**제목: DGA를 적용한 악성 도메인 분류기 및 알고리즘 성능비교**

**㈜넥스투비 연구소**

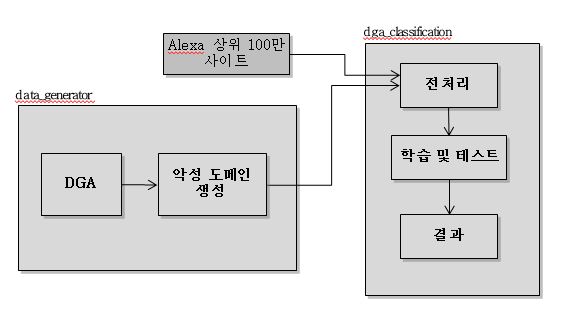
**㈜넥스투비 정윤수 차장**

**요약**

알려져 있는 DGA를 적용해(open source) 생성한 악성 도메인과 정상 도메인을 학습 데이터로 사용해 학습(Supervised learning) 및 분류(Classification) 하였습니다.

딥러닝 알고리즘(LSTM)과 머신러닝 알고리즘(Bigram) 적용하여 도메인 분류기 성능을 비교했으며

LSTM 결과가 우수함을 확인 하였습니다.



**1. 소개**

오늘날 사이버 범죄의 근원인 악성코드는 광범위한 연구 노력에도 불구하고 여전히 심각하며 감염은 계속 증가하고 있습니다.

또한 해커들이 악성코드에 적용한 DGA는 더욱 진화하면서 악성 도메인을 생성하고 있으며,

그것을 극복하기 위한 대안도 지속하는 발전하는 것이 매우 중요합니다.

따라서 악성 코드로부터 감염된 좀비PC 와 제어서버를 효과적으로 차단하는 선제대응 기술을 필요로 합니다.

본 실험에서는 악성코드 문제를 해결하기 위해 알려져 있는 DGA를 적용해 패턴을 파악하고

그 패턴에 의해 생성된 악성 도메인을 머신러닝 및 딥러닝 기술로 학습하고 제어서버 도메인을 분류합니다.

이는 미래의 악성 도메인 사전 예측을 의미하며, 향후 악성 도메인 자동 분류기로 발전할 수 있습니다.

또한 딥러닝 알고리즘과 머신러닝 알고리즘 성능 비교 합니다.

**2. 관련 연구**

- DGA

- 정의(제어서버와 관계)

악성 코드에 감염된 좀비PC 가 제어서버(C&C 서버)에 접속할 때 DNS를 요청하게 되는데 해커는 악성코드에 DGA를 적용 합니다.  
(DNS는 제어 서버의 실제 주소와 웹사이트이 별칭(domain)을 연결시켜주는 서비스)

DGA는 정해진 패턴에 의해 대량의 도메인을 생성하고 이를 제어서버에 접속하기 위한 도메인으로 사용하며, 이때 제어서버의 주소를 계속 바꾸게 되어 DNS 행위 분석을 통해 악용 당하고 있는 도메인에 대한 접근을 차단하거나 끊어버려도 다른 도메인을 사용하게 되어 제어서버의 차단을 피하게 됩니다.

예를 들어 제어서버의 도메인을 어제는 cnc1.com 로 사용하고, 오늘은 cnc2.com를 사용하고,  
내일은 cnc3.com을 사용해(주기적인 변경) 오늘 제어서버 접속이 차단 되더라도 내일은 접속  
이 가능한 식이기 때문에 항구적인 접속 차단이 어렵게 되는 것입니다.

- Bigram

- 정의 (unigram / bigram / n-gram)

유니그램 모형은 문장 내 하나하나의 단어는 어떤 확률 분포에서 독립적으로 추출되었다고 가정 합니다.

쉽게 말해 한 면마다 단어가 쓰인 거대한 주사위를 던져서, 온 단어들로 문장이 이뤄졌다고 보는 것 입니다. 단어마다 문장 내 출현 빈도를 알고 있으므로, 단어들을 이 확률에 따라 랜덤하게 생성할 수 있습니다.

유니그램 모형의 확장 모형으로 바이그램(bigram) 모형이 있습니다.

'바이(bi-)'는 '둘'이라는 뜻 입니다. 한 단어가 나타날 확률이 앞 단어에 영향을 받는다고 가정하는 것입니다.

같은 방식으로 바이그램 모형을 확장해서 트라이그램(trigram) 모형을 만들 수도 있습니다.

