# 4.3 프로세스 분석

## 4.3.2 기능 명세서

▷ 개발기능명세서

|  |  |  |  |
| --- | --- | --- | --- |
| Description | | | 비고  (신규/재개발) |
| Level1 | Level2 | Level3 |
| AI | 데이터셋 관리 | 중복데이터 삭제 |  |
| 데이터셋 통합 |  |
| 피처 추출 | 피처 전처리 |  |
| 피처 스케일링 |  |
| 피처값 저장 |  |
| 피처값 시각화 |  |
| 모델 관리 | 모델 학습 |  |
| 모델 저장/불러오기 |  |
| 모델 정확도 출력 |  |
| 실시간 판단 |  |

# 4.8 핵심 알고리즘(AI 피처)

## 4.8.1 TLD\_Index

1 if 도메인 TLD in TLD\_list :  
2 return TLD 순위  
3 else   
4 return 0

1. 도메인의 TLD가 TLD 순위 파일에 있으면
2. 해당 TLD 순위 반환
3. 도메인의 TLD가 TLD 순위 파일에 없으면
4. 0 반환

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

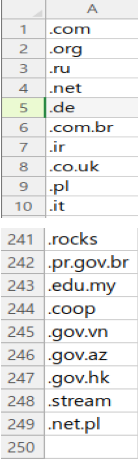


그림 1 TLD\_list

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

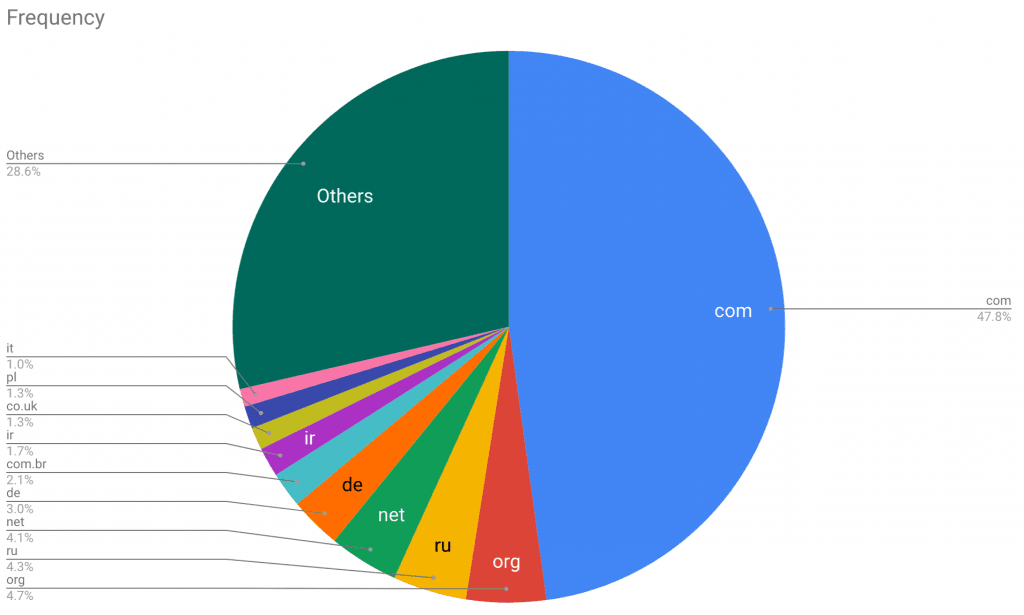


그림 2 Top 10 TLDs in the Alexa 1 Million

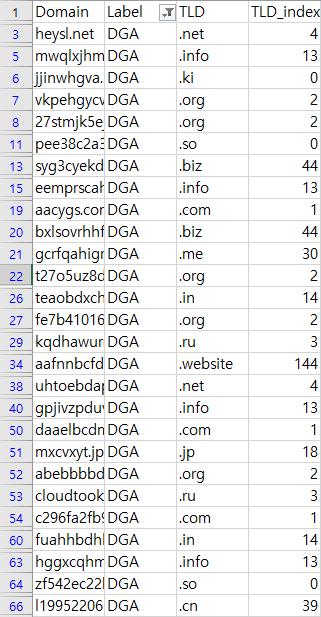
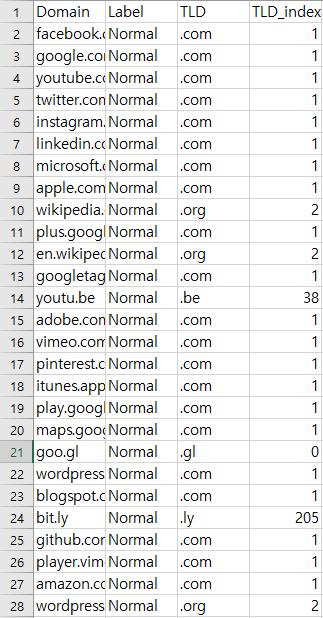


그림 3 정상/DGA 도메인 TLD

그림 3은 정상과 DGA 도메인의 TLD를 나타내는 표이다. 정상 도메인의 TLD는 자주 사용되는com, org가 많아 TLD\_index가 높지만 DGA 도메인은 com뿐만 아니라 여러 TLD도 사용하므로 TLD\_index의 분포가 큰 것을 확인할 수 있다.

