기계학습 프로젝트 제안서

아이디어 스케치 (초안)

팀 인절미

작성자:
201611144 한경욱
201611122 김민

목차

- 서비스
- 필요성
- 프로토타입
- 개발: 알고리즘, 데이터, 소스코드, 서베이, 개발 전략
- 사업화 가능성

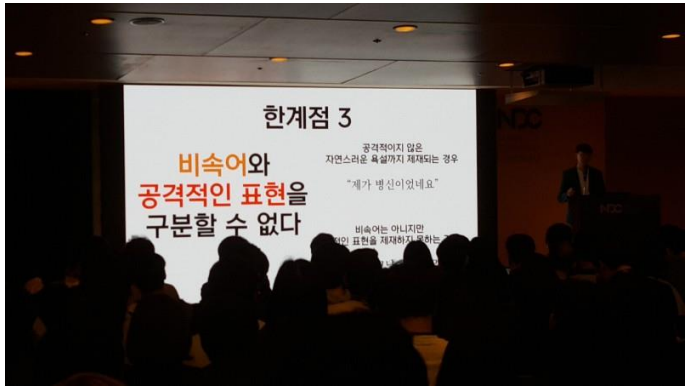
서비스

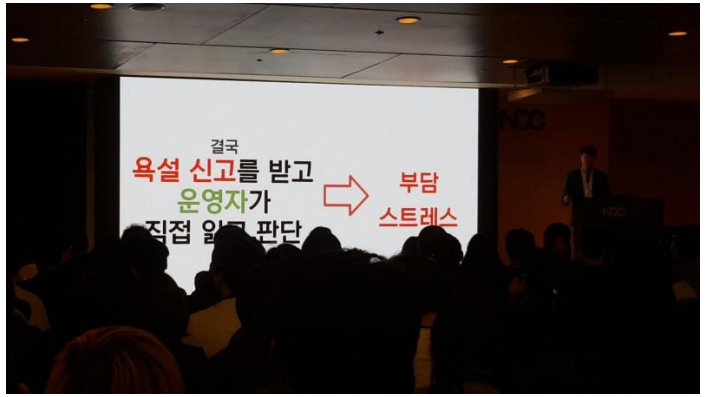
- 욕설 감지 프로그램 -> 어찌보면 좀 식상하지 않나?

https://m.blog.naver.com/PostView.naver?isHttpsRedirect=true&blogId=alice780&logNo=221260702527