'트라이(tri-)'는 3이라는 뜻이므로 이번엔 한 단어가 나타날 확률이 바로 앞단어만이 아니라 그 앞단어에도 영향을 받습니다.

유니그램, 바이그램, 트라이그램 같은 모형을 모두 합쳐서 N그램이라고 부르는데 N을 늘리면 늘릴수록 점점 더 말같은 소릴 하는 모형을 만들 수 있지만 N그램만으로는 완전한 문장을 쓸 수가 없습니다.

N그램 모형은 확률로만 처리하고 의미적인 부분을 처리하는 부분이 전혀 없기 때문에 아무리 N이 늘어나도 구문론적인 관계를 전혀 포착하지 못하기 때문이다.

예를 들어 "눈이 아파"라는 문장이 있다면 이 '눈'은 펄펄 내리는 눈(雪)이 아니라

사람의 몸에 있는 눈(目)일 확률이 높다는 것 정도는 N그램 모형으로 식별 할 수 있습니다.

- LSTM

- 정의 (RNN 과 LSTM)

RNN은 히든 노드가 방향을 가진 엣지로 연결돼 순환구조를 이루는(directed cycle) 인공신경망의 한 종류입니다.

음성, 문자 등 순차적으로 등장하는 데이터 처리에 적합한 모델로 알려져 있으며, 시계열 데이터 형태를 갖는 데이터의 패턴을 인식하는 인공신경망 입니다.

RNN은 지금 들어온 입력 데이터와 과거에 입력 받았던 데이터를 동시에 고려하며

마치 머릿속에 기억을 저장하고 있듯이 은닉층에 기억을 저장합니다.

사람은 생각하고 판단하는 과정에서 과거의 기억에 의존하는데, RNN이 하는 일도 이와 비슷합니다.

RNN의 변형인 LSTM은 오차의 그라디언트가 시간을 거슬러서 잘 흘러갈 수 있도록 도와줍니다 backprop하는 과정에서 오차의 값이 더 잘 유지되는데, 결과적으로 1000단계가 넘게 거슬러 올라갈 수 있습니다.

이렇게 그라디언트가 잘 흘러간다는 것은 다시 말해 RNNs가 더 오래 전 일도 잘 기억한다는 의미입니다.

LSTM 유닛은 여러 개의 게이트(gate)가 붙어있는 셀(cell)로 이루어져있으며 이 셀의 정보를

새로 저장/셀의 정보를 불러오기/셀의 정보를 유지하는 기능이 있습니다(컴퓨터의 메모리 셀과 비슷합니다).

셀은 셀에 연결된 게이트의 값을 보고 무엇을 저장할지, 언제 정보를 내보낼지, 언제 쓰고 언제 지울지를 결정합니다.

이 게이트가 열리거나(1) 닫히는(0) 디지털이 아니라 아날로그라는 점 주의하셔야 합니다. 즉, 각 게이트는 0에서 1사이의 값을 가지며 게이트의 값에 비례해서 여러 가지 작동을 합니다.

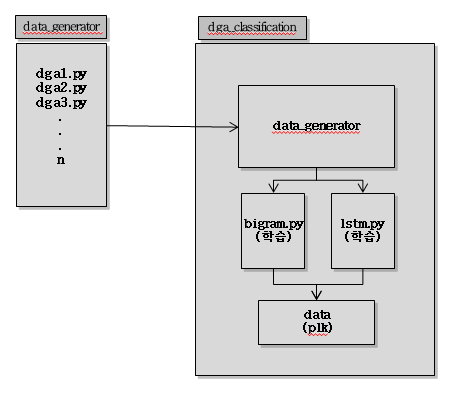
각 게이트가 갖는 값, 즉 게이트의 계수(또는 가중치, weight)는 은닉층의 값과 같은 원리로 학습됩니다.

즉 게이트는 언제 신호를 불러올지/내보낼지/유지할지를 학습하며 이 학습과정은 출력의 오차를 이용한 경사 하강법(gradient descent)을 사용합니다.

**3. 실험**

- 실험 모형

악성도메인 분류기의 처리 모형은 크게 2개의 모듈로 구성되어 있으며 다음 그림과 같습니다.



data\_generator 모듈은 정상적인 도메인 데이터를 수집하고 기존에 알려진 DGA를 이용해

학습용 악성 도메인 정보를 생성하며, 실험에서는 10개의 알려진 DGA를 사용 하였습니다.

DGA를 10개 사용했으므로 실제 100만개 데이터가 생성됩니다.

dga\_classification 모듈은 data\_generator 로부터 생성된 학습용 데이터와 Alexa 상위 100만건

데이터를 10만건 랜덤 추출해 최종 학습용 데이터로 생성합니다.

