 정상 도메인의 3~5 gram Score가 DGA의 3~5 gram Score보다 높은 것을 확인할 수 있다.

4.7.3 Length

```
1 Length = len(도메인)
2 return Length
```

[그림4-16] Length 피쳐

Length는 도메인의 전체 길이인 피쳐이다. 정상 도메인의 경우 편의성을 위해 길이가 짧으나 DGA 도메인은 무작위로 생성돼 대부분 길이가 긴 것이 특징이다.



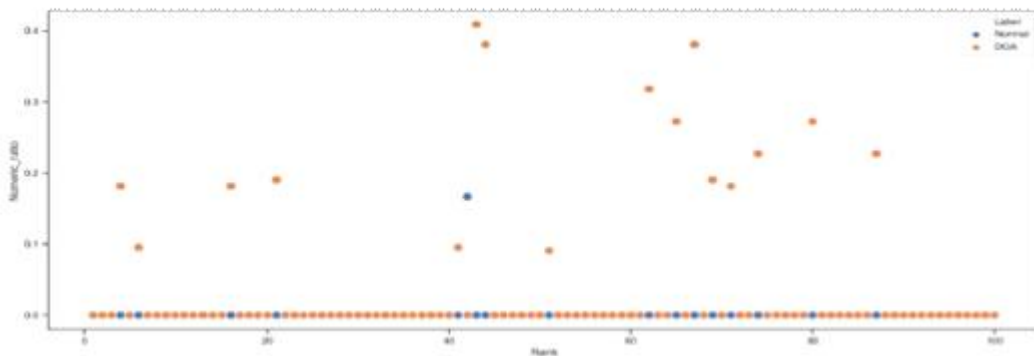
[그림4-17] 정상/DGA 도메인 길이

4.7.4 Numeric_ratio

```
1 Numeric_ratio = count(서브 도메인에 포함된 숫자(0~9)) / len(서브 도메인)
```

[그림4-18] Numeric_ratio 피쳐

Numeric_ratio는 서브 도메인의 숫자 비율로 서브 도메인에 포함된 숫자 개수들을 세고 서브 도메인의 길이로 나눈 피쳐이다. 정상 도메인의 경우 대부분 알파벳으로 이뤄져 있으나 몇몇 DGA 도메인은 알파벳과 숫자가 섞여있는 것이 특징이다.



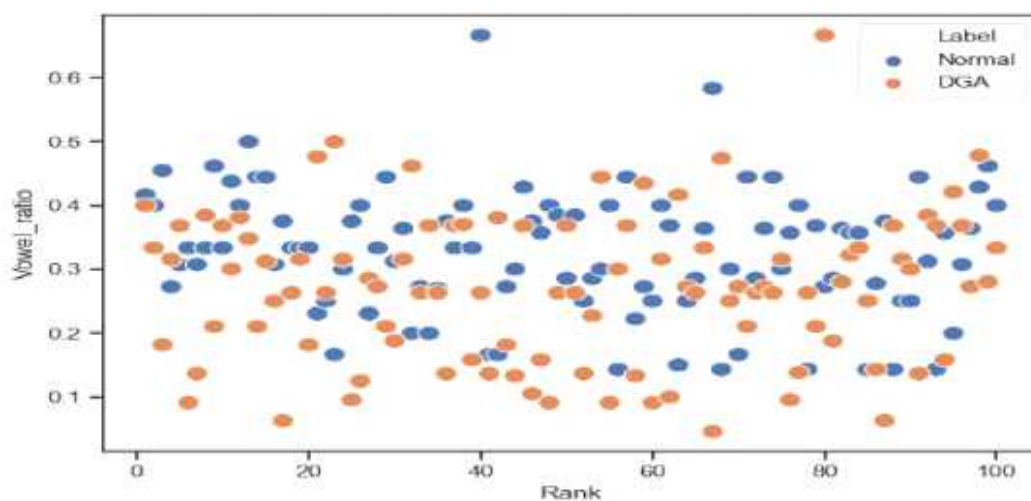
[그림4-19] 정상/DGA 서브 도메인 숫자 비율

4.7.5 Vowel_ratio

```
1 Vowel_ratio = count(서브 도메인에 포함된 모음(a,e,i,o,u)) / len(서브 도메인)
```

[그림4-20] Vowel_ratio 피쳐

Vowel_ratio는 서브 도메인의 모음 비율로 서브 도메인에 포함된 모음(a,e,i,o,u) 개수들을 세고 서브 도메인의 길이로 나눈 피쳐이다. 정상 도메인의 경우 자음과 모음의 비율이 비슷하지만 DGA 도메인의 경우 모음 개수보다는 자음의 개수가 더 많다.



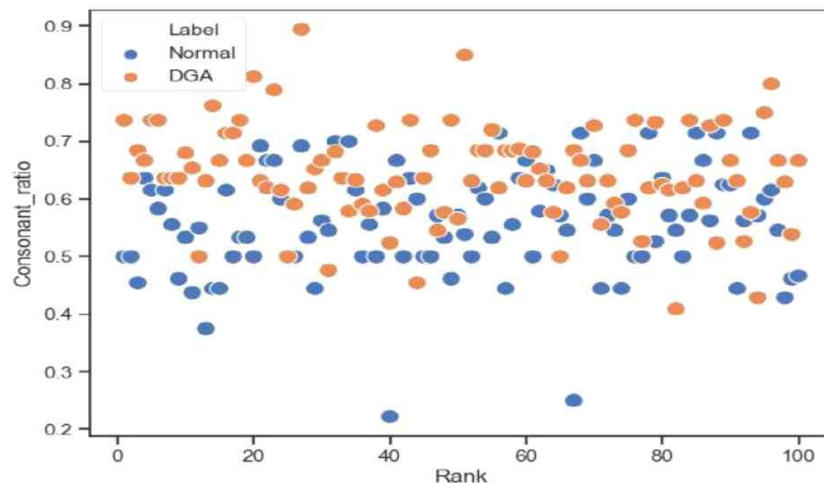
[그림4-21] 정상/DGA 서브 도메인 모음 비율

4.7.6 Consonant_ratio

```
1 Consonant_ratio = count(서브 도메인의 자음(^aeiou)) / len(서브 도메인)
```

[그림4-22] Consonant_ratio 피쳐

Consonant_ratio는 서브 도메인의 자음 비율로 서브 도메인에 포함된 자음 개수들을 세고 서브 도메인의 길이로 나눈 피쳐이다. 정상 도메인의 경우 자음과 모음의 비율이 비슷하지만 DGA 도메인의 경우 무작위로 생성되기 때문에 하나의 문자가 자음으로 생성될 확률이 모음으로 생성될 확률보다 더 높다.



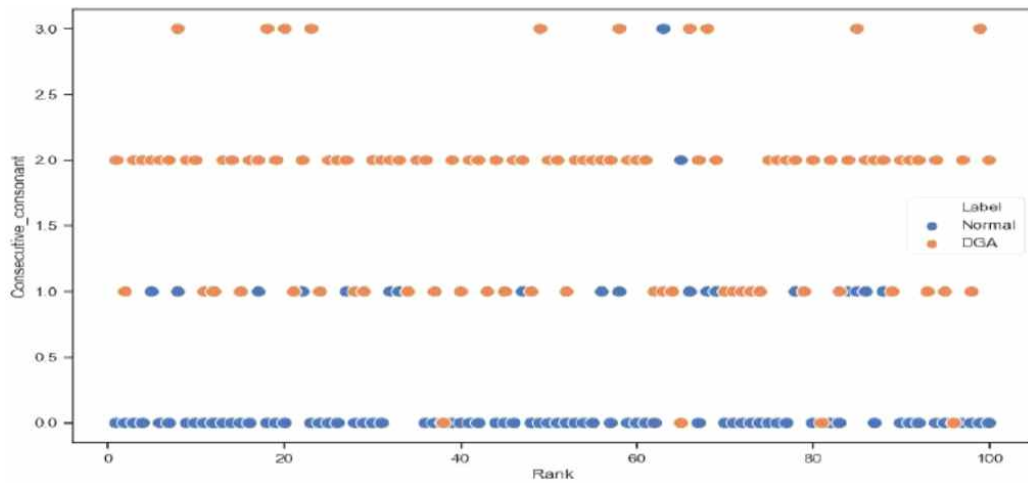
[그림4-23] 정상/DGA 서브 도메인 자음 비율

4.7.7 Consecutive_consonant

```
1 Consecutive_consonant = count(서브 도메인[^.aeiou]{3,})
```

[그림4-24] Consecutive_consonant 피쳐

Consecutive_consonant는 서브 도메인에서 3음절 이상 연속되는 자음, 숫자 문자열 개수를 센 피쳐이다. 알파벳에서 3음절 이상 연속되는 자음 문자열은 발음하기 어려워 정상 도메인에서는 잘 사용되지 않지만 DGA 도메인의 경우 무작위로 생성되고 발음할 이유가 없기 때문에 정상 도메인보다 3음절 이상 연속되는 자음, 숫자 문자열 개수가 많다.