https://www.inven.co.kr/webzine/news/?news=198156&site=honkai3rd

http://ndc.vod.nexoncdn.co.kr/NDC2018/slides/NDC2018_0033/index.html





서비스

- 욕설 감지 프로그램 -> 어찌보면 좀 식상하지 않나?

- 그래서 만들었습니다! 우리으 ㅣ서비스 :
- 서비스 이름: 쉿! (어라?, 어?, 다? 신고하겠습니노, 언급 금지, 병먹금, etc -> 이런 느낌)

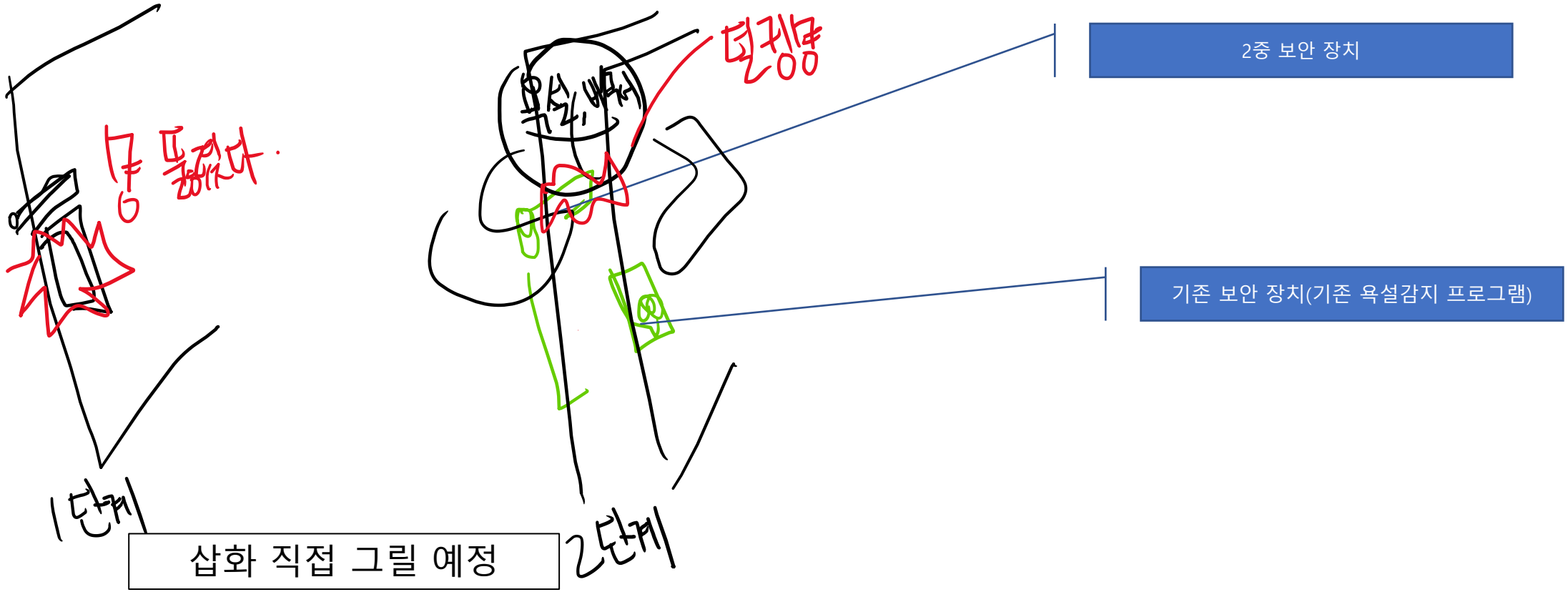
- 식상하지 않은 이유
- -> 기존 서비스는~ 운영자(관리자)의 부담 증대
- -> 우리 서비스는~ 이러한걸 추가할거에요