이 과정에서 각각 정상 데이터와 비정상 데이터를 Label 합니다.(White / Black list)

전처리 및 Labeling 과정을 거친 최종 데이터는 Supervised learning 으로 분류하며 성능

비교를 위해 LSTM, Bigram 알고리즘을 각각 적용하여 실행 하였습니다.

실행을 마치면 각각의 결과 ROC를 시각화 하여 표현하며, 분류를 마친 데이터는 파이썬

모듈 내 피클(pickle) 타입으로 떨어지게 됩니다.

- data의 이해

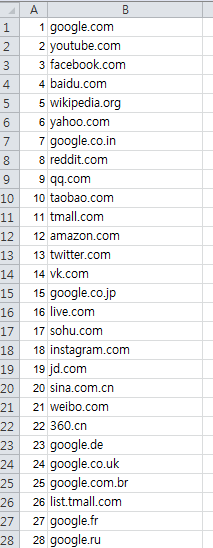
선별된 데이터에 대한 결과를 추출하기 위해서는 주제 또는 업무에 대한 높은 이해도를

필요로 하며 머신러닝에 적용하는 단계 이전에 수집된 데이터를 일차적으로 전처리 및 가공

하는 것이 선행되어야 합니다.

DGA를 통해 생성한 악성도메인과 정상도메인을 Label 하는 작업을 하여 딥러닝 학습을

위해 벡터화(Vectoring) 후 모델에 적용합니다.



정상(White) 데이터를 위해 Alexa 상위 100 만 사이트에서 10만건 랜덤 추출하여 학습 데이터

로 사용 하였습니다.



알려진 10가지 DGA를 사용해 악성 도메인 100만개를 생성하여 학습 데이터로 사용

하였습니다.

전처리 작업과 Label 작업을 마친 총 110만건의 데이터를 8:2 비율로 train set, test set

으로 구분 합니다.

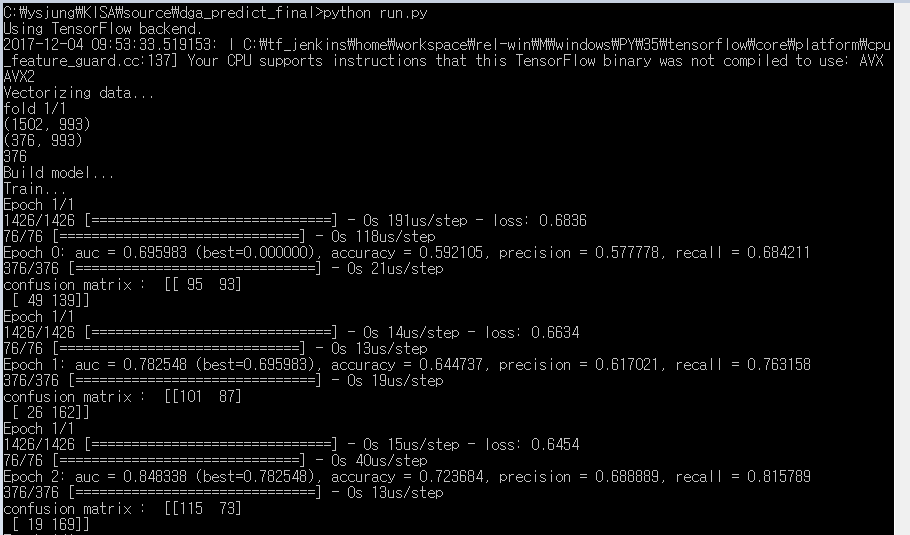
**4. 결과 및 분석**

- 실행 및 결과

파이썬으로 구현 되어진 코드는 위 모형에서 표현한 것처럼 기능별로 모듈화 되어 있으며

실행하게 되면 학습 과정 및 결과 관련된 지표 값을 찍을 수 있게 구현 되었습니다.

또한 실행이 완료되면 Bigram 과 LSTM 실행 결과의 ROC curve를 시각화 하여 표현합니다.



- Confusion matrix

학습 및 테스트 과정에서 Confusion matrix를 표현 합니다.

Confusion matrix 값은 참(Positive)인데 참, 참인데 거짓(Negative), 거짓인데 거짓, 거짓인데 참 이 4가지 경우 수로 예측한 결과가 몇 건인지 표현합니다.

- 정확도(Accuracy)

전체 샘플 중 맞게 예측한 샘플 수의 비율을 의미합니다.