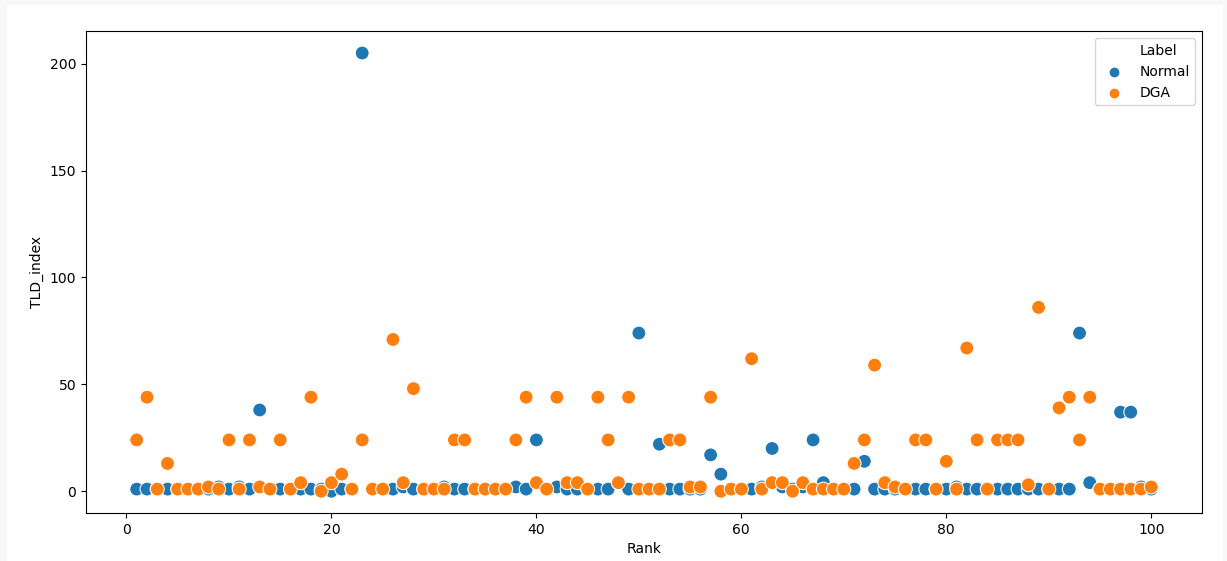


그림 4 정상/DGA 도메인 TLD 그래프

그림 4는 정상, DGA 도메인 TLD 분포를 나타낸 점 그래프인데, 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.

## 4.8.2 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 저장돼 있는 부분 문자열 빈도수를 더함
4. 전체 빈도수 합을 서브도메인 길이로 나눠 반환

N-gram score는 서브 도메인을 N-gram으로 자르고, 부분 문자열들의 빈도수를 각각 합해

서브 도메인의 길이로 나눈 값이다.



그림 5 N-gram Score 수식

부분 문자열들의 빈도수를 구하는 방법은 Majestic Top 1 Million에서 도메인의 서브도메인들을

N-gram으로 자르고, 모든 부분 문자열들의 빈도수 저장한 N-gram\_list 파일에서 불러온다.

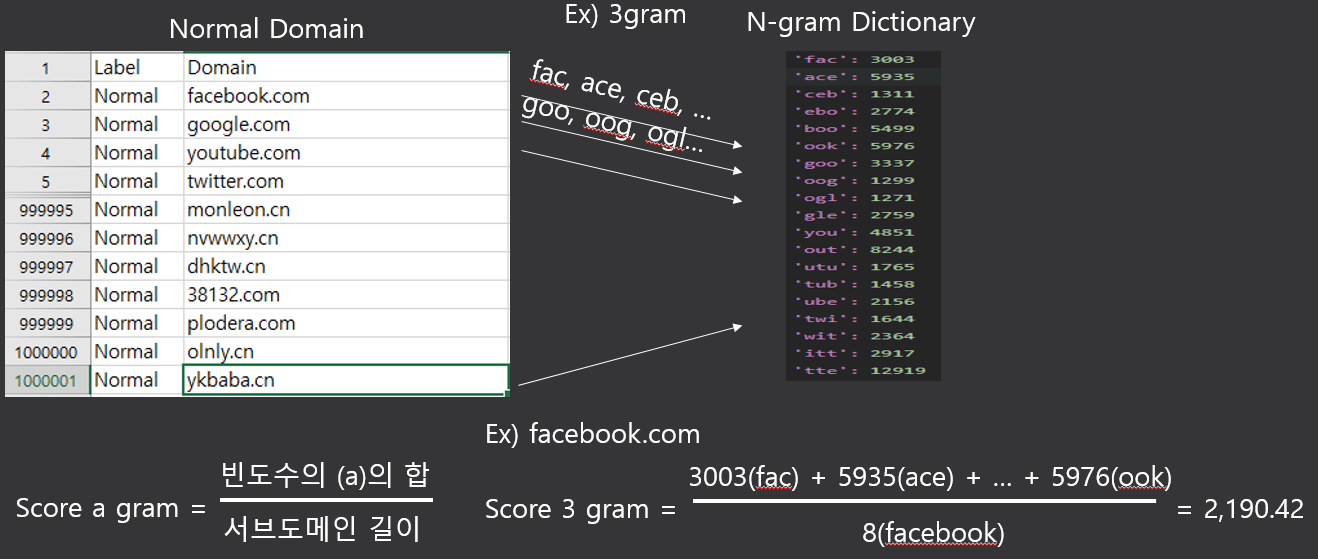


그림 6 3-gram\_list

그림 6은 정상 도메인에서 3-gram을 사용하여 3-gram\_list를 만드는 방법과 3-gram Score의 계산 방법을 예시로 보여주고 있다.