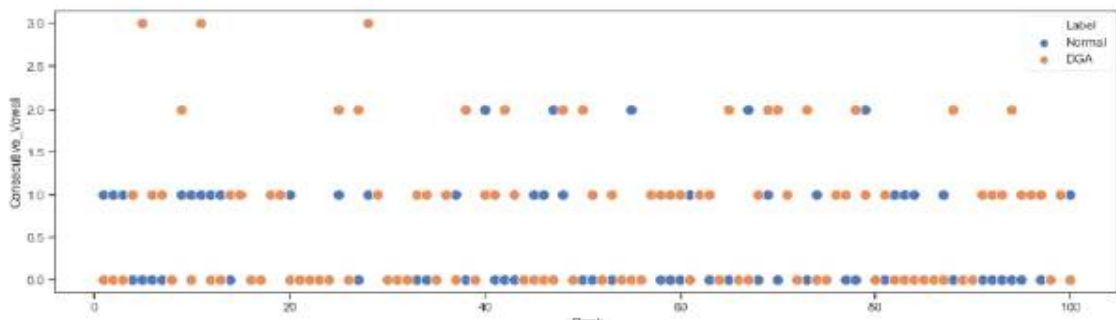
[그림4-25] 정상/DGA 서브 도메인 3음절 이상 연속되는 자음 문자열 개수

4.7.8 Consecutive_Vowel

```
1 Consecutive_Vowel = count(서브 도메인[aeiou]{2,})
```

[그림4-26] Consecutive_Vowel 피쳐

Consecutive_Vowel는 서브 도메인에서 2음절 이상 연속되는 모음 문자열 개수를 센 피쳐이다. 정상 도메인에서 2음절이상 연속되는 모음 문자열 개수는 예시로 facebook, youtube, google, kshielldr과 같이 보통 1개인 경우가 대부분이다. 하지만 DGA 도메인의 경우 랜덤한 위치에서 모음이 생성될 수 있으므로 2음절 이상 연속되는 모음 문자열 개수가 정상 도메인보다 많다.



[그림4-27] 정상/DGA 서브 도메인 2음절 이상 연속되는 모음 문자열 개수

4.7.9 period

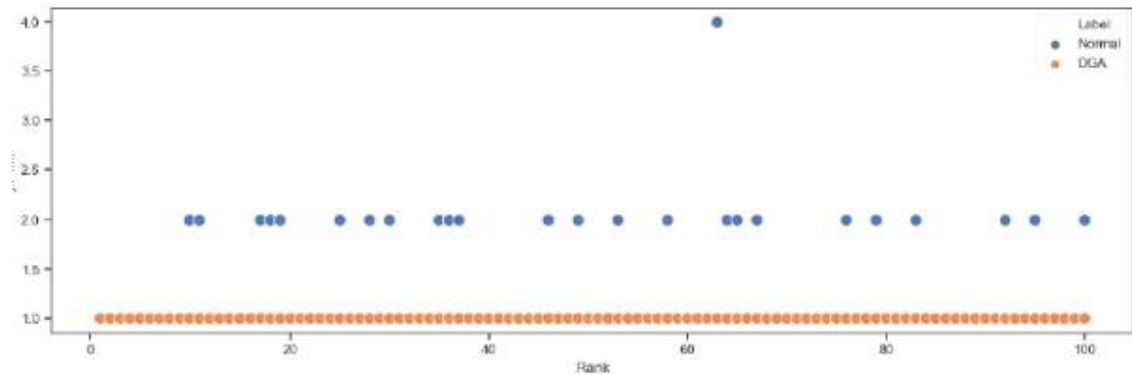
```
1 period = count(도메인[.])
```

[그림4-28] period 피쳐

period는 도메인에서 마침표(.)를 센 피쳐이다.

정상 도메인에서는 TLD뿐만 아니라 편의성을 위해 cafe.naver, blog.naver와 같이 SLD도 사용하기에 마침표(.)가 1개보다 많을 수 있다. 하지만 DGA 도메인의 경우 SLD를 사용할 필요가 없고 TLD에서만

마침표(.)를 사용하기 때문에 개수가 1개이다.



[그림4-29] 정상/DGA 도메인 마침표(.) 개수

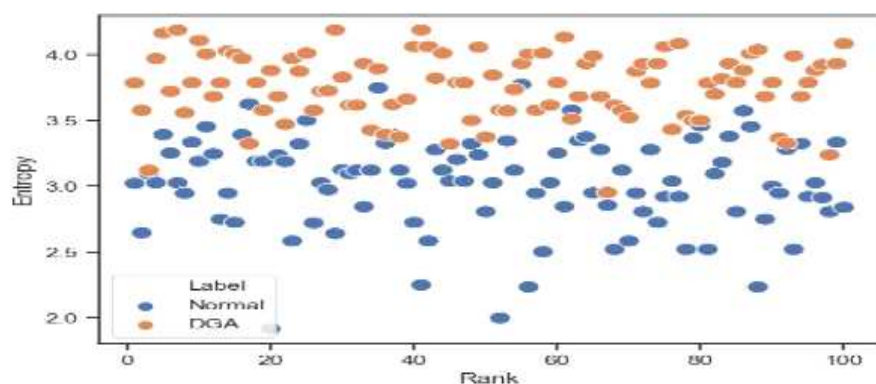
4.7.10 Entropy

1 counts = Counter(도메인)	① Counter를 사용해 도메인의 알파벳과 개수 counts에 저장
2 for alphabet in counts :	② Counts 크기 만큼 반복
3 prob = [alphabet / 도메인 길이]	③ 알파벳별로 확률을 구해 prob 배열에 저장
4 for p in prob:	④ prob 배열 크기 만큼 반복
5 Entropy -= sum(p * log(p) / log(2.0))	⑤ Entropy 공식을 사용해 계산
6 return Entropy	⑥ Entropy 값 반환

[그림4-30] Entropy 피쳐

Entropy는 도메인의 Shannon Entropy 피쳐이다. Shannon Entropy는 모든 사건 정보량의 기대값으로 문자열에서 알파벳의 종류가 많을수록 다음에 오는 문자를 예측할 수 있는 확률이 낮아져 Entropy가 높게 나타나게 된다.

정상 도메인에서는 영어에서 사용 빈도수가 높은 알파벳을 주로 쓰기 때문에 알파벳 종류가 적어 Entropy가 낮지만 DGA 도메인은 무작위의 알파벳과 숫자를 사용하므로 경우의 수가 많아 다음에 올 문자를 예측할 수 있는 확률이 낮기 때문에 상대적으로 정상 도메인에 비해 Entropy가 높게 나타난다.



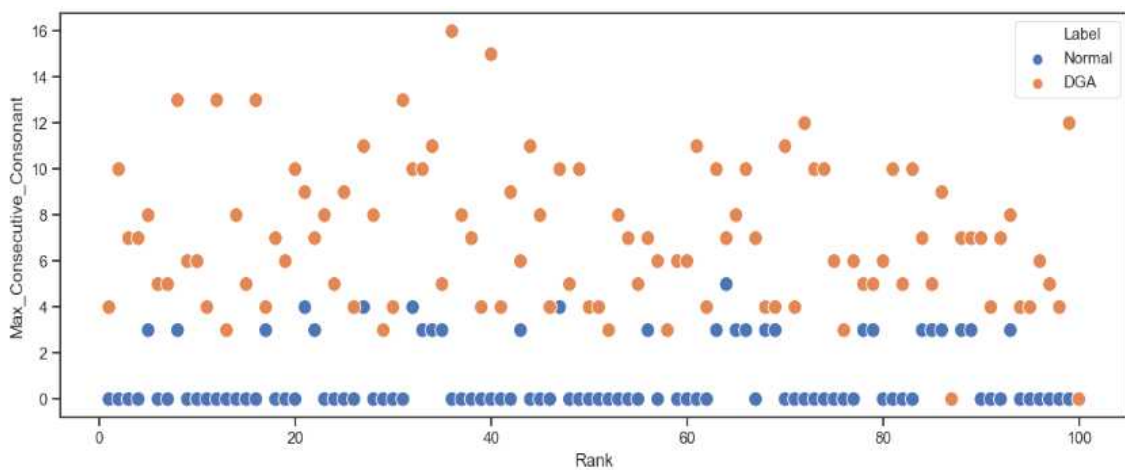
[그림4-31] 정상/DGA 도메인 Entropy

4.7.11 Max_Consecutive_Consonant

```
1 Max_Consecutive_Consonant= len(max(서브 도메인.findall([^\aeiou]{3,})))
```

[그림4-32] Max_Consecutive_Consonant 피쳐

Max_Consecutive_Consonant는 서브 도메인에서 3음절 이상 연속되는 자음, 숫자 문자열 중 최대 길이를 찾는 피쳐이다. Consecutive_Consonant 피쳐처럼 정상 도메인에서는 발음상의 이유로 3음절 이상 연속된 문자열의 최대 길이가 짧다. 하지만 DGA 도메인의 경우 무작위로 생성되고 발음해야 할 이유가 없어 3음절 이상 연속되는 자음 문자열들이 길 확률이 크다.



