기계학습 및 데이터 마이닝

Term Project Proposal

그거 맞아? (피싱 사이트 탐지)

팀 명: 트릴리언

안윤회 201620899

김두원 201620630

유상정 201620641

김상우 201620631

이찬호 201620648

1. 배경

최근 피싱 사이트에 대한 피해사례가 굉장히 늘고 있다. 실제로 과학기술정보통신부에서 실시한 조사결과에 따르면 2016년부터 2019년 8월까지의 4년간 피싱 사이트 신고·차단 건수가 31,000건이 넘어선 것으로 확인됐다. 또한 작년 하반기에는 국내 포털사이트 점유율 2위인 네이버의 로그인 창을 사칭하여 조작된 피싱 사이트로 인하여 개인정보가 유출되는 등 많은 피해사례가 나타났다. 2017년 기준 피싱, 파밍 피해액은 이전 3년간 5405억원으로 절대 무시할 수 없는 수준의 피해가 발생하고 있음을 확인할 수 있다.

2. 개요

본 프로젝트는 '기계학습을 이용한 피싱 사이트의 탐지' 로 피싱 사이트의 특징에 대해 살펴보고 사이트 구분을 위한 feature 선정을 우선적으로 고려하였다. 피싱 사이트의 내부는 일부 정상적인 사이트를 모방하여 사용자를 속이는 것을 목적으로 하기 때문에 사이트 내부의 내용을 통한 분류는 힘들 것으로 판단된다. 따라서 가장 두드러지는 특징인 URL을 이용하여 피싱 사이트를 분류하고자 한다.

3. 목표

1) EDA를 통한 피싱 사이트 URL의 특징파악
2) 알고리즘별 평가를 통한 가장 적합한 알고리즘 판단
3) URL 기반의 피싱 사이트와 정상사이트 구별

4. Data 분석 및 평가

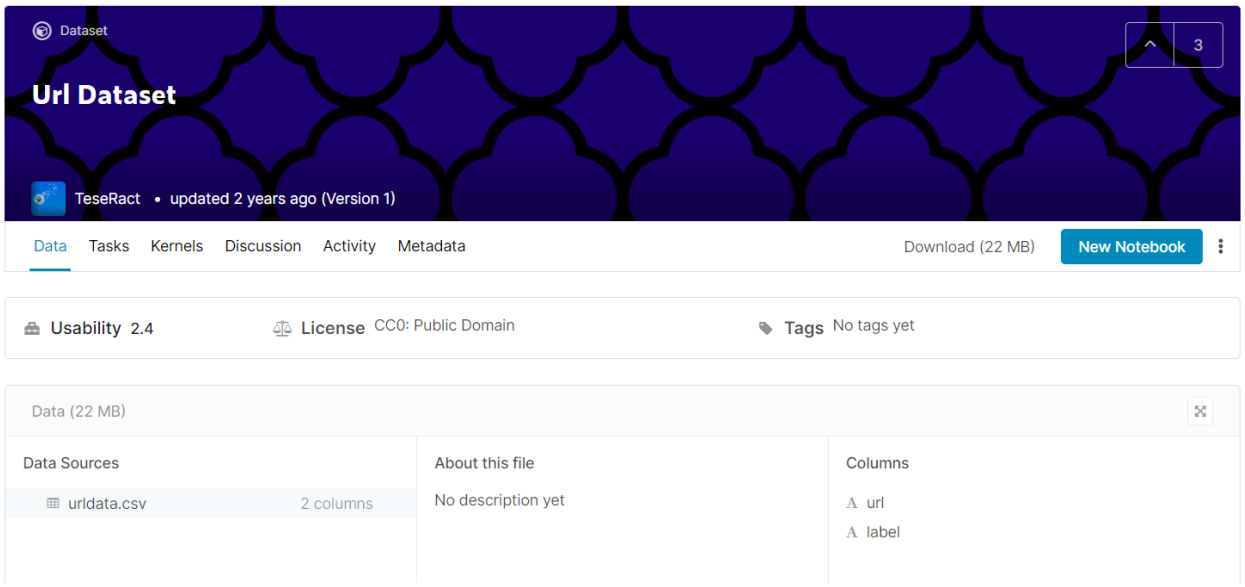
4.1. 분석할 Data set

Recent Submissions

You can help! Sign in or register (free! fast!) to verify these suspected phishes.

ID	URL	Submitted by
6545494	http://deerguardian.com/	verifrom
6545491	https://tera-no-mi.com/?Vic=info%40eurotexbg.eu	CSIRTUMINHO
6545479	http://redtulip.in/	verifrom
6545478	http://chriscullenmayor.com/	verifrom
6545477	http://happychampion.com/	verifrom
6545476	http://888showbis.com/	verifrom
6545475	https://winsecurity.fr/frontend/assets/files/custo...	verifrom
6545474	http://homesforsale-summitcounty.com/	verifrom
6545473	http://bkdphotos.com/	verifrom
6545472	http://fedds.net/	verifrom
6545471	http://miprofeonline.com/	verifrom
6545470	http://on0.org/	verifrom
6545469	http://altavia.me/	verifrom

[사진 1] 피싱 사이트 URL PhishTank



[사진 2] 정상 사이트 URL Url Dataset

4.2. 분석 및 평가 방법

분석 과정은 다음과 같다. 먼저, 정상사이트와 피싱 사이트 URL의 특징을 파악하고 각 사이트를 구별 지을 수 있는 feature를 선정한다. 이때, EDA를 통해 outlier data를 제거하고 feature간의 관계를 파악한다. 그 후, 전체 data set 중 train data를 90%로 test data를 10%로 나누고, 분류 알고리즘인 KNN, SVM, RandomForest, DecisionTree 와 회귀 알고리즘인 Logistic Regression 등을 GridSearchCV로 최적의 하이퍼파라미터를 찾고 학습하고, test set을 Accuracy, f1-score, Recall score, Precision score 등을 활용하여 성능 평가를 할 예정이다. 이러한 성능 평가 결과를 활용하여 피싱 사이트 탐지에 어떤 알고리즘이 적합한지 판단 후 가장 적합한 알고리즘을 사용해 문제를 해결할 예정이다.