단어 군집화를 통한 스미싱 탐지 규칙 추천

: 신종 스미싱 단어 탐지를 중심으로

방지훈 정구섭 박재민
곽동환 민평홍 오치현

INDEX

- 문제 배경
- 문제 정의
- 프로젝트 목표

프로젝트 개요

문제 해결 프로세스

- 데이터수집 및 전처리
- 키워드 중요도 산출
- 시계열 패턴 유사도 산출
- 시계열 패턴 군집화

- 군집 카테고리화
- 유효성 검증

프로젝트 결과

기대 효과 및 고찰

문제 배경

● 스미싱 정의

SMS와 피싱(Phishing)의 합성어로 문자메시지를 이용한 휴대폰 해킹 기법

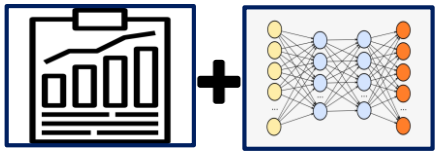
● 스미싱 유형

[Web발신]
[KISA 보안공지] 중요공지사항 http://sge3s.xyz

[Web발신]
[CJ대한통운]주문하신물품.미배달 도로명불일치.수정하세요 https://han.gl/zt6Uz

● 스미싱 피해 사례

● 현재 스미싱 탐지 시스템



이상탐지가 분류 성능은 뛰어나지만 오류에 대한 위험부담 및 탐지 속도 향상을 위해 오용탐지가 필요

● 규칙 기반 오용탐지

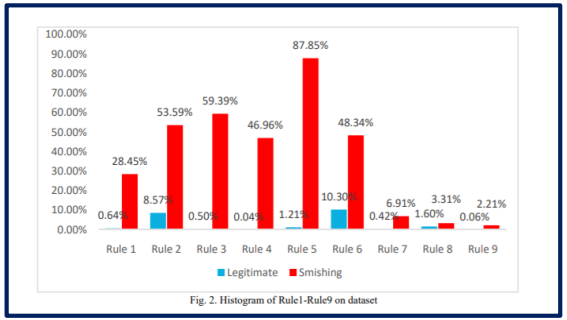


Fig. 2. Histogram of Rule1-Rule9 on dataset

규칙 기반 탐지 시스템에서는 스미싱 의심 단어 기반 규칙이 탐지 성능에 가장 큰 요인 (Ankit Kumar Jain et al. 2018)

문제 정의

● 선행 연구

논문 제목	요약
Knowledge Graph-based Korean New Words Detection Mechanism for Spam Filtering.	스미싱 차단 방지 위해 변화된 방식으로 스미싱 전송.
A Normalization Method of Distorted Korean SMS Sentences for Spam Message Filtering	문자 메시지 자동 차단 회피 위해 문자 내용을 다양한 형태로 변형하거나 왜곡시키고 있다.
A SVM-based spam filtering system for short message service (SMS)	현재 스미싱 필터링 기능은 발송된 다양한 종류의 스미싱 메시지를 적절히 필터링할 수 없다.

다수의 논문에서 **단어의 유형 변화**를 문제점으로 제시

● 신종 스미싱 유형

(출처 : KISA 탐지팀)

2017년	2019년	2020년
요금미납, 쿠폰, 대통령, 탄핵	버닝썬, 동영상	n번방, 마스크, 코로나, 재난지원금
공통 유형		
청첩장, 택배		

● 단어 규칙의 문제점

단어 규칙 시스템

신종 스미싱 키워드 →

탐지 실패

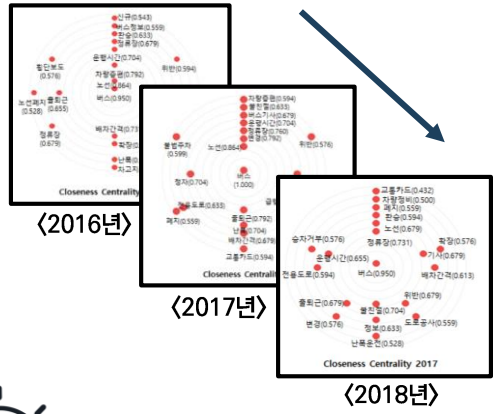
단어 규칙 시스템은 탐지 속도가 빠르고 오류가 적지만
단어 **유형 변화 대응에 한계점이 존재**.