기존 사례, 일반적으로 쉽게 생각할 수 있는 경우
- 채팅 자체를 가지고 판단

우리 팀에서 개발하고자 하는 모델
- 채팅 자체를 가지고 판단
- 채팅이 이뤄지고 난 후 문맥, 상황 파악

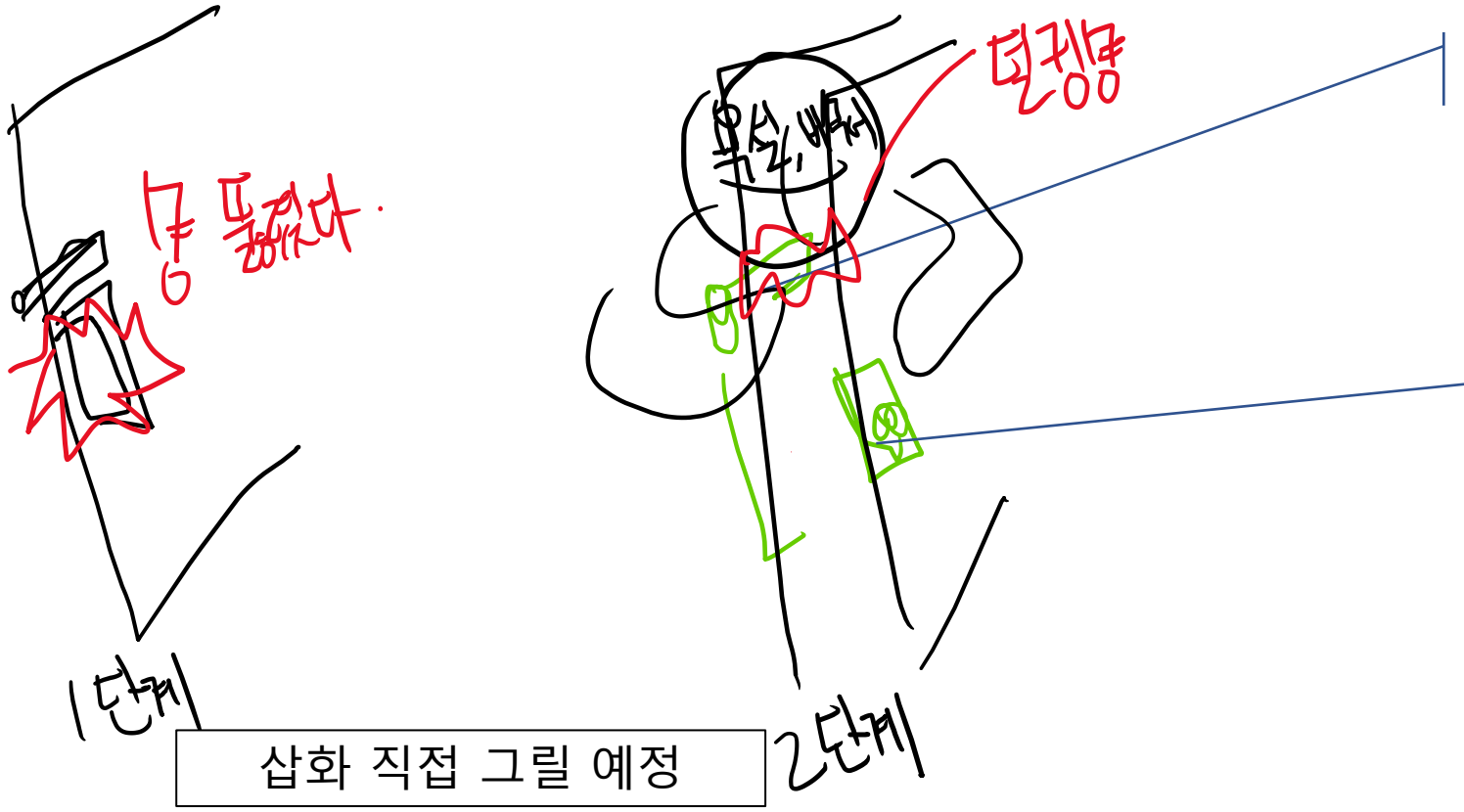
필요성

- 실시간으로 나쁜 언행을 막아주는 프로그램 vs (로그를 통해서) 나쁜 언행을 사용한 사람을 효과적으로 검출하는 프로그램



필요성

- 실시간으로 나쁜 언행을 막아주는 프로그램 vs (로그를 통해서) 나쁜 언행을 사용한 사람을 효과적으로 검출하는 프로그램



삽화 직접 그릴 예정

2중 보안 장치

이게 우리가 추가하고자 하는 기능
1) 채팅이 이뤄지고 난 후의 전후 상황을 파악

기존 보안 장치(기존 욕설감지 프로그램)

이게 기존 비속어 필터 시스템
1) 직접 작성: 욕설
2) 특수 문자: 욕@설 / 욕 " 설 / 욕!설 / 욕1설
3) 단타: 욕
   설
4) 욕설욕설
5) 우회 및 변형: OHㅁI (오 에이치 미음 아이)