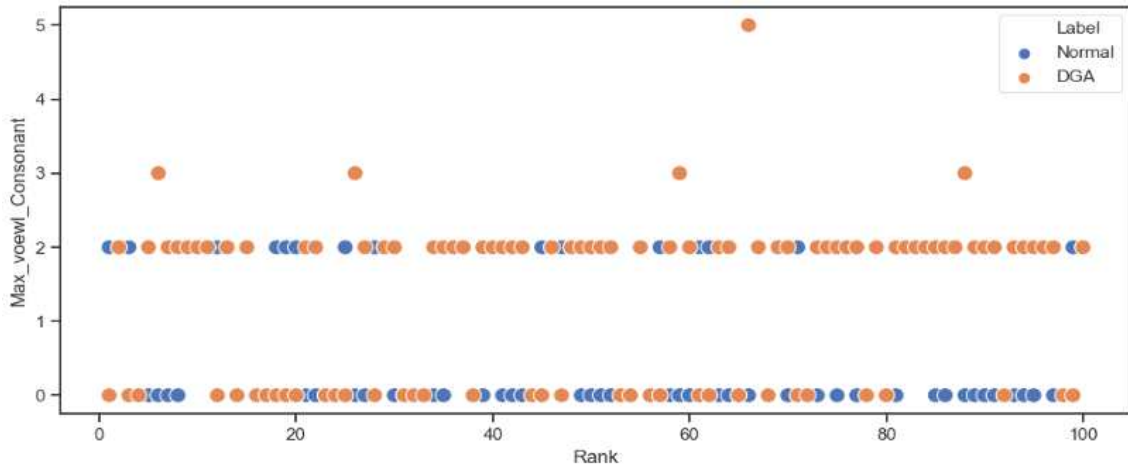
[그림4-33] 정상/DGA 도메인 Max_Consecutive_Consonant

4.7.12 Max_Vowel_Consonant

```
1 Max_Vowel_Consonant= len(max(서브도메인.findall([aeiou]{2,})))
```

[그림4-34] Max_Vowel_Consonant 피쳐

Max_Vowel_Consonant는 서브 도메인에서 2음절 이상 연속되는 모음 문자열 중 최대 길이를 찾는 피쳐이다. 보통 정상 도메인에서는 2음절 이상 연속되는 모음 문자열의 최대 길이가 2음절을 초과하지 않지만 DGA 도메인의 경우 자음, 모음 상관없이 랜덤하게 생성되므로 2음절 이상 연속되는 모음 문자열들의 최대 길이가 길 확률이 크다.



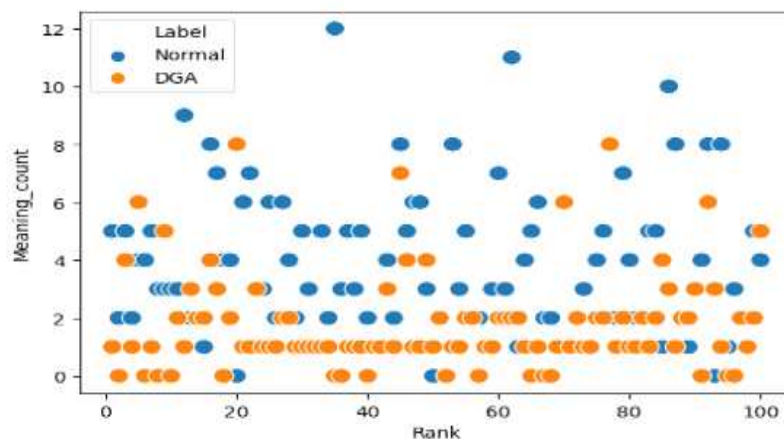
[그림4-35] 정상/DGA 도메인 Max_Vowel_Consonant

4.7.13 Meaning_count

1 for word in word_list :	① 단어 사전 크기 만큼 반복
2 if word in 서브도메인	② 서브 도메인에 단어가 포함돼 있으면
3 counts++	③ 개수를 세는 counts 증가
4 return counts	④ counts 반환

[그림4-36] Meaning_count 피쳐

Meaning_count는 서브 도메인에서 의미있는 단어 개수를 찾는 피쳐이다. 단어 사전은 1500~2008년까지의 서적에서 특정 단어가 얼마나 많이 사용됐는지 검색할 수 있는 엔진인 Google Books Ngram Viewer를 사용했는데 3음절 이상의 영어 단어 중 상위 10,000개를 추출해서 만들었다. 정상 도메인은 자주 사용하는 영어 단어들로 이루어져 있지만 DGA 도메인은 의미 없이 랜덤하게 생성되므로 도메인에 의미 있는 단어가 있을 확률이 낮다.



[그림4-37] 정상/DGA 도메인 Meaning_count

05

시스템 구현

5.1 개발 언어 및 라이브러리

5.2 구현 결과

5.3 시스템 시험

05 시스템 구현

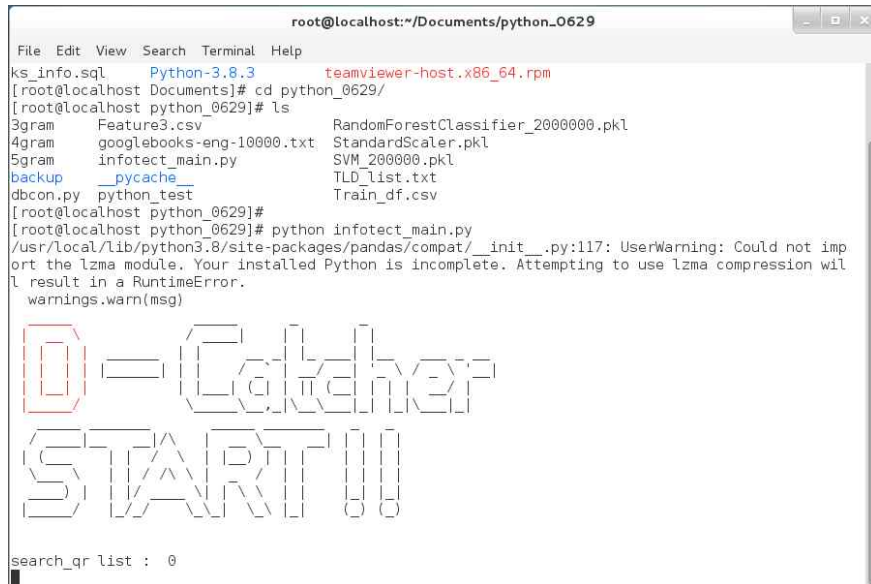
| 5.1 개발 언어 및 라이브러리

구분	개발언어	라이브러리	설명
DNS 패킷 캡처링 모듈, 판별 결과 전달 모듈	Python 3.8.3	Datetime	날짜/시간 관련 라이브러리
		Time	시간 관련 라이브러리
		Multiprocessing	다중 처리를 위한 라이브러리
		Threading	쓰레딩을 위한 라이브러리
		Scapy	패킷을 수집, 변조하는 라이브러리
		Pymysql	MySQL을 사용하기 위한 라이브러리
데이터베이스	MySQL		
관리자 대시보드	HTML		
	PHP		
	Javascript		
DGA 도메인 판별 모듈(AI)	Python 3.8.3	Collections	Entropy 피쳐 계산을 위해 리스트 원소의 개수를 세는 라이브러리
		Joblib	모델을 .pkl 파일로 저장, 불러오는 라이브러리
		Itertools	피쳐 알고리즘에서 반복자를 만드는 라이브러리
		Math	Entropy 계산을 위한 수학 라이브러리
		Matplotlib	데이터 시각화 라이브러리
		Numpy	행렬, 리스트를 처리하는 라이브러리
		Pickle	피쳐 계산에 필요한 파일(단어 사전, TLD_list, N-gram_list)과 피쳐 스케일러를 .pkl 파일로 저장, 불러오는 라이브러리
		Pandas	데이터 조작 및 분석 라이브러리
		Sklearn	머신러닝 모델 라이브러리
		Seaborn	Matplotlib을 기반으로 향상된 데이터 시각화를 위한 라이브러리