프로젝트 목표

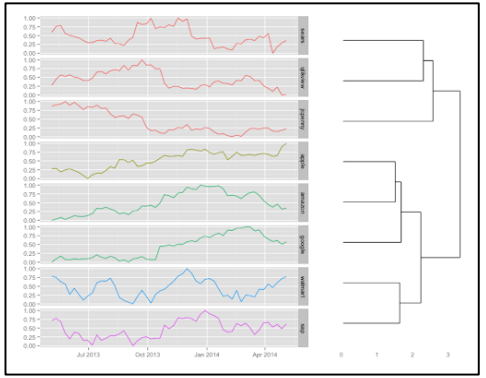
단어 규칙의 개념변화 문제

사회적인 이슈 단어가
들어간 신규 스미싱 단어가
지속적으로 생성됨.

단어 중요도



시계열 패턴 군집화



“ 단어 중요도의 시계열 패턴 군집화를 통한
신종 단어 군집 탐지 및 단어 규칙 시스템 개선 ”

문제 해결 프로세스

- 기존 시스템

단어 규칙 기반 분류 모델

〈빈도수 기반〉

〈스미싱 이진 분류〉

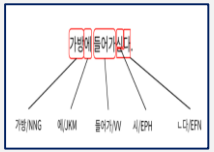
- Process

step1

데이터 수집 &전처리

KISA 한국인터넷진흥원

스미싱 문자



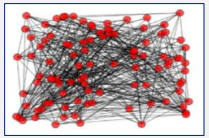
Tokenization

주 단위 split

step2

단어 중요도 산출

주별 동시 출현 행렬



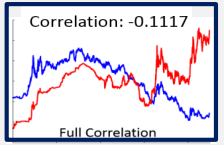
단어 네트워크 생성

$$C_i = \frac{1}{\sum_{i,j \in v} d_{ij}}$$

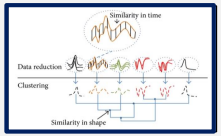
근접 중심도 계산

step3

시계열 패턴 군집화



단어 간 상관계수

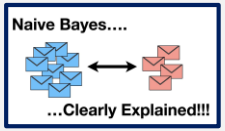


K-means clustering

step4

신종 단어 군 선택 및 검증

상승, 빈출, 신규 단어 군 선택



Naïve-bayes BoW 검증

- Ours 개선 시스템

〈빈도수 + 신종 단어 기반〉

〈스미싱 이진 분류〉

데이터

● 데이터 확보 현황 : 약 100만개의 스미싱 문자 데이터

수집 기간	범위	수집 경로
2016년 11월 ~ 현재	스미싱 문자 내용, URL	KISA에서 데이터를 제공받음

KISA 한국인터넷진흥원

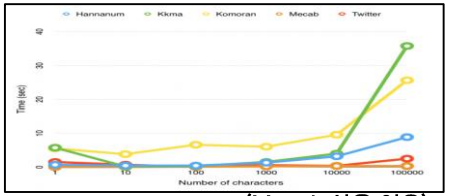
idsmlist.mess: only_hangul
2020-03-04 [한진택배] 상품 보관 완료 (입구 보관 완료)사진보기
2020-03-04 송장번호 [566*****6] 미확인입니다. 반송처리하오니 주소 확인
2020-03-04 [롯데택배] 오후 16:35 완료상품 배송. 고객 문 앞 2박스. 사진
2020-03-04 택배입니다. 배송지주소 확인부탁드립니다.

● 데이터 전처리

SMS 데이터

Preprocessing

한글만 남기고 삭제 → Mecab Pos-Tagging

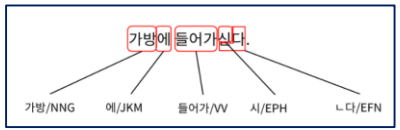


〈Mecab 사용 이유〉

Tokenization

Noun 추출
('NNG','NNP','NNB','NP')

주 단위 split



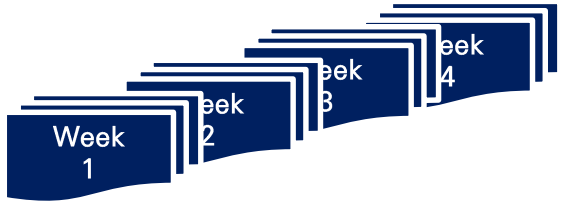
〈품사 태깅〉

키워드 중요도 산출

(1) 단어 네트워크, 근접 중심성

● 단어 네트워크

주 단위 Split 데이터

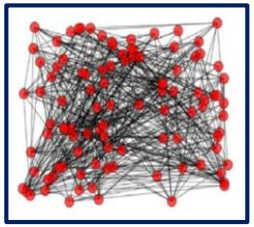


Co-occurrence matrix

	한진택배	정보	주소	배달	수정
한진택배	0	4	25	1	4	3	3
정보	4	0	8	4	7	6	5
주소	25	8	0	8	15	5	23	2
배달	1	4	8	0	4	1	1
수정	4	7	5	4	0	3	3