- 정밀도(Precision)

학습 및 테스트 과정에서 전체 샘플 중 맞게 예측한 샘플 수의 비율 입니다.

참이라고 판단한 값 중에서 실제 참인 값의 비율을 의미 합니다.

- 재현율(Recall)

TPR: true positive rate

실제 클래스에 속한 샘플 중에 클래스에 속한다고 출력한 샘플의 수 입니다.

실제 참인 값 중에서 참으로 예측한 값의 비율이며, sensitivity(민감도)

라 표현 하기도 합니다.

- 위양성율(Fall-Out)

FPR: false positive rate

실제 클래스에 속하지 않는 샘플 중에 클래스에 속한다고 출력한 샘플의 수 입니다.

실제 거짓인 값 중에서 참으로 예측한 값의 비율이며, specificity(특이도)라 표현 하기도

합니다.

- AUC (Area Under the Curve)

AUC는 ROC curve의 면적을 의미합니다.

Fall-Out 대비 Recall 값이 클 수록 AUC가 1에 가까운 값이며 민감한 모형 입니다..

- ROC(Receiver Operator Characteristic) Curve

ROC(Receiver Operator Characteristic) 커브는 클래스 판별 기준값의 변화에 따른

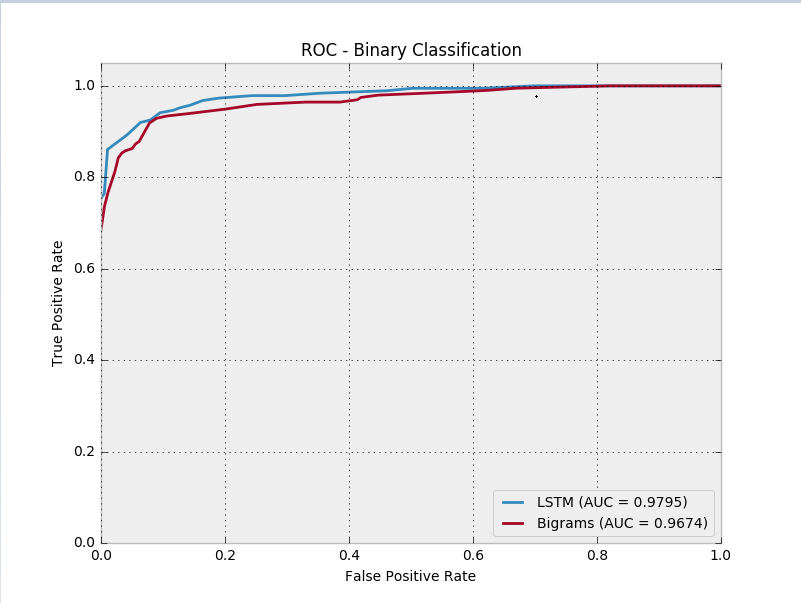
Fall-out과 Recall의 변화를 시각화한 것입니다.

일반적으로 클래스 판별 기준이 변화함에 따라 Recall과 Fall-out은 같이 증가하거나

감소 합니다.

일반적으로 Fall-out보다 Recall이 더 빠르게 증가하는 모형은 좋은 모형으로

판단할 수 있습니다.



ROC의 X축은 특이도(false positive rate) 이며, Y축은 민감도(true positive rate) 입니다.

LSTM의 AUC는 0.9795 이며 Bigram의 AUC 는 0.9674 로 LSTM 의 AUC 값이 Bigram 의 AUC 값 보다 높게 나타내므로 LSTM 이 Bigram 보다 좋은 모델이라 할 수 있습니다.

- 향후 연구를 제안

향후 DGA 분석(Reverse engineering) 모델을 고도화하고, 더 다양한 분류 Algorithm 사용으로

향후 악성 도메인 분류기 성능 향상을 기대 할 수 있습니다

**5.결론**

본 실험에서 DGA 패턴 분석 및 학습해 악성 도메인을 분류 하였습니다.

이는 악성 도메인 자동 분류기 형태이며, 기존 악성 도메인 차단 한계 극복 가능성 확인을

의미 합니다.

또한 성능 평가를 위해 딥러닝 알고리즘(LSTM)과 머신러닌 알고리즘(Bigram) 모델을 각각

실행하여 성능 평가 했으며 LSTM 모델의 성능이 우수한 것을 확인하였습니다.

향후 DGA 분석(Reverse engineering) 모델을 고도화하고, 더 다양한 분류 Algorithm 사용으로

향후 악성 도메인 분류기 성능 향상을 기대 할 수 있습니다.