[표 5-1] 개발 언어 및 라이브러리

| 5.2 구현 결과

5.2.1 DNS 패킷 캡처링 모듈



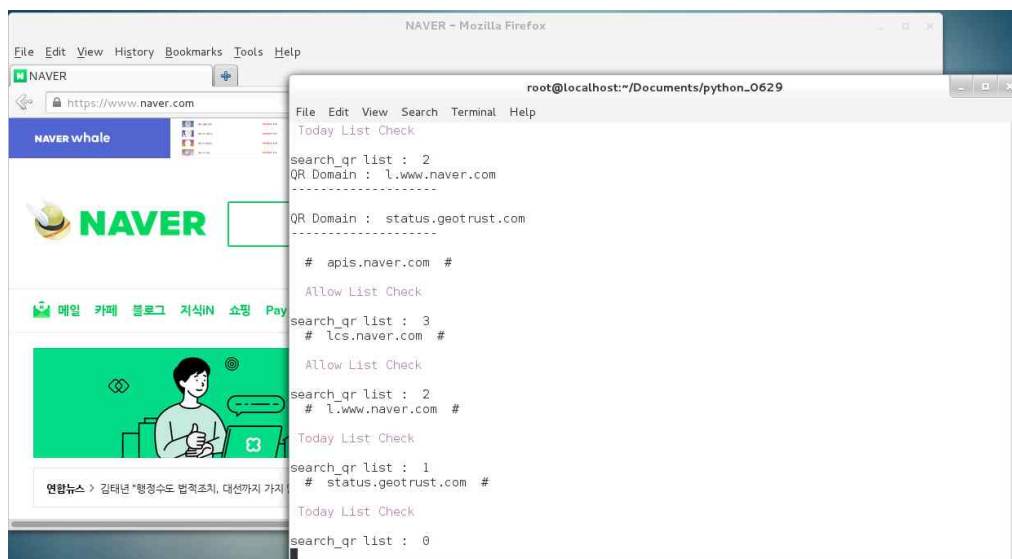
A terminal window titled 'root@localhost:~/Documents/python_0629' showing the execution of a Python script. The script lists files in the current directory, including 'ks_info.sql', 'Python-3.8.3', 'teamviewer-host.x86_64.rpm', '3gram', 'Feature3.csv', 'RandomForestClassifier_2000000.pkl', '4gram', 'googlebooks-eng-10000.txt', 'StandardScaler.pkl', '5gram', 'infotect_main.py', 'SVM_2000000.pkl', 'backup', '__pycache__', 'TLD_list.txt', 'dbcon.py', 'python_test', and 'Train_df.csv'. It then runs 'python infotect_main.py', which displays a large ASCII art graphic that reads '0-Catcher START!!'. Below the graphic, it shows 'search_qr list : 0'.

```
root@localhost:~/Documents/python_0629
File Edit View Search Terminal Help
ks_info.sql Python-3.8.3 teamviewer-host.x86_64.rpm
[root@localhost Documents]# cd python_0629/
[root@localhost python_0629]# ls
3gram      Feature3.csv      RandomForestClassifier_2000000.pkl
4gram      googlebooks-eng-10000.txt  StandardScaler.pkl
5gram      infotect_main.py   SVM_2000000.pkl
backup     __pycache__       TLD_list.txt
dbcon.py   python_test       Train_df.csv
[root@localhost python_0629]#
[root@localhost python_0629]# python infotect_main.py
/usr/local/lib/python3.8/site-packages/pandas/compat/_init_.py:117: UserWarning: Could not import the lzma module. Your installed Python is incomplete. Attempting to use lzma compression will result in a RuntimeError.
  warnings.warn(msg)

0-Catcher
START!!

search_qr list : 0
```

[그림5-1] 프로그램 실행



A terminal window titled 'root@localhost:~/Documents/python_0629' showing the execution of a Python script. The script performs a 'Today List Check' and displays the following output: 'search_qr list : 2', 'QR Domain : l.www.naver.com', 'QR Domain : status.geotrust.com', '# apis.naver.com #', 'Allow List Check', 'search_qr list : 3', '# lcs.naver.com #', 'Allow List Check', 'search_qr list : 2', '# l.www.naver.com #', 'Today List Check', 'search_qr list : 1', '# status.geotrust.com #', 'Today List Check', and 'search_qr list : 0'. In the background, a Mozilla Firefox browser window shows the Naver homepage.

```
NAVER - Mozilla Firefox
File Edit View History Bookmarks Tools Help
NAVER
https://www.naver.com
NAVER whale
NAVER
메일 카페 블로그 지식IN 쇼핑 Pay
연립뉴스 > 김태년 "행정수도 법적조치, 대선까지 가지"

root@localhost:~/Documents/python_0629
File Edit View Search Terminal Help
Today List Check
search_qr list : 2
QR Domain : l.www.naver.com
QR Domain : status.geotrust.com
# apis.naver.com #
Allow List Check
search_qr list : 3
# lcs.naver.com #
Allow List Check
search_qr list : 2
# l.www.naver.com #
Today List Check
search_qr list : 1
# status.geotrust.com #
Today List Check
search_qr list : 0
```

[그림5-2] 프로그램 동작

5.2.2 데이터베이스

```
MariaDB [KS_INFO1]> DESC CAL;
+-----+-----+-----+-----+-----+-----+
| Field | Type      | Null | Key | Default      | Extra      |
+-----+-----+-----+-----+-----+-----+
| a_no  | int(11)   | NO   | PRI | NULL         | auto_increment |
| a_domain | varchar(50) | NO   |     | NULL         |              |
| a_time | timestamp | NO   |     | current_timestamp() |              |
+-----+-----+-----+-----+-----+-----+
3 rows in set (0.00 sec)
```

[그림5-3] 화이트리스트 테이블

```
MariaDB [KS_INFO1]> DESC CDL;
+-----+-----+-----+-----+-----+-----+
| Field | Type      | Null | Key | Default      | Extra      |
+-----+-----+-----+-----+-----+-----+
| d_no  | int(11)   | NO   | PRI | NULL         | auto_increment |
| d_domain | varchar(50) | NO   |     | NULL         |              |
| d_ip  | varchar(50) | NO   |     | NULL         |              |
| d_time | timestamp | NO   |     | current_timestamp() |              |
+-----+-----+-----+-----+-----+-----+
4 rows in set (0.00 sec)
```

[그림5-4] 블랙리스트 테이블

```
MariaDB [KS_INFO1]> DESC CTL;
+-----+-----+-----+-----+-----+-----+
| Field | Type      | Null | Key | Default      | Extra      |
+-----+-----+-----+-----+-----+-----+
| t_no  | int(11)   | NO   | PRI | NULL         | auto_increment |
| t_domain | varchar(50) | NO   |     | NULL         |              |
| t_sip | varchar(50) | NO   |     | NULL         |              |
| t_dip | varchar(50) | NO   |     | NULL         |              |
| t_time | timestamp | NO   |     | current_timestamp() |              |
+-----+-----+-----+-----+-----+-----+
5 rows in set (0.00 sec)
```

[그림5-5] 탐지된 도메인 리스트 테이블

```
MariaDB [KS_INFO1]> DESC TODAY_CTL;
+-----+-----+-----+-----+-----+-----+
| Field | Type      | Null | Key | Default      | Extra      |
+-----+-----+-----+-----+-----+-----+
| to_no | int(11)   | NO   | PRI | NULL         | auto_increment |
| to_domain | varchar(50) | NO   |     | NULL         |              |
| to_sip | varchar(50) | NO   |     | NULL         |              |
| to_dga | varchar(50) | NO   |     | NULL         |              |
| to_time | timestamp | NO   |     | current_timestamp() |              |
+-----+-----+-----+-----+-----+-----+
5 rows in set (0.01 sec)
```

[그림5-6] 금일 탐지 리스트 테이블

```
MariaDB [KS_INFO1]> DESC MONTH_C;
+-----+-----+-----+-----+-----+-----+
| Field | Type      | Null | Key | Default      | Extra      |
+-----+-----+-----+-----+-----+-----+
| m_no  | int(11)   | NO   | PRI | NULL         | auto_increment |
| m_year | int(11)   | NO   |     | NULL         |              |
| m_month | int(11)   | NO   |     | NULL         |              |
| m_count | int(11)   | NO   |     | NULL         |              |
+-----+-----+-----+-----+-----+-----+
4 rows in set (0.01 sec)
```

[그림5-7] 월간 탐지 리스트 테이블

```
MariaDB [KS_USER]> DESC CUSER;
```

Field	Type	Null	Key	Default	Extra
c_no	int(11)	NO	UNI	NULL	auto_increment
c_id	varchar(50)	NO	PRI	NULL	
c_pw	varchar(50)	NO		NULL	

3 rows in set (0.01 sec)