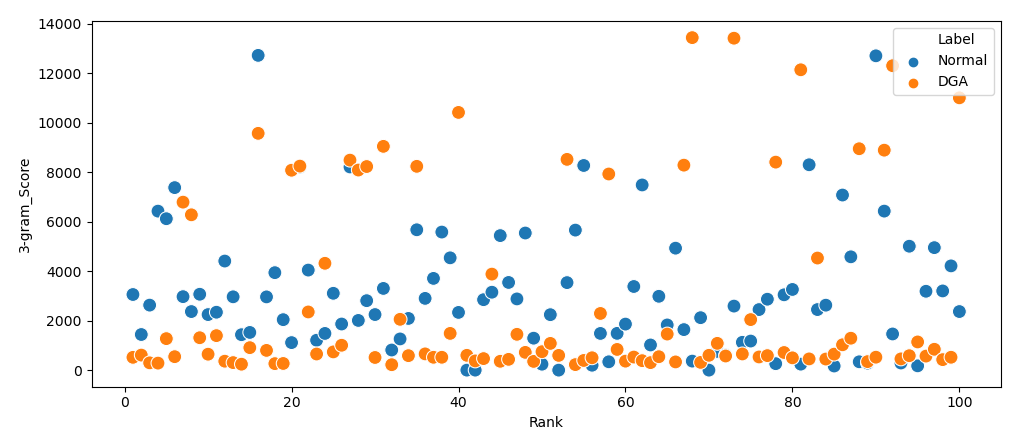


그림 7 정상/DGA 도메인 3-gram score

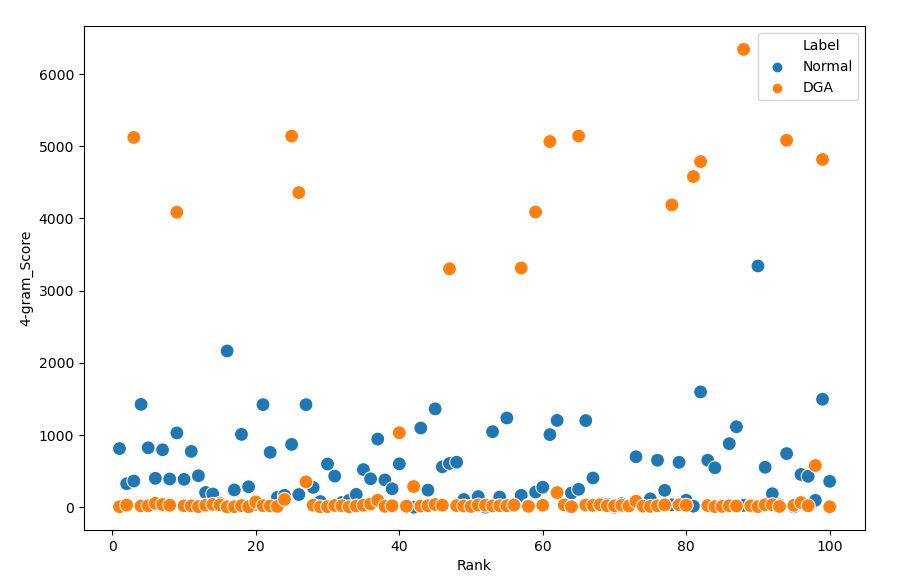


그림 8 정상/DGA 도메인 4-gram score

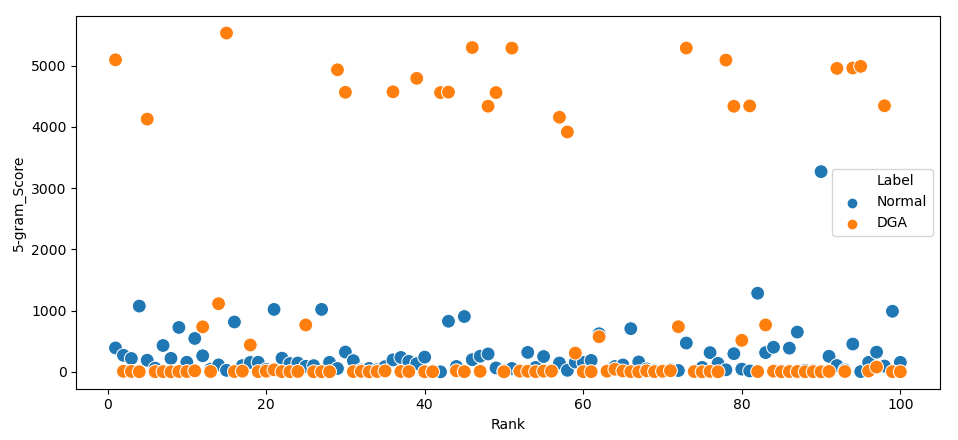


그림 9 정상/DGA 도메인 5-gram score

그림 7~9는 정상, DGA 도메인들의 3~5 gram score를 나타낸 점 그래프이다. 그래프를 보면 정상 도메인의 3~5 gram score가 DGA의 3~5 gram score보다 높다는 것을 확인할 수 있다.