필요성

- 실시간으로 나쁜 언행을 막아주는 프로그램 vs (로그를 통해서) 나쁜 언행을 사용한 사람을 효과적으로 검출하는 프로그램

삽화 직접 그릴 예정

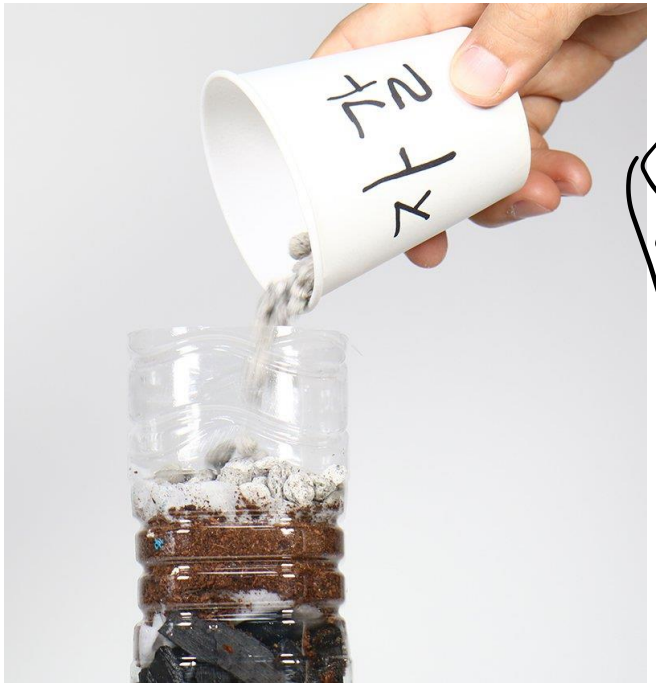
2중 보안 장치

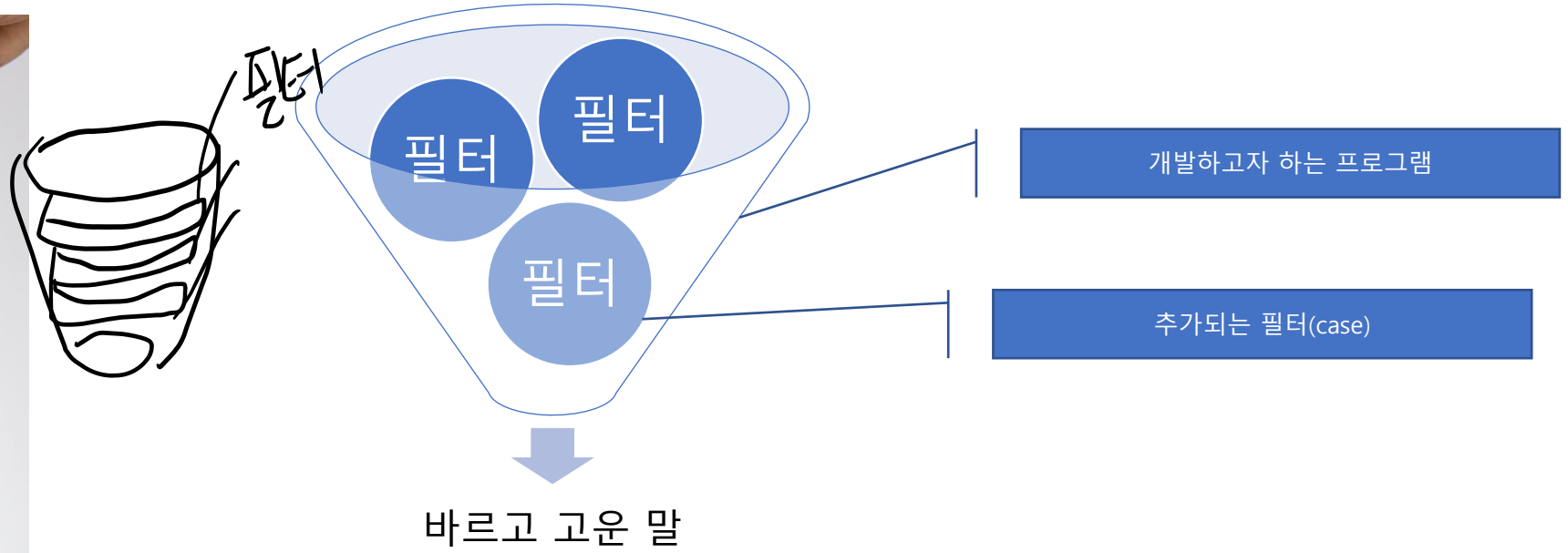
이게 우리가 추가하고자 하는 기능
1) 채팅이 이뤄지고 난 후의 전후 상황을 파악

Ex)
1. 특정 어휘를 사용한 이용자가 다른 유저에게 차단을 당하는 경우 (신고를 당한 경우)
2. 특정 어휘가 사용된 이후 채팅량 급증
3. ~~특정 행동을 유도한다.~~(어려움 / 어떤 시스템에 사용될지_이식성 문제)
   롤을 대상으로 만들면 메이플에 적용 불가능
   게임을 대상으로 만들면 유튜브에 사용 불가능
4. 해당 어휘를 본 다른 사람이 채팅방에서 퇴장하는 경우
5. 특정 어휘를 사용한 사람이 강퇴를 당하는 경우