[그림5-8] 관리자 계정정보 테이블

5.2.3 판별 결과 전달 모듈

root@localhost:~/Documents/python_0629

File Edit View Search Terminal Help

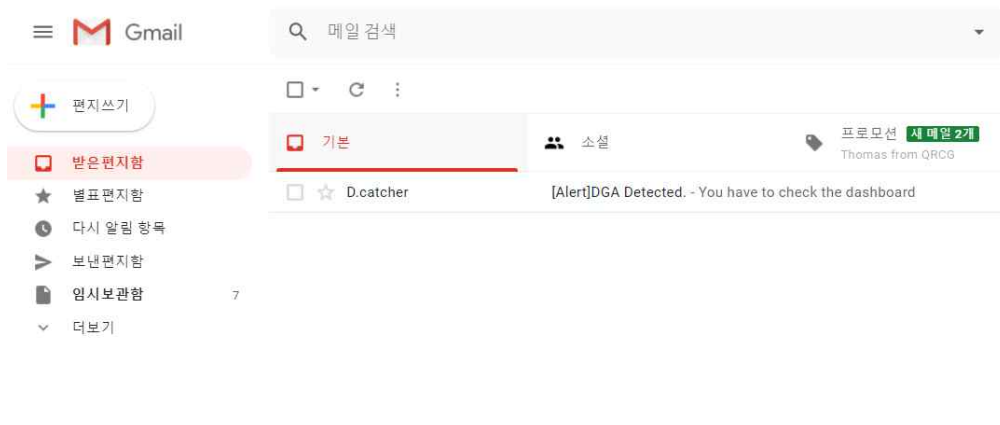
pen a new tab

126	ftp.nara.wide.ad.jp	192.168.193.150	DGA	2020-07-16 16:01:32
127	mirror.navercorp.com	192.168.193.150	Normal	2020-07-16 16:01:33
127	mirrors.cat.net	192.168.193.150	Normal	2020-07-16 16:01:33
128	mirror.opensourcelab.co.kr	192.168.193.150	Normal	2020-07-16 16:01:33
129	ftp.jaist.ac.jp	192.168.193.150	Normal	2020-07-16 16:01:33
130	mirrors.fedoraproject.org	192.168.193.150	Normal	2020-07-16 16:01:34
131	ftp.iiij.ad.jp	192.168.193.150	DGA	2020-07-16 16:01:34
132	yum.mariadb.org	192.168.193.150	Normal	2020-07-16 16:01:34
133	linux.teamviewer.com	192.168.193.150	Normal	2020-07-16 16:01:35
134	mirrors.tuna.tsinghua.edu.cn	192.168.193.150	Normal	2020-07-16 16:01:35
135	mirror.earthlink.iq	192.168.193.150	Normal	2020-07-16 16:01:35
136	fedora.ipserverone.com	192.168.193.150	Normal	2020-07-16 16:01:35
137	my.mirrors.thegigabit.com	192.168.193.150	Normal	2020-07-16 16:01:35
138	download.nus.edu.sg	192.168.193.150	Normal	2020-07-16 16:01:35
139	mirrors.thzhost.com	192.168.193.150	Normal	2020-07-16 16:01:35
140	hk.mirrors.thegigabit.com	192.168.193.150	Normal	2020-07-16 16:01:35
141	mirrors.bestthaihost.com	192.168.193.150	Normal	2020-07-16 16:01:35
142	ctftime.localdomain	168.126.63.1	Normal	2020-07-26 12:01:47
143	ctftime	168.126.63.1	Normal	2020-07-26 12:01:48
144	www.gstatic.com	192.168.193.150	Normal	2020-07-26 12:01:51
145	ctftime.org	192.168.193.150	Normal	2020-07-26 12:01:52
146	ocsp.int-x3.letsencrypt.org	192.168.193.150	Normal	2020-07-26 12:01:52
147	platform.twitter.com	192.168.193.150	Normal	2020-07-26 12:01:56
148	mc.yandex.ru	192.168.193.150	Normal	2020-07-26 12:01:56
149	syndication.twitter.com	192.168.193.150	Normal	2020-07-26 12:01:57
150	yandex.ocsp-responder.com	192.168.193.150	Normal	2020-07-26 12:01:57

115 rows in set (0.04 sec)

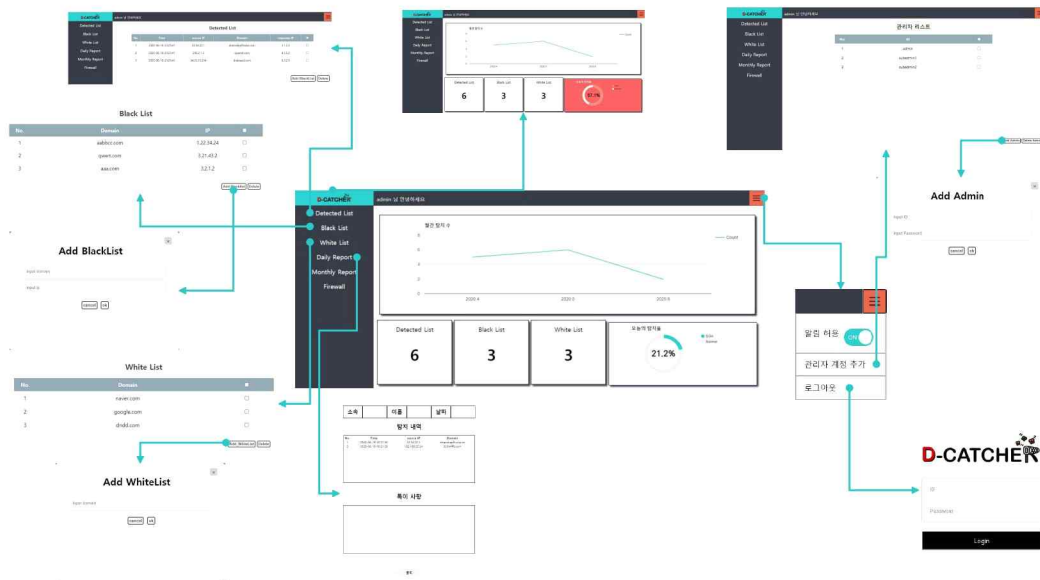
MariaDB [KS_INF0]>

[그림5-10] 판별 결과 DB에 저장



[그림5-11] 관리자 메일로 탐지 사실 전달

5.2.4 관리자 대시보드



[그림5-12] 관리자 대시보드 화면 구조도



- 1 관리자의 ID 입력
- 2 관리자의 Password 입력
- 3 로그인 수행 버튼

[그림5-13] 관리자 대시보드 로그인 화면



- 1 통계 화면
- 4 White List 출력
- 7 방화벽 연동(미구현)
- 2 Detected List 출력
- 5 일간 보고서 출력
- 8 설정 메뉴 바
- 3 Black List 출력
- 6 월간 보고서 출력(미구현)
- 9 List 및 보고서 출력 화면

[그림5-14] 관리자 대시보드 홈 화면



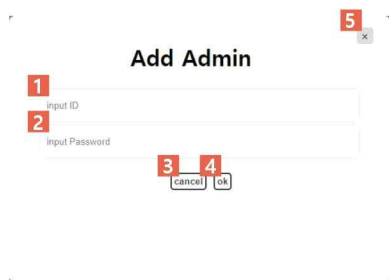
- 1 DGA 탐지 시 알림 허용/차단
- 2 서브 관리자 관리
- 3 로그아웃

[그림5-15] 관리자 대시보드 메뉴 탭



- 1 서버 관리자 목록
- 2 서버 관리자 추가 팝업창 열림
- 3 서버 관리자 삭제

[그림5-16] 관리자 대시보드 관리자 관리 화면



- 1 추가할 관리자 ID 입력
- 2 추가할 관리자 Password 입력
- 3 관리자 추가 취소
- 4 관리자 추가
- 5 팝업창 닫힘

[그림5-17] 관리자 대시보드 관리자 계정 추가 팝업

No.	Time	source IP	Domain	response IP	
1	2020-06-18 23:25:41	22.34.22.1	chirokepfake.com	3.12.3	<input type="checkbox"/>
2	2020-06-18 23:25:41	292.21.3	opend.com	4.3.5.2	<input type="checkbox"/>
3	2020-06-18 23:25:41	34.23.33.234	dnfeed.com	6.3.2.5	<input type="checkbox"/>

- 1 탐지된 DGA 도메인 목록
- 2 선택된 목록 블랙 리스트에 추가
- 3 선택된 목록 탐지 목록에서 삭제

[그림5-18] 관리자 대시보드 탐지리스트 화면

No.	Domain	
1	naver.com	<input type="checkbox"/>
2	google.com	<input type="checkbox"/>
3	dndd.com	<input type="checkbox"/>

- 1 허용할 도메인 목록
- 2 추가할 도메인 입력할 팝업창 열림
- 3 선택된 도메인 삭제

[그림5-19] 관리자 대시보드 화이트리스트 화면



[그림5-20] 관리자 대시보드 화이트리스트 추가 팝업

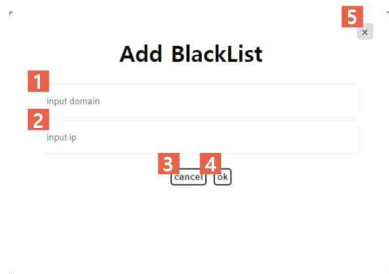
1

No.	Domain	IP	
1	aabcc.com	1.22.34.24	<input type="checkbox"/>
2	qwert.com	3.21.43.2	<input type="checkbox"/>
3	aaa.com	3.2.1.2	<input type="checkbox"/>

2 3
Add Blacklist Delete

-
- 1 차단할 도메인 목록
 - 2 추가할 도메인과 IP 입력할 팝업창 열림
 - 3 선택된 도메인과 IP 삭제

[그림5-21] 관리자 대시보드 블랙리스트 화면



Add BlackList

1 Input domain

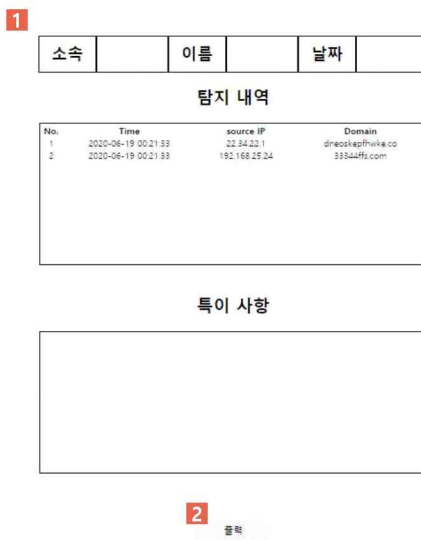
2 Input ip

3 cancel 4 ok

5 X

- 1 차단하고 싶은 도메인 입력
- 2 차단하고 싶은 IP 입력
- 3 Black List에 추가 취소
- 4 Black List에 추가
- 5 팝업창 닫힘