## 4.8.3 Length

1 Length = len(도메인)  
2 return Length

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

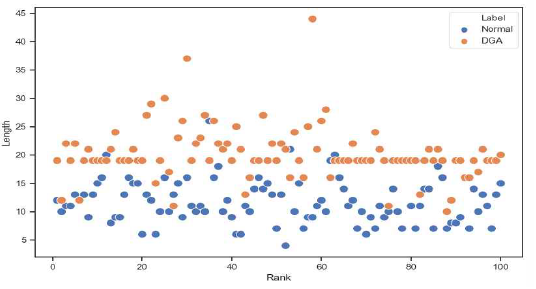


그림 10 정상/DGA 도메인 길이

## 4.8.4 Numeric\_ratio

1 Numeric\_ratio = count(서브 도메인에 포함된 숫자(0~9)) / len(서브 도메인))

Numeric\_ratio는 서브도메인의 숫자 비율로 서브 도메인에 포함된 숫자 개수들을 세고 서브 도메인의 길이로 나눈 피처이다. 정상 도메인의 경우 대부분 알파벳으로 이뤄져 있으나 몇몇 DGA 도메인은 알파벳과 숫자가 섞여있는 것이 특징이다.

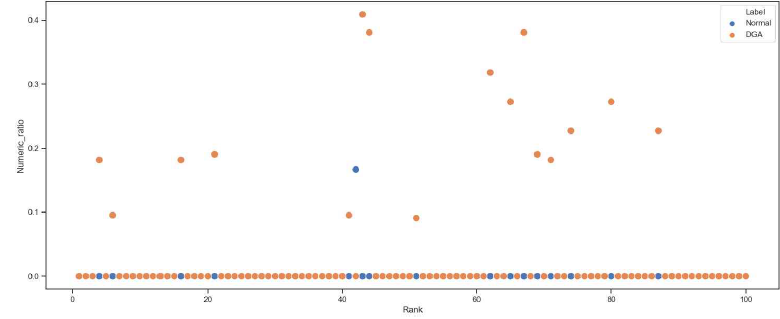


그림 11 정상/DGA 서브도메인 숫자 비율

## 4.8.5 Vowel\_ratio

1 Vowel\_ratio = count(서브 도메인에 포함된 모음(a,e,i,o,u)) / len(서브 도메인))

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

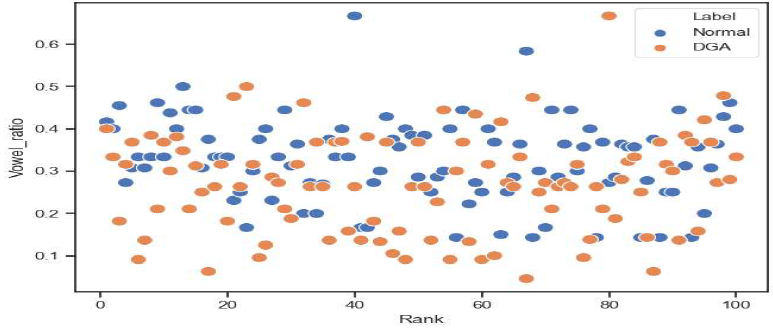


그림 12 정상/DGA 서브도메인 모음 비율

## 4.8.6 Consonant\_ratio

1. Consonant\_ratio = count(서브 도메인의 자음(^aeiou)) / len(서브 도메인))

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

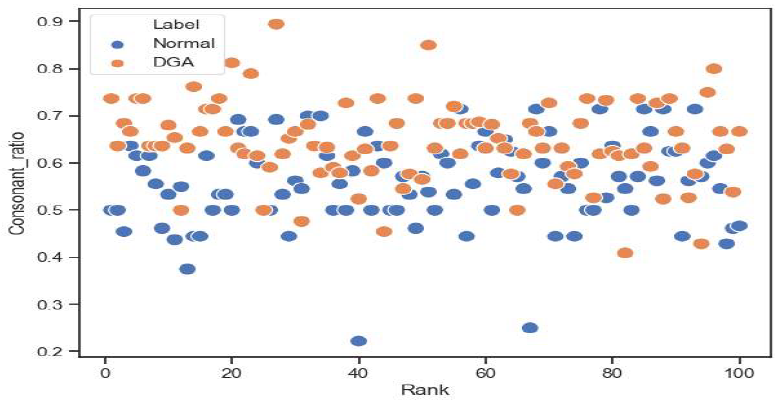


그림 13 정상/DGA 서브도메인 자음 비율

## 4.8.7 Consecutive\_consonant

1. Consecutive\_consonant = count(서브 도메인[^.aeiou]{3,})

Consecutive\_consonant는 서브도메인에서 3음절 이상 연속되는 자음, 숫자 문자열 개수를 센 피처이다. 알파벳에서 3음절 이상 연속되는 자음 문자열은 발음하기 어려워 정상 도메인에서는 잘 사용되지 않지만 DGA 도메인의 경우 무작위로 생성되고 발음할 이유가 없기 때문에 정상도메인보다 3음절 이상 연속되는 자음, 숫자 문자열 개수가 많다.