개발 전략

- 실시간으로 나쁜 언행을 막아주는 프로그램 vs (로그를 통해서) 나쁜 언행을 사용한 사람을 효과적으로 검출하는 프로그램



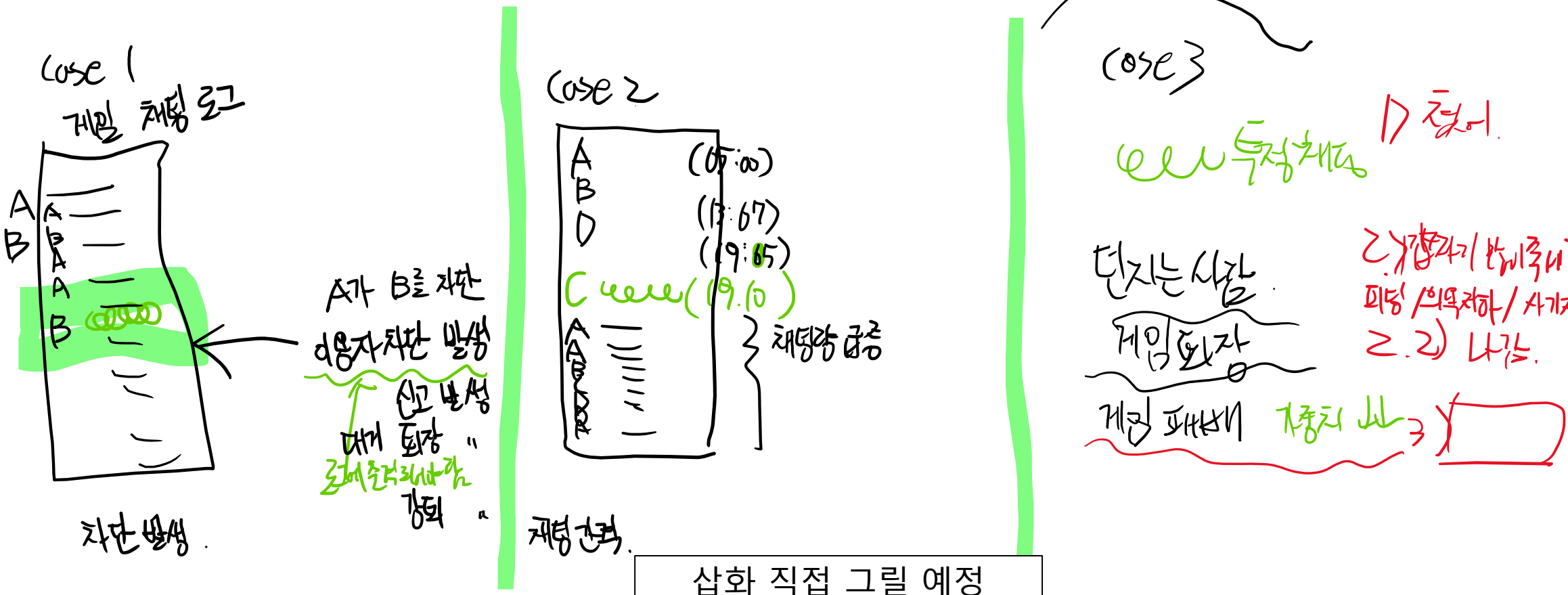


https://post.naver.com/viewer/postView.nhn?volumeNo=26275452&memberNo=9406188

삽화 직접 그릴 예정

개발 전략

- 예시) LoL



삽화 직접 그릴 예정

개발 전략

- 데이터 수집

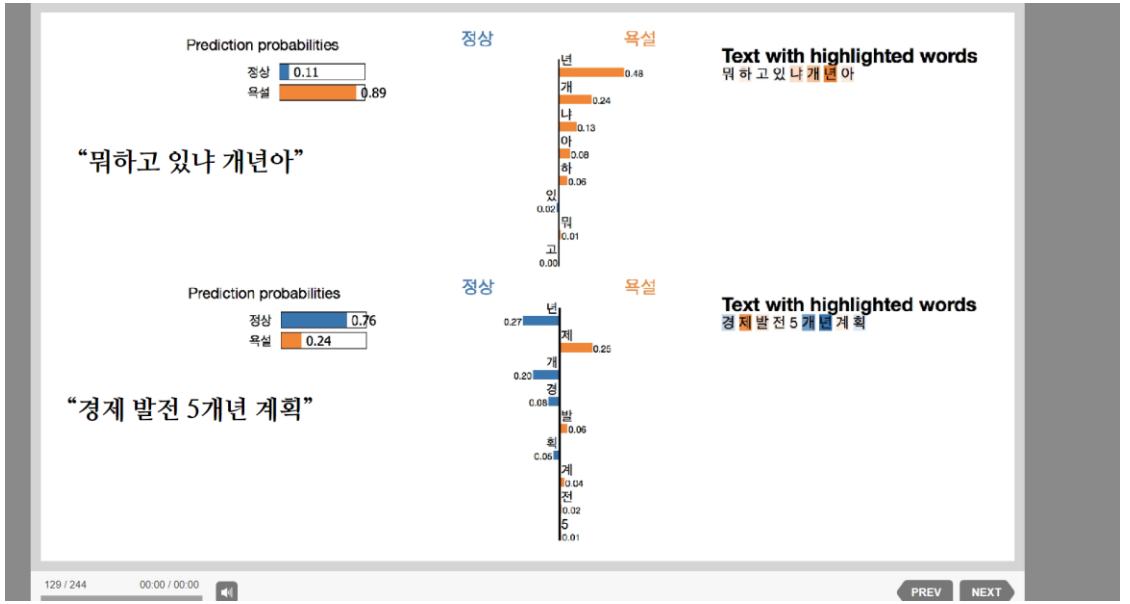
- -> 직접 만들수도 있고
- 채팅 로그 다운 받을 수도 이쏙
- 김민이 친구랑 칼바람 나락 즐기고 직접 데이터 만들어서 찾아내오기
- 만들어 낼꺼면 -> 우리가 데이터를 일부 조작(우리 서비스를 위해서) -> 차단 내역을 채팅 로그에 추가