[그림5-22] 관리자 대시보드 블랙리스트 추가 팝업



1

소속	이름	날짜	
탐지 내역			
No.	Time	source IP	Domain
1	2020-06-19 00:21:33	22.34.22.1	dineosapphulie.co
2	2020-06-19 00:21:33	192.168.25.24	333a4ffz.com

특이 사항

2 출력

- 1 일간 보고서
- 2 일간 보고서 출력

[그림5-23] 관리자 대시보드 일간 보고서 화면

5.2.5 DGA 도메인 판별 모듈(AI)

```

DGA Domain List
# The list contains four columns:
#   DGA family, Domain, Start and end of valid time(UTC)
#
# Feed Provided By: netlab.360
# netlab@360.cn
#
# Mirai scanner daily statistics and bot IP check
# data.netlab.360.com/mirai-scanner
# DGA domain data feed
# data.netlab.360.com/dga
# Exploit Kit data feed
# data.netlab.360.com/ek
# All data provided by netlab@360.cn
# data.netlab.360.com
# About Network Security Research Lab at 360
# netlab.360.com
nyman: jwxygozsf.info 2020-07-26 00:00:00 2020-07-26 23:59:59
nyman: ulicwledoms.org 2020-07-26 00:00:00 2020-07-26 23:59:59
nyman: aligier.net 2020-07-26 00:00:00 2020-07-26 23:59:59
nyman: byethditz.net 2020-07-26 00:00:00 2020-07-26 23:59:59
nyman: bmoanngm.com 2020-07-26 00:00:00 2020-07-26 23:59:59
nyman: lagfbowvru.net 2020-07-26 00:00:00 2020-07-26 23:59:59
nyman: somarrts.net 2020-07-26 00:00:00 2020-07-26 23:59:59
nyman: rikouhe.com 2020-07-26 00:00:00 2020-07-26 23:59:59
nyman: wqkxh.org 2020-07-26 00:00:00 2020-07-26 23:59:59
nyman: bnawee.info 2020-07-26 00:00:00 2020-07-26 23:59:59
nyman: nffqzqs.net 2020-07-26 00:00:00 2020-07-26 23:59:59

```

```

#####
1 ## Domain feed of known DGA domains from -2 to +3 days
2 ## HIGH CONFIDENCE DOMAINS ONLY
3 ##
4 ## Feed generated at: Sun Jul 26 00:40:02 UTC 2020
5 ##
6 ## Feed Provided By: John Bambenek of Bambenek Consulting
7 ## jcb@bambenekconsulting.com // http://bambenekconsulting.com
8 ##
9 ##
10 ## Use of this feed is governed by the license here:
11 ## http://osint.bambenekconsulting.com/license.txt
12 ## For more information on this feed go to:
13 ## http://osint.bambenekconsulting.com/manual/dga-feed.txt
14 ##
15 #####
16 Jmksoull Domain u ##### http://osint.bambenekconsulting.com/manual/cl.txt
17 wqjuahtl Domain u ##### http://osint.bambenekconsulting.com/manual/cl.txt
18 knokmim Domain u ##### http://osint.bambenekconsulting.com/manual/cl.txt
19 stjgwigw Domain u ##### http://osint.bambenekconsulting.com/manual/cl.txt
20 ndgexnm Domain u ##### http://osint.bambenekconsulting.com/manual/cl.txt
21 onvgbden Domain u ##### http://osint.bambenekconsulting.com/manual/cl.txt
22 ouesdofy Domain u ##### http://osint.bambenekconsulting.com/manual/cl.txt
23 pfngtpak Domain u ##### http://osint.bambenekconsulting.com/manual/cl.txt
24 rlukxferm Domain u ##### http://osint.bambenekconsulting.com/manual/cl.txt

```

	A	B	C
1	Label	Domain	Class
2	DGA	ab7bec0d0ad6.com	ccleaner
3	DGA	ab15aa3ee0ad3.com	ccleaner
4	DGA	bcmvhyiplypyn.com	xshellghos
5	DGA	jepafezejsz.com	xshellghos
6	DGA	lyjqrulavyral.com	xshellghos
7	DGA	xsjczwxjcd.com	xshellghos
8	DGA	fcj2lyhr.ru	dromedan
9	DGA	kow19fe7ak.ru	dromedan
10	DGA	9jdc01e.ru	dromedan
11	DGA	xg0aya3my.ru	dromedan
12	DGA	hyj6f2a8dtr.ru	dromedan
13	DGA	ekic4bfsc.ru	dromedan
14	DGA	pxrsvgai.ru	dromedan
15	DGA	4wcbwk02.ru	dromedan
16	DGA	www.7f3gmwnnkn.net	madmax
17	DGA	www.jdn1p4p4b3.net	madmax
18	DGA	www.yz1sg9jntf.org	madmax
19	DGA	www.pafsc00v94.info	madmax
20	DGA	vaxswshr6q.info	madmax

[그림5-24] 데이터셋 통합

정상 도메인은 The Majestic Million에서 백만개를 수집하였으며 DGA 도메인은 netlab.360과 osint.bambenekconsulting에서 수집했다. DGA 도메인을 수집한 두 곳의 데이터 형식이 서로 달라 필요없는 필드와 중복 데이터들을 삭제했으며 DGA 도메인마다 DGA 알고리즘별로 라벨링하여 데이터셋을 통합시켰다.

TLD_index	3-gram_Score	4-gram_Score	5-gram_Score	Length	Numeric_ratio	Vowel_ratio	...	Consecutive_consonant	Consecutive_Vowel	period	Entropy	Max_Consecutive_Consonant	Max_vowel_Consonant	Meaning_count
0	-0.469568	-0.87709	-0.292795	-0.355201	-0.540403	-0.414497	...	0.923754	0.933805	-0.18609	0.333528	-0.145129	1.228286	-0.340711

[1 rows x 15 columns]

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	TLD_index	3-gram_Score	4-gram_Score	5-gram_Score	Length	Numeric_r	Vowel_rati	Consonan	Consecuti	Consecuti	period	Entropy	Max_Cons	Max_voew	Meaning_count	
2	0	-0.46057	-0.07709	-0.29279	-0.3552	-0.5404	-0.4145	-0.26395	0.659976	0.923754	0.933805	-0.18609	0.333528	-0.14513	1.228286	-0.34071

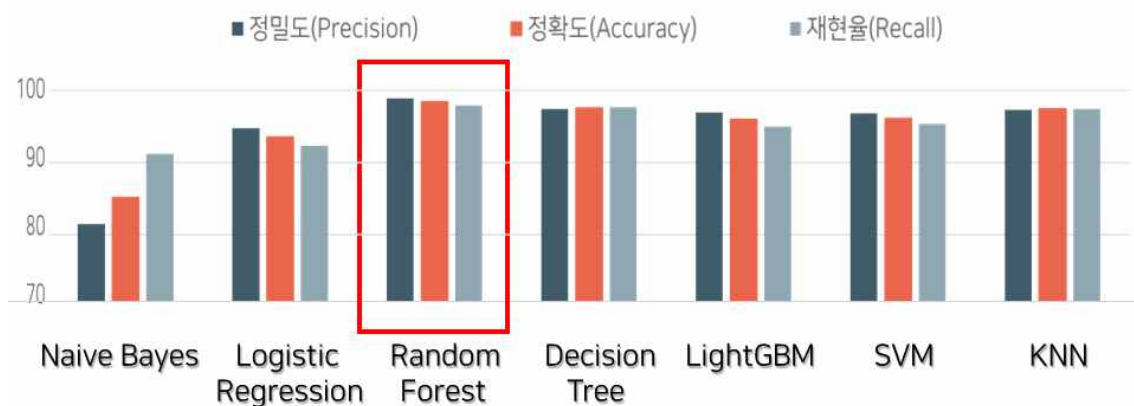
[그림5-25] 피쳐 스케일링

피쳐 값을 추출한 원 상태는 피쳐마다 값의 범위가 큰데 바로 모델에 판단을 요청하면 판단의 정확도가 낮아져 오탐, 미탐이 발생할 수 있다. StandardScaler를 사용해 피쳐값들을 평균 0, 표준편차가 1이 되도록 스케일링한다.

	Naive Bayes	Logistic Regression	Random Forest	Decision Tree	Light GBM	SVM	KNN
정밀도 (Precision)	81.40%	94.68%	98.80%	97.44%	96.91%	96.73%	97.32%
정확도 (Accuracy)	85.19%	93.56%	98.35%	97.51%	95.99%	96.06%	97.36%
재현율 (Recall)	91.22%	92.31%	97.89%	97.58%	95.02%	95.34%	97.40%

[표 5-2] 모델 성능 평가 결과

정상과 DGA 도메인 2,000,000개 중 75%는 학습 데이터, 25%는 검증용 데이터로 정하여 여러 모델 알고리즘에 대한 성능평가를 진행하였다.