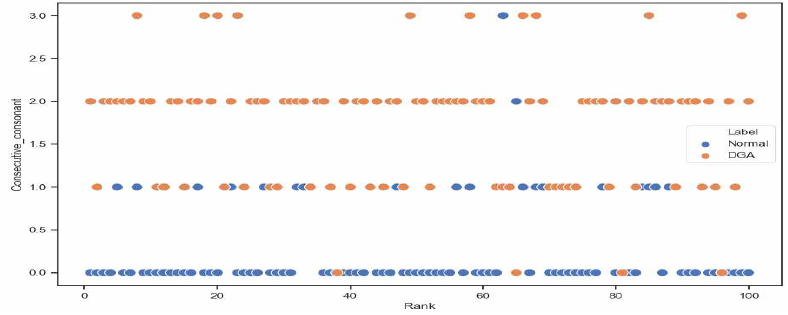


그림 14 정상/DGA 서브도메인 3음절 이상 연속되는 자음 문자열 개수

## 4.8.8 Consecutive\_Vowel

1. Consecutive\_Vowel = count(서브 도메인[aeiou]{2,})

Consecutive\_Vowel는 서브도메인에서 2음절 이상 연속되는 모음 문자열 개수를 센 피처이다.

정상 도메인에서 2음절 이상 연속되는 모음 개수는 예시로 facebook, youtube, google, kshieldjr과 같이 보통 1개인 경우가 대부분이다. 하지만 DGA 도메인의 경우 무작위 위치에서 모음이 생성될 수 있으므로 2음절 이상 연속되는 모음 문자열 개수가 정상도메인보다 많다.

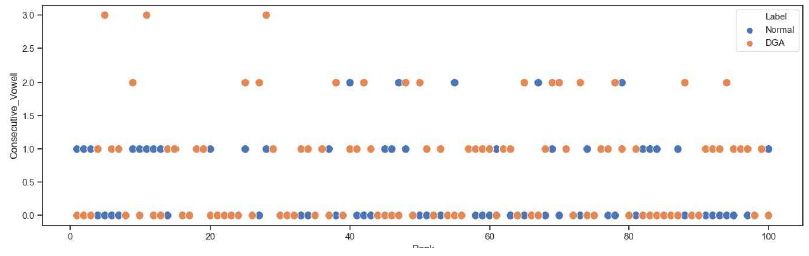


그림 15 정상/DGA 서브도메인 2음절 이상 연속되는 모음 문자열 개수

## 4.8.9 period

1. period = count(도메인[.])

period는 도메인에서 마침표 .를 센 피처이다.

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.

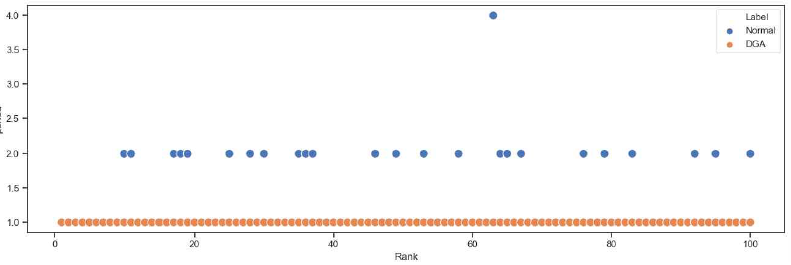


그림 16 정상/DGA 도메인 마침표 . 개수

## 4.8.10 Entropy

1 counts = Counter(도메인)  
2 for alphabet in counts :  
3 prob = [alphabet / 도메인 길이]  
4 for p in prob:  
5 entropy -= sum(p \* log(p) / log(2.0))   
6 return entropy

1. Counter를 사용해 도메인의 알파벳과 개수 counts에 저장
2. Counts 크기 만큼 반복
3. 알파벳별로 확률을 구해 prob 배열에 저장
4. Prob 배열 크기 만큼 반복
5. Entropy 공식을 사용해 계산
6. Entropy 값 반환

Entropy는 도메인의 Shannon 엔트로피 피처이다. 섀넌 엔트로피(Shannon entropy)는 모든 사건 정보량의 기대값으로 문자열에서 알파벳의 종류가 많을수록 다음에 오는 문자를 예측할 수 있는 확률이 낮아져 엔트로피가 높게 나타나게 된다. 정상 도메인에서는 영어에서 사용 빈도수가 높은 알파벳을 주로 쓰기 때문에 알파벳 종류가 적어 엔트로피가 낮지만 DGA 도메인은 무작위의 알파벳과 숫자를 사용하므로 경우의 수가 많아 다음에 올 문자를 예측할 수 있는 확률이 낮기 때문에 상대적으로 정상 도메인에 비해 엔트로피가 높게 나타난다.