문서 내 단어간 동시출현을 고려하여 동시 출현 행렬 생성

단어 네트워크



동시 출현 행렬과 가중치 행렬을 요소별로 곱하여 단어 네트워크 생성

근접 중심성

$$C_i = \frac{1}{\sum_{i,j \in v} d_{ij}}$$

시계열 단어 근접 중심성

	10개	11번가	CJ통운	G마켓	N번방
2016.04.01	0.000000	0.000000	0.000000	0.000000	0.000000
2016.04.08	0.000000	0.000000	0.000000	0.000000	0.000000
2016.04.15	0.000000	0.000000	0.000000	0.000000	0.000000
2016.04.22	0.000000	0.000000	0.000000	0.417526	0.000000
2016.04.29	0.000000	0.000000	0.000000	0.411043	0.000000
2016.05.06	0.000000	0.000000	0.000000	0.419708	0.000000

산출한 근접중심도를 통하여 시계열데이터 생성

d_{ij}: shortest distance between node i and node j

시계열 데이터 유사도 산출

(1) 시계열 패턴 유사도

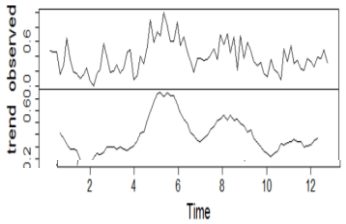
● 단어 네트워크

단어 근접 중심성

	10개	11번가	CJ통운	GD마켓	N번방
2016.04.01	0.000000	0.000000	0.000000	0.000000	0.000000
2016.04.08	0.000000	0.000000	0.000000	0.000000	0.000000
2016.04.15	0.000000	0.000000	0.000000	0.000000	0.000000
2016.04.22	0.000000	0.000000	0.000000	0.417526	0.000000
2016.04.29	0.000000	0.000000	0.000000	0.411043	0.000000
2016.05.06	0.000000	0.000000	0.000000	0.419708	0.000000

산출한 근접중심도를 통하여 시계열데이터 생성

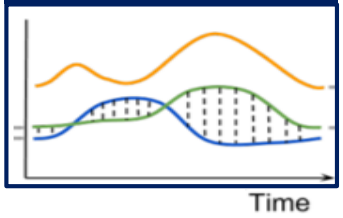
MA (Moving Average)



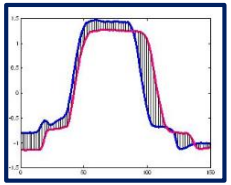
MA로 noise제거
Trend 확인

● 단어 근접 중심성 시계열 간의 Similarity measure 산출 및 선택

시계열 데이터

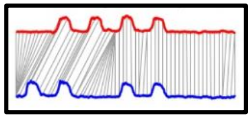


Euclidean



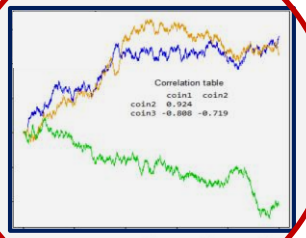
$$ED(Q,C)=\sqrt{\sum_{i=1}^{n}(q_i-c_i)^2}$$

DTW (Dynamic Time Wraping)



$$DTW(X,Y)=\min_{r\in M}\left(\sum_{m=1}^{M}|x_{im}-y_{jm}|\right)$$

Correlation



$$COR(X,Y)=\frac{\sum_{n=1}^{N}(x_n-\bar{X})(y_n-\bar{Y})}{\sqrt{\sum_{n=1}^{N}(x_n-\bar{X})^2}\sqrt{\sum_{n=1}^{N}(y_n-\bar{Y})^2}}$$

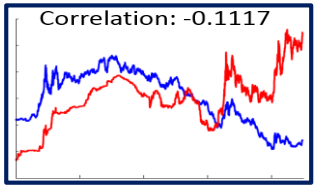
Correlation을 Similarity measure로 선택.
(n → n x n matrix)