[그림5-26] 모델 성능 평가 결과 그래프

```

테스트 세트의 정확도: 0.98
오차 행렬:
[[244723  5272]
 [ 2955 247050]]
정밀도(precision) = 0.9880691866051889
정답률(accuracy) = 0.983546
재현율(Recall) = 0.9789115782315646
precision    recall  f1-score   support

   DGA      0.9881    0.9789    0.9835     249995
  Normal      0.9791    0.9882    0.9836     250005

 accuracy          0.9835     500000
 macro avg      0.9836    0.9835    0.9835     500000
weighted avg      0.9836    0.9835    0.9835     500000

End

```

[그림5-27] Random Forest 모델 학습 결과

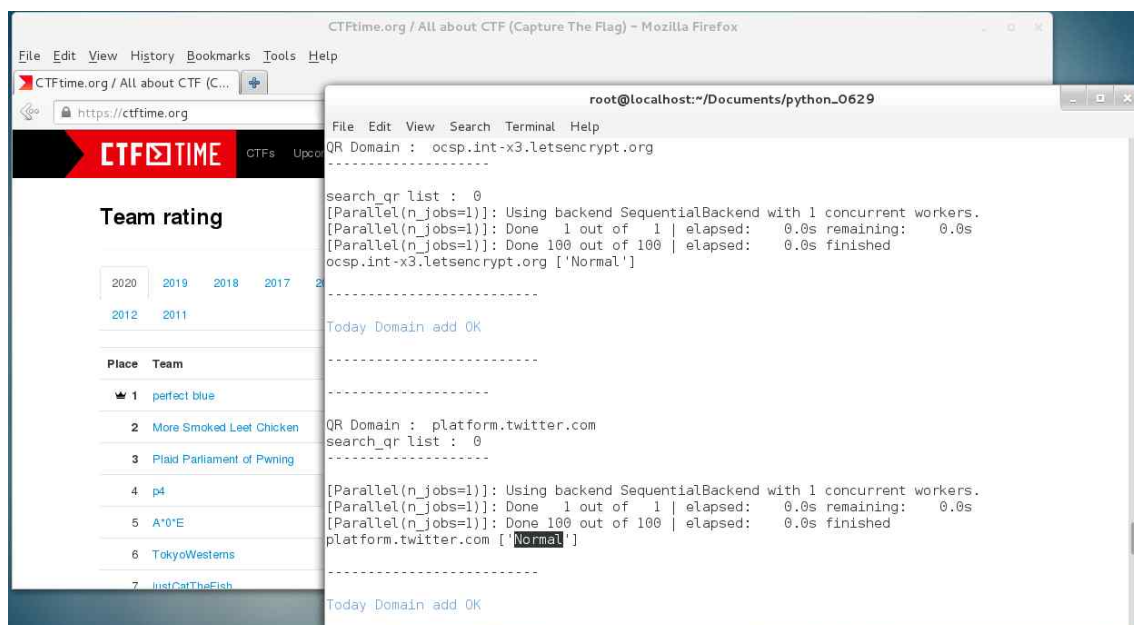
Random Forest 모델이 다른 모델보다 정확도, 정밀도, 재현율이 모두 우수하여 시스템에서 사용할 모델로 선정하였다.

```

creativefan.com ['Normal']
lonerwolf.com
lonerwolf.com ['DGA']
coolblue.be
coolblue.be ['Normal']
superlib.net
superlib.net ['Normal']
bkaj.net
bkaj.net ['DGA']
dpunkt.de
dpunkt.de ['DGA']
easyvoyage.com
easyvoyage.com ['Normal']
malcare.com
malcare.com ['Normal']
dyndns.dk
dyndns.dk ['DGA']
docshop.com
docshop.com ['Normal']
eduiso.org
eduiso.org ['DGA']

```

[그림5-28] 도메인 실시간 판단 결과



[그림5-29] 도메인 실시간 판별 결과 2

전달받은 도메인을 DGA 도메인인지 정상 도메인인지 실시간으로 판별한다.

5.3 시스템 시험

5.3.1 시험 환경

가상 공격자 네트워크	가상 회사 네트워크
가상 네트워크 시뮬레이터 (EVE-NG)	
가상 머신 (VMware)	
데스크탑	

[그림5-30] 시스템 시험 환경 구조도

5.3.1.1 시험 환경 사양

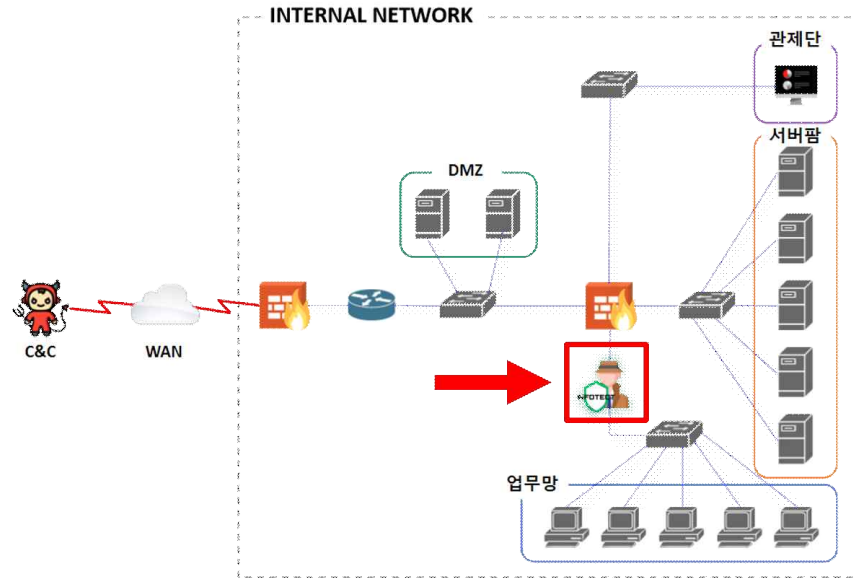
구분	사양
데스크탑	OS: Windows 10 Pro K Version 1909(Build 18363.836) CPU: AMD FX-8300(8-Core, 8-Thd, 3.3~4.2GHz, L3 8MB Cache) RAM: DDR3-16GB SSD: 120GB HDD: 500GB (2EA)

[표 5-3] 하드웨어 사양

구분	사양
가상머신	VM: VMware Workstation 15 Pro(15.5.2 build-15785246) OS: EVE-NG 4.20.17 CPU: 4 Processors (2 cores each) RAM: 8GB HDD: 300GB

[표 5-4] 가상머신 사양

5.3.2 가상 네트워크 구성



[그림5-31] 시스템 시험 환경 구조도

구분	종류	사양
공격자 명령제어 서버	데스크탑	VM: QEMU 2.4.0 OS: Kali-linux-large-2019.3 CPU: 2 Processor(QEMU Virtual CPU v2.5+) RAM: 2GB

[표 5-4] 공격자 데스크탑 사양

구분	세부구분	구성
가상 회사 네트워크	DMZ	WEB서버: VM: QEMU 2.4.0 OS: Cent OS 7.4.1708 CPU: 1 Processor(QEMU Virtual CPU v2.5+) RAM: 1GB
		DNS서버: VM: QEMU 2.4.0 OS: Windows Server 2008 lite-u1 CPU: 1 Processor(QEMU Virtual CPU v2.5+) RAM: 2GB
	업무망	업무용 데스크탑(5EA): VM: QEMU 2.4.0 OS: Windows 7 Ultimate CPU: 2 Processor(QEMU Virtual CPU v2.5+) RAM: 4GB
	서버팜	서버(5EA): VM: QEMU 2.4.0 OS: Cent OS 7.4.1708 CPU: 1 Processor(QEMU Virtual CPU v2.5+) RAM: 2GB
	관제망	관제용 데스크탑: VM: QEMU 2.4.0 OS: Windows 7 Ultimate CPU: 2 Processor(QEMU Virtual CPU v2.5+) RAM: 4GB
	네트워크 장비	L2스위치: VM: QEMU 2.4.0 IOS: Cisco-I2-iol-image NVRAM: 1GB RAM: 1GB
		라우터: VM: QEMU 2.4.0 IOS: Cisco-I3-iol-image NVRAM: 1GB RAM: 1GB
		방화벽: VM: QEMU 2.4.0 OS: Sophosutm-UTM-9.510-5.1 CPU: 1 Processor(QEMU Virtual CPU v2.5+) RAM: 2GB
	시험용 서버	VM: QEMU 2.4.0 OS: Cent OS 7.4.1708 CPU: 2 Processor(QEMU Virtual CPU v2.5+) RAM: 8GB