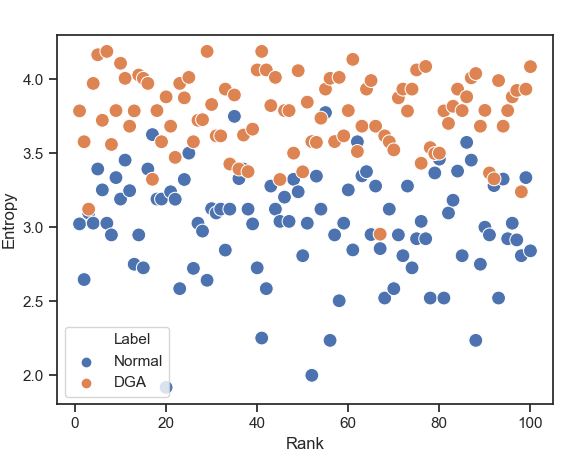


그림 17 정상/DGA 도메인 Entropy

## 4.8.11 Max\_Consecutive\_Consonant

1.Max\_Consecutive\_Consonant= len(max(서브도메인.findall([^.aeiou]{3,})))

Max\_Consecutive\_Consonant는 서브도메인에서 3음절 이상 연속되는 자음, 숫자 문자열 중 최대 길이를 찾는 피처이다. Consecutive\_Consonant 피처처럼 정상 도메인에서는 3음절 이상 연속되는 자음, 숫자 문자열 개수가 적고 발음상의 이유로 3음절이상 연속되더라도 최대 길이가 짧지만 DGA 도메인의 경우 무작위로 생성되고 발음해야 할 이유가 없어 3음절 이상 연속되는 자음 문자열들이 길 확률이 크다.

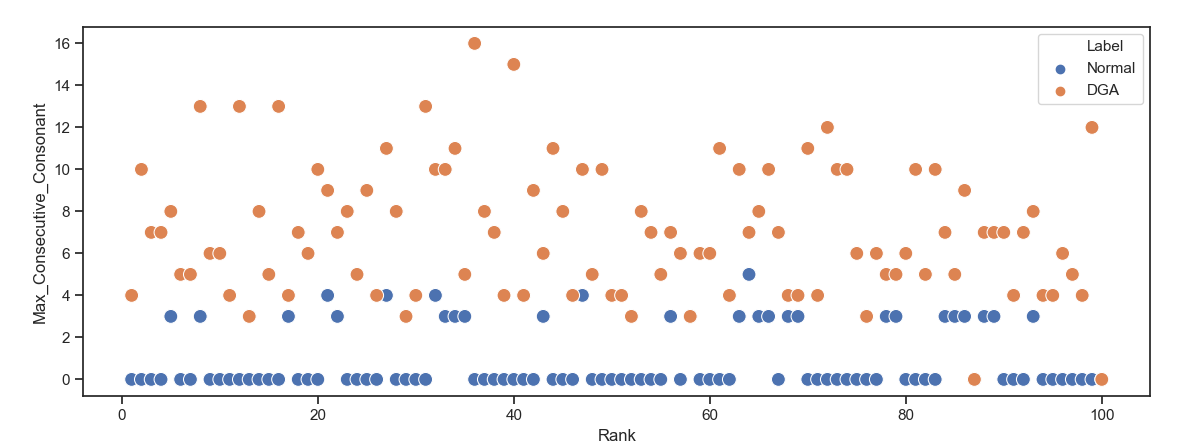
****

그림 18 정상/DGA 도메인 Max\_Consecutive\_Consonant

## 4.8.12 Max\_Voewl\_Consonant

1.Max\_Voewl\_Consonant= len(max(서브도메인.findall([aeiou]{2,})))

Max\_Voewl\_Consonant는 서브도메인에서 2음절 이상 연속되는 모음 문자열 중 최대 길이를 찾는 피처이다. 보통 정상 도메인에서는 2음절 이상 연속되는 모음 문자열 길이가 2음절을 초과하지 않지만 DGA 도메인의 경우 자음, 모음 상관없이 랜덤하게 생성되므로 2음절 이상 연속되는 모음 문자열들이 길 확률이 크다.