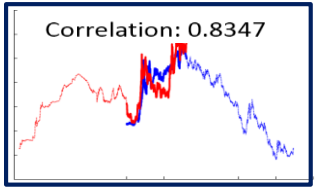
시계열 데이터 군집화

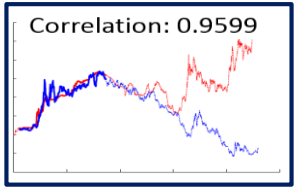
(1) 시계열 패턴 군집화

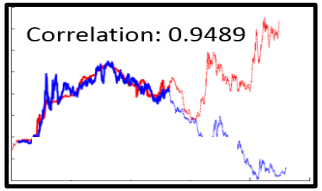
● 시계열 데이터들의 Correlation 값

예시



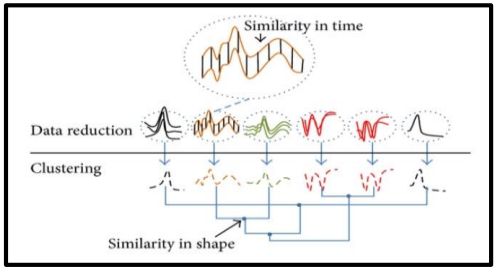




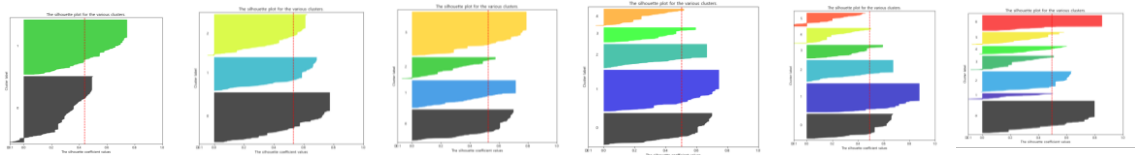


시계열 데이터들의 Correlation 값을 통해 군집화 준비

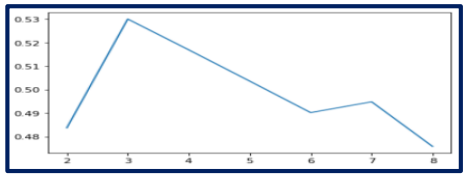
● 시계열 데이터 군집화



k-means와 계층적 군집화 중 정량적으로 정확한 군집을 정의해 줘야 하기 때문에 k-means 선택



〈K = 2〉 〈K =3 〉 〈K = 4〉 〈K = 5〉 〈K = 6〉 〈K = 7〉



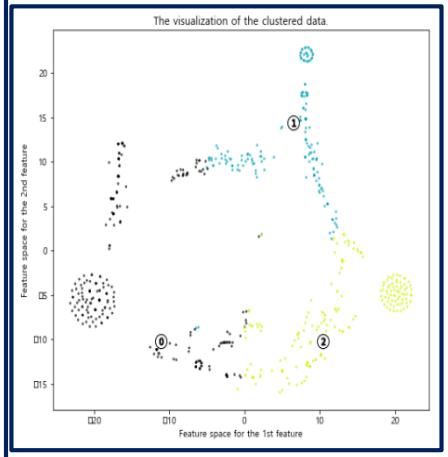
실루엣 스코어를 통해 Best K인 3을 군집 수로 선택

시계열 데이터 군집화

(2) 시계열 패턴 시각화

빈출
신규
급상승

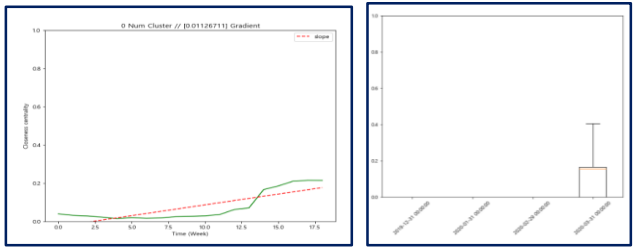
● 군집화된 시계열 데이터 시각화



K= 3일 때 Best 군집 수

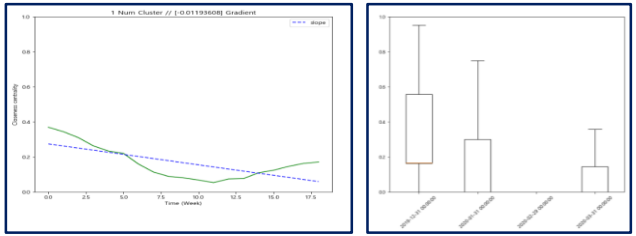
● 군집 별 단어들의 근접중심도 변화

1번 군집 + 곡선, box plot



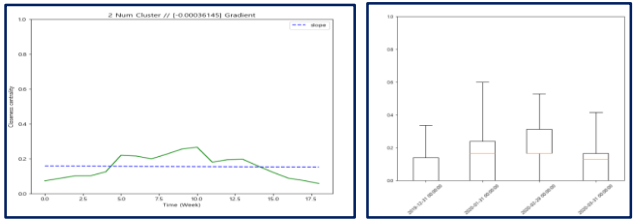