- Etc

- 사례 찾아야할 듯

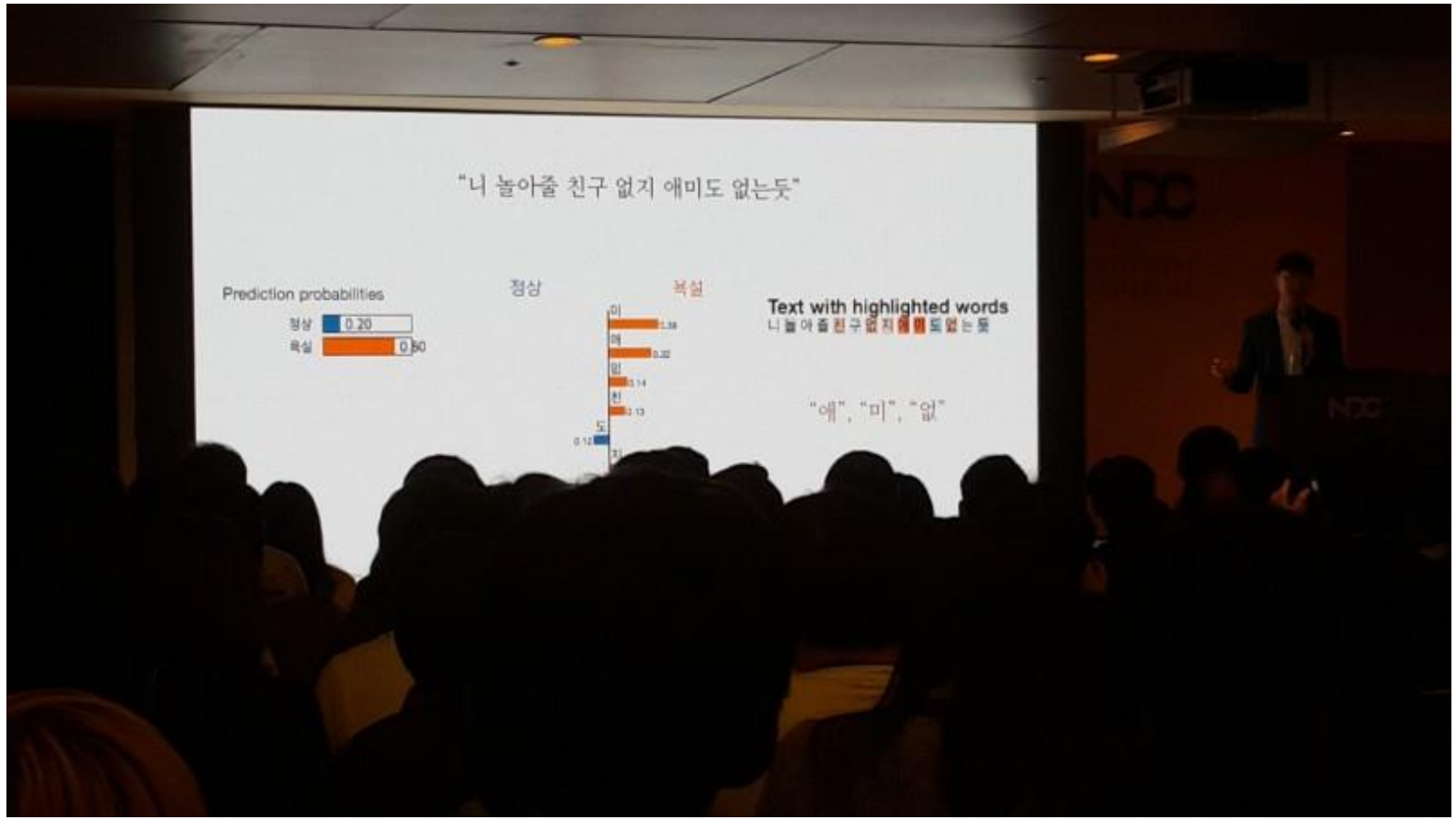
개발 전략

- 알고리즘
- 가장 많이 이용된 알고리즘은 CNN이라던데,,, 찾아봐야할듯
- 특이점 감지? 알고리즘?