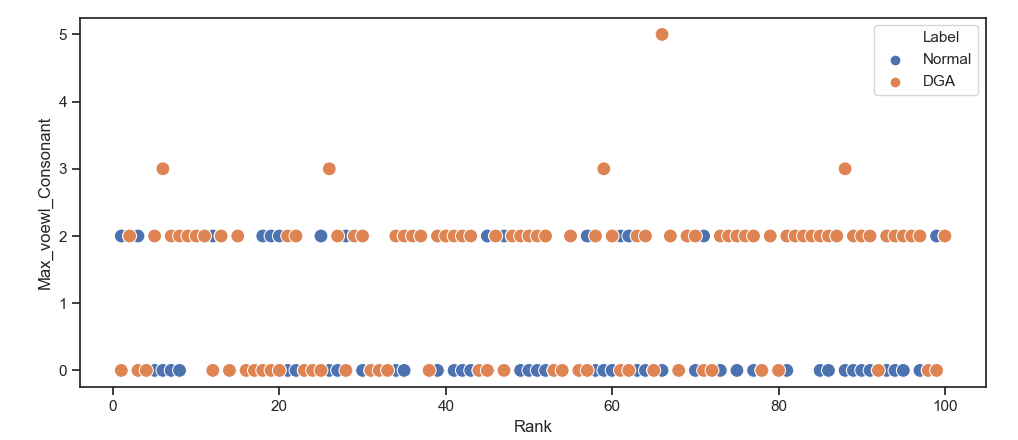
****

그림 19 정상/DGA 도메인 Max\_Voewl\_Consonant

## 4.8.13 Meaning\_count

1 for word in word\_list :  
2 if word in 서브도메인  
3 counts++  
4 return counts

1. 단어 사전 크기 만큼 반복
2. 서브도메인에 단어가 포함돼 있으면
3. 개수를 세는 counts 증가
4. Counts 반환

Meaning\_count는 서브도메인에서 의미있는 단어 개수를 찾는 피처이다.

단어 사전은 1500~2008년까지의 서적에서 특정 단어가 얼마나 많이 사용됐는지 검색할 수 있는 엔진인 Google Books Ngram Viewer를 사용했는데 3음절 이상의 영어 단어 상위 10,000개를 추출해서 만들었다. 정상 도메인은 자주 사용하는 영어 단어들로 이루어져 있지만 DGA 도메인 은 의미없이 랜덤하게 생성되므로 도메인에 단어가 있을 확률이 적다.

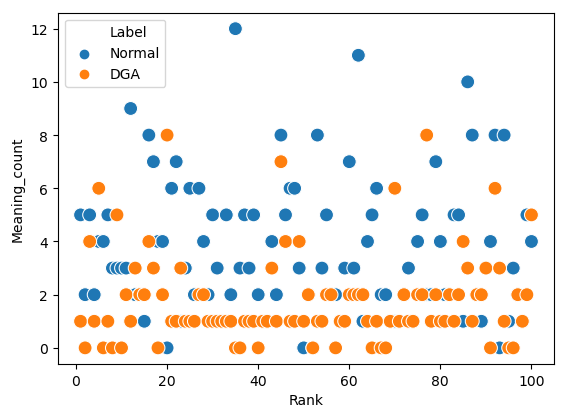


그림 20 정상/DGA 도메인 Meaning\_count

# 5.2 구현 결과

## 5.2.4 AI 모델

1. 데이터셋 관리

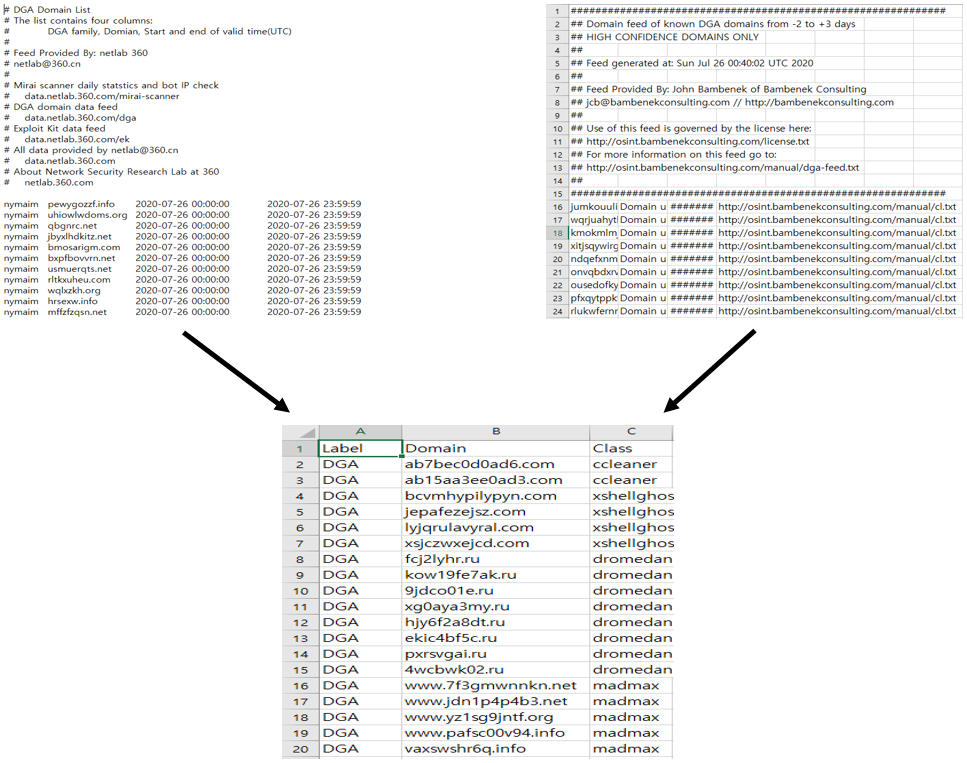


그림 21 데이터셋 통합

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

2. 피처 추출



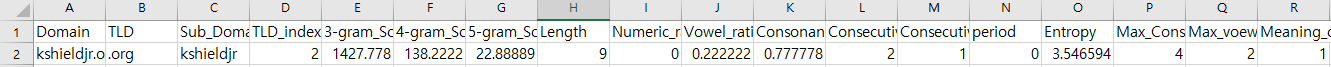
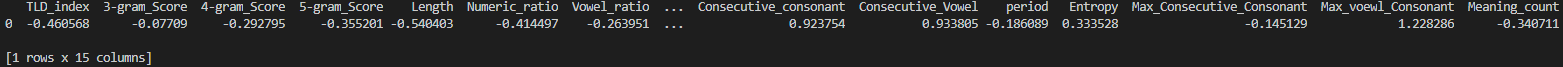