11번가 G마켓 N번방 가계 가능 가맹점 개월 거래 거부 거주지 대한통운 주소 상환 등기 등록 루트 링크 마스크 결제 결제시 배송 결혼 결혼식 겸용 경채 계약 한진택배 국내외 국외 국제 기간 코로나 대금 대리 대상 경비실 경우 경제 만원 모즈 미지 미확인 바이러스 박스 반품 확인 우한 군산시 귀하 그림 금리 금액 금융

2번 군집 + 곡선, box plot



10개 가입 동의 라이프 로젠택배 롯데캐피탈 매장 서비스 세트 센터 소재 소포 속달 손해 쇼핑몰 수령인 수자 메시지 메일 명칭 문의 물건 물류 물품 배달 보험 보험금 본부 본인 부탁 불가 이름 불명 사유 사이트 가정 개인 거리 고객 고침 과장 국민 국민카드 도로명 상세 상태 상품 미배달 미배송 미수령 바람 배송 배송원 번호 변경 보상 반송 반송물 색상 서류 서명

3번 군집 + 곡선, box plot

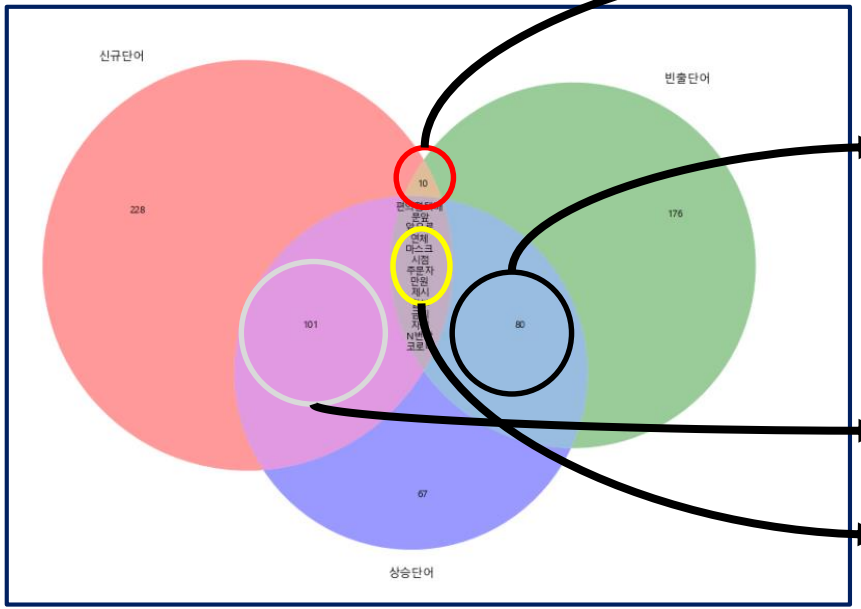


CJ통운 가운데 가지 간직 감사 갤럭시 거름 거목 걱정 건강 걸음 관원 관절 관절염 광고 구절초 구형 그대 그루 때문 리아 마음 겁니다 격려 결실 겸손 고가 고유 공지 관리 만남 말씀 맹세 모바일 모습 무릎 문구 문제 미납 미래 부분 부작용 부족 불법 불편 글루코사민 기기 기능 다운 다운로드 다짐 당귀 독성 동안 비법 뿌리 사랑 사은품 상담 상환

군집 카테고리화

(1) 군집 별 카테고리 부여

● 군집 별 카테고리



〈기간: 2020.01~2020.04〉

빈출, 신규 단어

효과,창구, 우정사업본부, 보험금,관절, 여사…….

상승, 빈출 단어

바이러스, 경비실, 반품, 약정, 이자, 축하, 현관, 실시간, 현대, 배달원, 국제…….

상승, 신규 단어

동양, 신천지, 연회비, 신상, 닷컴, 11번가, 군산시, 청원, 체결……

상승 , 빈출, 신규 단어

편의점택배, 쿠팡, 연체, 마스크, N번방,코로나, 금리, 할부…….