개발 전략

- 서베이


"니 놀아줄 친구 없지 애미도 없는듯"
Prediction probabilities
정상
욕설
0.20
0.80
정상
욕설
Text with highlighted words
"애", "미", "없"

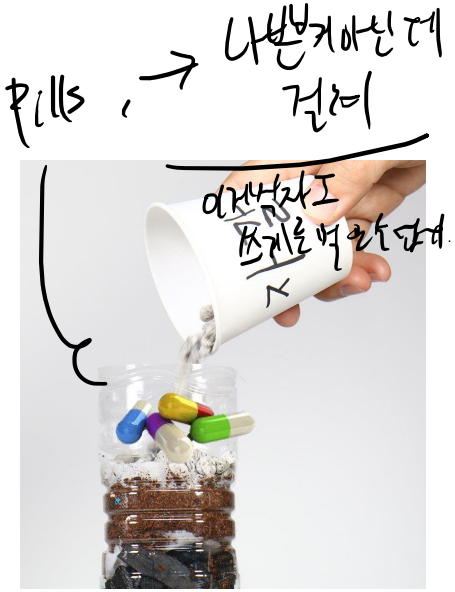
필요성

- 성인에 비해서 판단력이 부족한 어린이나 청소년이 무분별하게 컨텐츠와 사람들의 반응에 노출될 수 있다.
- 은어와 비속어는 빠르게 등장한다.
- 직접 단어를 추가하고 제어하는 형식의 서비스 관리는 장기적인 관점에서 어려움에 처할 가능성이 높아진다.
- + 텍스트(채팅)만을 이용하는 것은 그 한계가 뚜렷하다.
- + 우리가 사용할 수 있는 채팅 외적인 문맥을 파악하면 더 효과적으로 막아낼 수 있을 것이다.
- + (추가 필요)
- + (추가 필요)

- 운영자에게 부담과 스트레스를 가중시킨다.
- https://m.blog.naver.com/PostView.naver?isHttpsRedirect=true&blogId=alice780&logNo=221260702527

주안점

알약 먹자고 쓰레기를 먹을
수는 없기 때문
-> 쓰레기를 완벽히 배척하자
-> 이게 우리 모토



- 주안점
- 목표: 욕하는 사람을 잡아내는 것
- 억울하게 채팅 권한을 상실하는 이용자가 발생할 수 있다. -> 억울한 사람이 발생하더라도 강하게 대처하자.
  -> 전체적으로 클린한 대화 환경을 만들기 위해서
- 교묘하게 우회하는 나쁜 언행을 일삼는 사람이 많다.
- 욕설을 사용하지 않지만 상대방에게 공격적인 언행을 하는 것

- 관리자의 노고? -> 억울한 풀어주려고 노력하는 관리자으 | 노고? X -> 억울한 애들(알약같은 존재들)은 그냥 빌미를 주지마라
  -> 점진적으로 줄여나가기 위해서 욕설하는 하는 사람들이 발생했을 때 전후 상황을 정확하게 파악 -> 장기적으로 좋은 시스템

- 실시간으로 나쁜 언행을 막아주는 프로그램 vs (로그를 통해서) 나쁜 언행을 사용한 사람을 효과적으로 검출하는 프로그램

삽화 직접 그릴 예정

프로토타입

- 실시간으로 나쁜 언행을 막아주는 프로그램 vs (로그를 통해서) 나쁜 언행을 사용한 사람을 효과적으로 검출하는 프로그램

- 트래픽이 가장 많은 시점 혹은 채널의 채팅 로그 입력 (채팅을 포함한 문맥적 데이터가 수집된 경우)
- -> 데이터 취합하여 학습
- -> 서비스 내 적용
- -> 나쁜 언어를 사용하는 이용자 발생
- -> 이용자 계정 정보 추적
- -> 계정 정보를 채팅 관리 시스템에 전송
- -> 해당 이용자의 채팅 시스템 차단

- -->> 각기 다른 적용 방법 (컨텐츠 차단, 이용자 차단, 로그 전송etc)

프로토타입

- 실시간으로 나쁜 언행을 막아주는 프로그램 vs (로그를 통해서) 나쁜 언행을 사용한 사람을 효과적으로 검출하는 프로그램

- 트래픽이 가장 많은 시점 혹은 채널의 채팅 로그 입력 (채팅을 포함한 문맥적 데이터가 수집된 경우)
- -> 데이터 취합하여 학습
- -> 서비스 내 적용
- -> 나쁜 언어를 사용하는 이용자 발생
- -> 이용자 계정 정보 추적
- -> 계정 정보를 채팅 관리 시스템에 전송
- -> 해당 이용자의 채팅 시스템 차단

실시간 감지 및 차단

사후 관리 시스템

- -->> 각기 다른 적용 방법 (컨텐츠 차단, 이용자 차단, 로그 전송etc)