그림 22 피처 추출

새로운 도메인이 들어오면 TLD와 서브 도메인을 분리하고 도메인에 대한 피처 값을 계산하여 출력한다.

3. 피처 스케일링



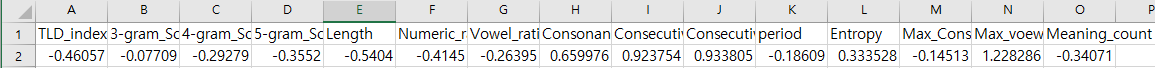


그림 23 피처 스케일링

피처 값을 추출한 원 상태는 피처 마다 값의 범위가 큰데 바로 모델에 판단을 요청하면 판단의 정확도가 낮아져 오탐, 미탐이 발생할 수 있으므로 StandardScaler를 사용해 피처값들을 평균 0, 표준편차가 1이 되도록 스케일링한다.

4. 모델 학습 결과

정상과 DGA 도메인 2,000,000개 중 75%는 학습 데이터 500,000개(25%)를 검증용 데이터로 정하여 여러 모델 알고리즘에 대한 성능평가를 진행했다.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Naive Bayes | Logistic Regression | Random Forest | Decision Tree | LightGBM | 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% |

표1 모델 성능 평가 결과

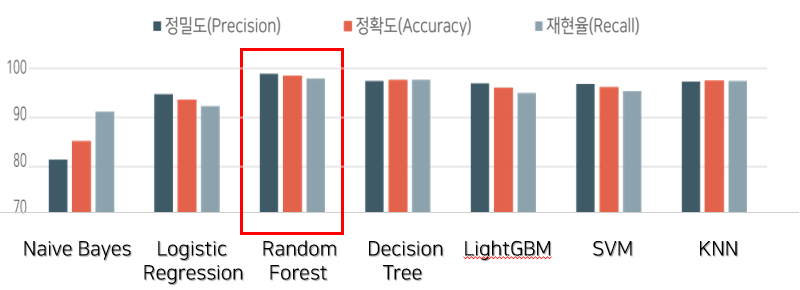


그림 24 모델 성능 평가 결과 그래프

그 결과 Random Forest 모델이 다른 모델보다 정확도, 정밀도, 재현율이 모두 우수하여 시스템에서 사용할 모델로 선정했다.

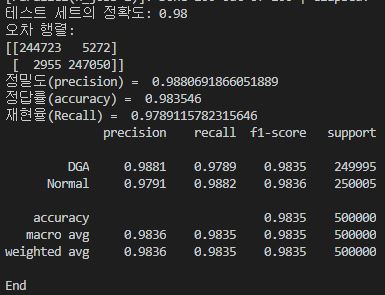


그림 25 Random Forest 모델 학습 결과

5. 실시간 판단

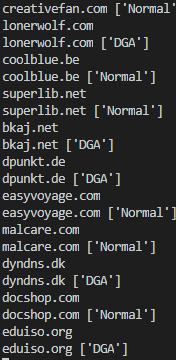


그림 26 도메인 실시간 판단 결과

패킷 미러링을 통해 DNS에 도메인이 오면 머신러닝 모델이 해당 도메인이 정상인지 DGA 도메인인지 실시간으로 판단하게 된다.

# 5. 시스템 구현

## 5.1 개발 언어 및 라이브러리

|  |  |
| --- | --- |
| 개발 언어 및 라이브러리 | 설명 |
| Python 3.8.3 |  |
| * Collections | Entropy 피처 계산을 위해 리스트 원소의 개수를 세는 라이브러리 |
| * Joblib | 모델을 .pkl 형태로 저장, 불러오는 라이브러리 |
| * Itertools | 피처 알고리즘에서 반복자를 만드는 라이브러리 |
| * Math | Entropy 계산을 위한 수학 라이브러리 |
| * Matplotlib | 데이타 시각화 라이브러리 |
| * Numpy | 행렬, 리스트를 처리하는 라이브러리 |
| * Pickle | 피처 계산에 필요한 파일(단어 사전, TLD\_list, N-gram\_list)과 피처 스케일러를 .pkl 형태로 저장, 불러오는 파일 라이브러리 |
| * Pandas | 데이터 조작 및 분석 라이브러리 |
| * Sklearn | 머신러닝 모델 라이브러리 |
| * Seaborn | Matplotlib을 기반으로 향상된 데이터 시각화을 위한 라이브러리 |