빈출, 신규, 상승 단어들을 규칙기반 분류 모델에 신종 스미싱 단어로 판단

유효성 검증

(1) 신종 단어 적용 스미싱 필터링(Naïve Bayes BoW)

● Naïve Bayes 필터링 모델(기존 BoW 방식 vs 신종 단어 적용 방식)

〈기존 BoW 방식-빈도수 기반〉

번호	타입	분류결과	텍스트
1	스미싱	스미싱	['촛불', '시위', '의무', '경찰', '폭행', '영상', '아래', '링크', '포함']
2	정상	스미싱	['광고', '신한', '카드', '일상', '숨', '필요', '시점', '신한', '가입', '이디야', '커피', '제공','이벤트','수신', '거부', '무료']
~~			
120,653	스미싱	스미싱	['속보', 'N번방', '전체', '회원', '신상', '공개']
120,654	정상	스미싱	['용산', '구청', '정정', '코로나', '번', '번', '확진', '남영동', '해외', '입국', '발생', '홈페이지', '블로그']

N= 120,654		예측	
		스미싱	정상
실제	스미싱	96,544	4,507
	정상	1,051	18,551

F1-Score : 0.87

〈신종 단어 적용 방식〉

번호	타입	분류결과	텍스트
1	스미싱	스미싱	['촛불', '시위', '의무', '경찰', '폭행', '영상', '아래', '링크', '포함']
2	정상	정상	['광고', '신한', '카드', '일상', '숨', '필요', '시점', '신한', '가입', '이디야', '커피', '제공','이벤트','수신', '거부', '무료']
~~			
120,653	스미싱	스미싱	['속보', 'N번방', '전체', '회원', '신상', '공개']
120,654	정상	스미싱	['용산', '구청', '정정', '코로나', '번', '번', '확진', '남영동', '해외', '입국', '발생', '홈페이지', '블로그']

상승 , 빈출, 신규 단어

편의점택배, 쿠팡, 연체, 마스크, N번방,코로나, 금리, 할부…….

N= 120,654		예측	
		스미싱	정상
실제	스미싱	96,992	4,059
	정상	550	19,052

F1-Score : 0.89

기대 효과 및 고찰

● 기대 효과

신종 스미싱을
정상문자로 분류하는 오탐률 감소

신종 스미싱 발생을
빠르게 탐지하여 조기 대응

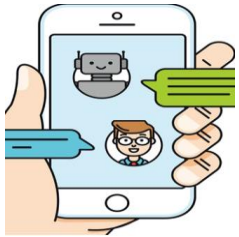
신종 스미싱으로 인한
금전적 피해 예방 및 개인정보 유출 감소

● 고찰

단어의 전처리 품질에 영향을 받는다.
다른 규칙과 함께 사용될 때의 효과 또한
고려해 볼 필요가 있다.

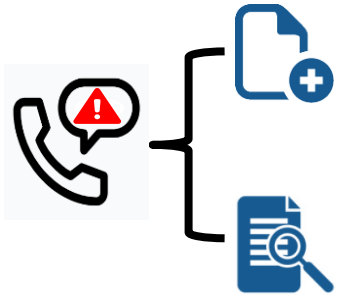
● 적용 가능성

챗봇 학습



지능형 대화형 챗봇에 적용하여,
대화에 신조어가 탐지되었을 때 자동으로
지식 그래프에 추가하여 그 의미를
이해할 수 있도록 적용

보이스 피싱



새롭게 스미싱 규칙에 추가된 단어들을
보이스 피싱 탐지 규칙에도 추가

보이스 피싱 음성을 텍스트로 변환 후
프로젝트의 프로세스를 적용하여 신규
단어 규칙 추천.

REFERENCES

▶ Alain Saas, Anna Guitart, África Periáñez. Discovering Playing Patterns: Time Series Clustering of Free-To-Play Game Data

▶ 유승의, 홍순구, 이태헌, 김나랑. A Pattern Analysis of Bus Civil Complaint in Busan City Using the Text Network Analysis

▶ Kim, Jihye and Okran Jeong. “Knowledge Graph-based Korean New Words Detection Mechanism for Spam Filtering.” (2020).

▶ Yong-Bum Cha , Won-Yong Choi , Gang-Seok Lee
Korea University , Korea , Korea Financial Security Institute
Proactive Response Against Smishing Using Shorten URL

▶ Kang, Seungshik. (2014). A Normalization Method of Distorted Korean SMS Sentences for Spam Message Filtering.

▶ Joe, Inwhee & Shim, Hye-Taek. (2009). A SVM-based Spam Filtering System for Short Message Service (SMS).