[표 5-5] 가상 회사 네트워크 구성

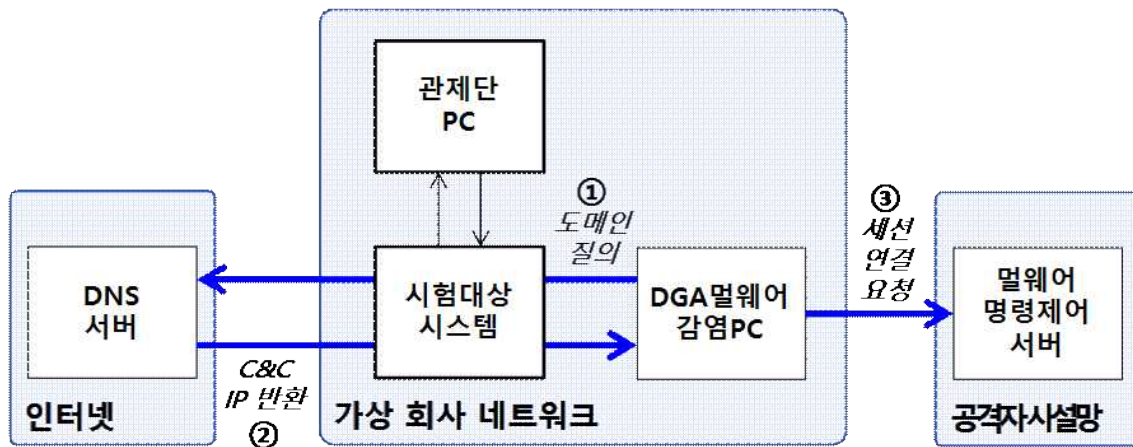
5.3.3 기능시험

5.3.3.1 시험 요구사항 정의

구분	기능	요구사항
DNS 패킷 캡처링 모듈	DNS 패킷 캡처링	1. 특정 네트워크 구간의 DNS 패킷 캡처 2. DNS 패킷에서 질의 도메인, 출발지IP, 응답IP 추출
	DB조회	1. 화이트리스트 조회 2. 탐지된 도메인 리스트 조회 3. 블랙리스트 조회 4. 금일 탐지 리스트 조회
	다중 처리	멀티 프로세싱을 구현하여 캡처링 및 추출을 동시에 처리
판별 결과 전달 모듈	데이터 전달	1. 금일 탐지 리스트에 데이터 추가 2. 탐지된 도메인 리스트에 데이터 추가
	알림	관리자 메일로 탐지 사실 전달
	표준 출력 메시지	1. 금일 탐지 리스트에 신규 데이터로 추가되는 경우, 파란색 중복 데이터 존재할 경우 보라색으로 라벨링 2. 탐지된 도메인리스트에 신규 데이터로 추가되는 경우 또는 중복 데이터 존재할 경우 빨간색으로 라벨링
데이터베이스	데이터 저장	1. 탐지된 도메인 리스트 저장 2. 화이트리스트 저장 3. 블랙리스트 저장 4. 금일 탐지 리스트 저장 5. 월간 탐지 리스트 저장 6. 관리자 계정 정보 관리 저장
	데이터 삭제	매일 자정이 되면 금일 탐지 리스트 초기화
관리자 대시보드	관리자 계정 관리	1. 로그인 계정정보 검증 2. 계정 조회 3. 신규 계정 등록 4. 기존 계정 삭제
	도메인 리스트 관리	1. 탐지된 도메인 리스트 조회/삭제 2. 블랙리스트 조회 및 도메인 추가/삭제 3. 화이트리스트 조회 및 도메인 추가/삭제
	탐지 현황 대시보드	2. 금일 탐지율을 원형그래프 형태로 표기 2. 최근 AI판별을 거친 10개의 데이터 중 50% 이상 DGA로 탐지되었을 경우 대시보드의 원형그래프 배경색을 붉게 변경 3. 월간 탐지 현황 표기
	탐지 현황 보고서	일간 보고서 제공
DGA 도메인 판별 모듈 (AI)	데이터셋 관리	1. 중복데이터 삭제 2. 학습에 필요한 데이터셋을 하나의 파일로 통합
	머신러닝 모델	1. 모델 학습 2. 모델 저장/불러오기
	피쳐 추출	1. 15개의 피쳐값 추출 2. 피쳐 스케일링 과정을 통해 정규화 3. 피쳐값 저장 4. 피쳐값 시각화 그래프 출력
	판별 모듈(AI모델)	실시간 판별

[표 5-6] 시험 요구사항 정의

5.3.3.2 시험 방법



[그림5-32] 시험 케이스 구조도

기능 시험 목적
시험대상 시스템이 개발자가 의도한 기능을 모두 정상적으로 수행하는지 검증하기 위함
시험 사전 준비
1. 가상 회사 네트워크 내에 시험대상 시스템 배치 및 시스템 구동 2. 공격자 사설망에 C&C서버를 구축하고 희생자PC가 리버스 ssh 세션 연결 요청할 때까지 해당 포트 listen 상태로 대기 3. 가상 회사 네트워크 내에 배치된 PC중 하나에 DGA 멀웨어 주입
기능 시험 절차
1. 시험대상 시스템 정상 작동 여부 확인 2. 관제단PC에서 시험대상 시스템의 웹서버(관리자 대시보드)로 접속가능 여부 확인 3. 희생자PC에서 여러 정상사이트 접속 4. 희생자PC에서 DGA 멀웨어 실행 5. 공격자C&C서버에서 희생자PC와의 세션 수립 여부 확인 6. 관제단PC에서 대시보드 확인 7. 데이터베이스에 저장된 데이터 확인 6. 시험대상 시스템의 시스템 시간을 변경하여 자정이 되었을 때, 금일 탐지 리스트 초기화 여부 확인

[표 5-7] 시험 방법

5.3.3.3 시험 결과

기능	시험 항목	정상 작동 여부
DNS 패킷 캡처링	1. 특정 네트워크 구간의 DNS 패킷 캡처	YES
	2. DNS 패킷에서 질의 도메인, 출발지IP, 응답IP 추출	YES
DB조회	1. 화이트리스트 조회	YES
	2. 탐지된 도메인 리스트 조회	YES
	3. 블랙리스트 조회	YES
	4. 금일 탐지 리스트 조회	YES
다중 처리	멀티 프로세싱을 구현하여 캡처링 및 추출을 동시에 처리	YES
데이터 전달	1. 금일 탐지 리스트에 데이터 추가	YES
	2. 탐지된 도메인 리스트에 데이터 추가	YES
알림	관리자 메일로 탐지 사실 전달	YES
표준 출력 메시지	1. 금일 탐지 리스트에 신규 데이터로 추가되는 경우 파란색 중복 데이터 존재할 경우 보라색으로 라벨링	YES
	2. 탐지된 도메인 리스트에 신규 데이터로 추가되는 경우 또는 중복 데이터 존재할 경우 빨간색으로 라벨링	YES
데이터 저장	1. 탐지된 도메인 리스트 저장	YES
	2. 화이트리스트 저장	YES
	3. 블랙리스트 저장	YES
	4. 금일 탐지 리스트 저장	YES
	5. 월간 탐지 리스트 저장	YES
	6. 관리자 계정 정보 관리 저장	YES
데이터 삭제	1. 매일 자정이 되면 금일 탐지 리스트 초기화	YES
관리자 계정 관리	1. 로그인 계정정보 검증	YES
	2. 계정 조회	YES
	3. 신규 계정 등록	YES
	4. 기존 계정 삭제	YES
도메인 목록 관리	1. 탐지된 도메인 리스트 조회/삭제	YES
	2. 블랙리스트 조회 및 도메인 추가/삭제	YES
	3. 화이트리스트 조회 및 도메인 추가/삭제	YES
탐지 현황 대시보드	1. 금일 탐지율을 원형그래프 형태로 표기	YES
	2. 최근 시판별을 거친 10개의 데이터 중 50%이상 DGA로 탐지되었을 경우 대시보드의 원형그래프 배경색을 붉게 변경	YES
	3. 월간 탐지 현황 표기	YES
	4. 일간 보고서 제공	YES
데이터셋 관리	1. 중복데이터 삭제	YES
	2. 학습에 필요한 데이터셋을 하나의 파일로 통합	YES
머신러닝 모델	1. 모델 학습	YES
	2. 모델 저장/불러오기	YES
피쳐 추출	1. 15개의 피쳐값 추출	YES
	2. 피쳐 스케일링 과정을 통해 정규화	YES
	3. 피쳐값 저장	YES
	4. 피쳐값 시각화 그래프 출력	YES
판별 모듈(AI모델)	실시간 판별	YES

[표 5-8] 시험 결과

06

결론

6.1 향후계획

06 결론

| 6.1 향후계획

현재 제작된 AI 모델은 DGA 도메인이 가지는 언어적 특징을 토대로 판별하는 모델이다. 하지만 DGA 멀웨어는 언어적인 특징 외에도 시간적, 수량적인 DNS 트래픽 패턴을 가진다. 따라서 향후에는 DGA 멀웨어가 가지는 시간적, 수량적인 DNS 트래픽 패턴을 분석하고 이를 탐지하는 기능을 추가하여 고도화할 계획이다. 기능 뿐만 아니라 UI적인 면에서도 ELK스택을 이용하여 더 다양한 정보를 제공하는 대시보드를 제작할 계획이다. 더하여 관리자가 블랙리스트 또는 화이트리스트에 등록한 도메인을 방화벽에 정책으로 자동 등록시키는 기능을 구현할 계획이다.