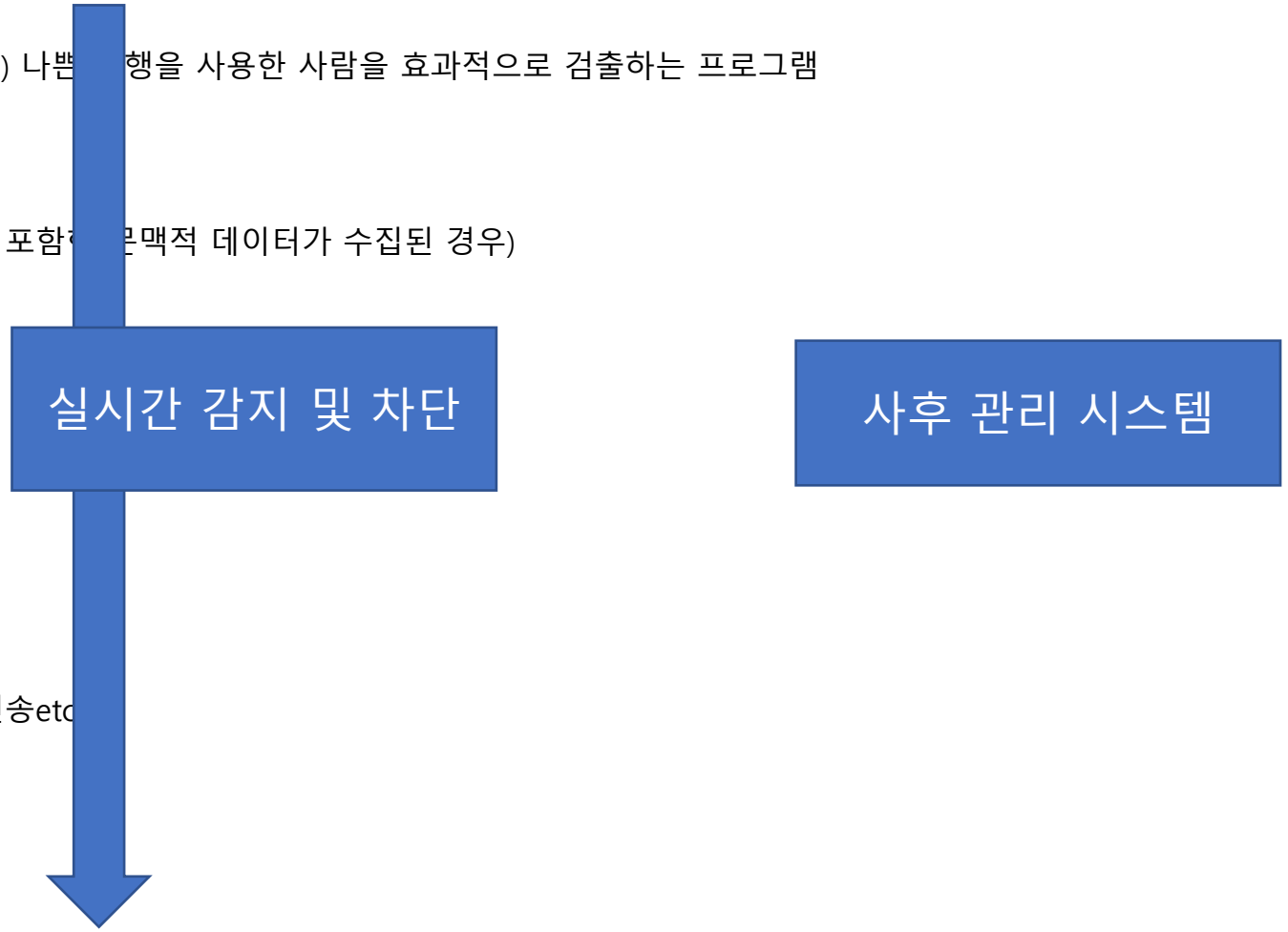
삽화 직접 그릴 예정

프로토타입

- 실시간으로 나쁜 언행을 막아주는 프로그램 vs (로그를 통해서) 나쁜 언행을 사용한 사람을 효과적으로 검출하는 프로그램

- 트래픽이 가장 많은 시점 혹은 채널의 채팅 로그 입력 (채팅을 포함한 문맥적 데이터가 수집된 경우)
- -> 데이터 취합하여 학습
- -> 서비스 내 적용
- -> 나쁜 언어를 사용하는 이용자 발생
- -> 이용자 계정 정보 추적
- -> 계정 정보를 채팅 관리 시스템에 전송
- -> 해당 이용자의 채팅 시스템 차단

- -->> 각기 다른 적용 방법 (컨텐츠 차단, 이용자 차단, 로그 전송etc



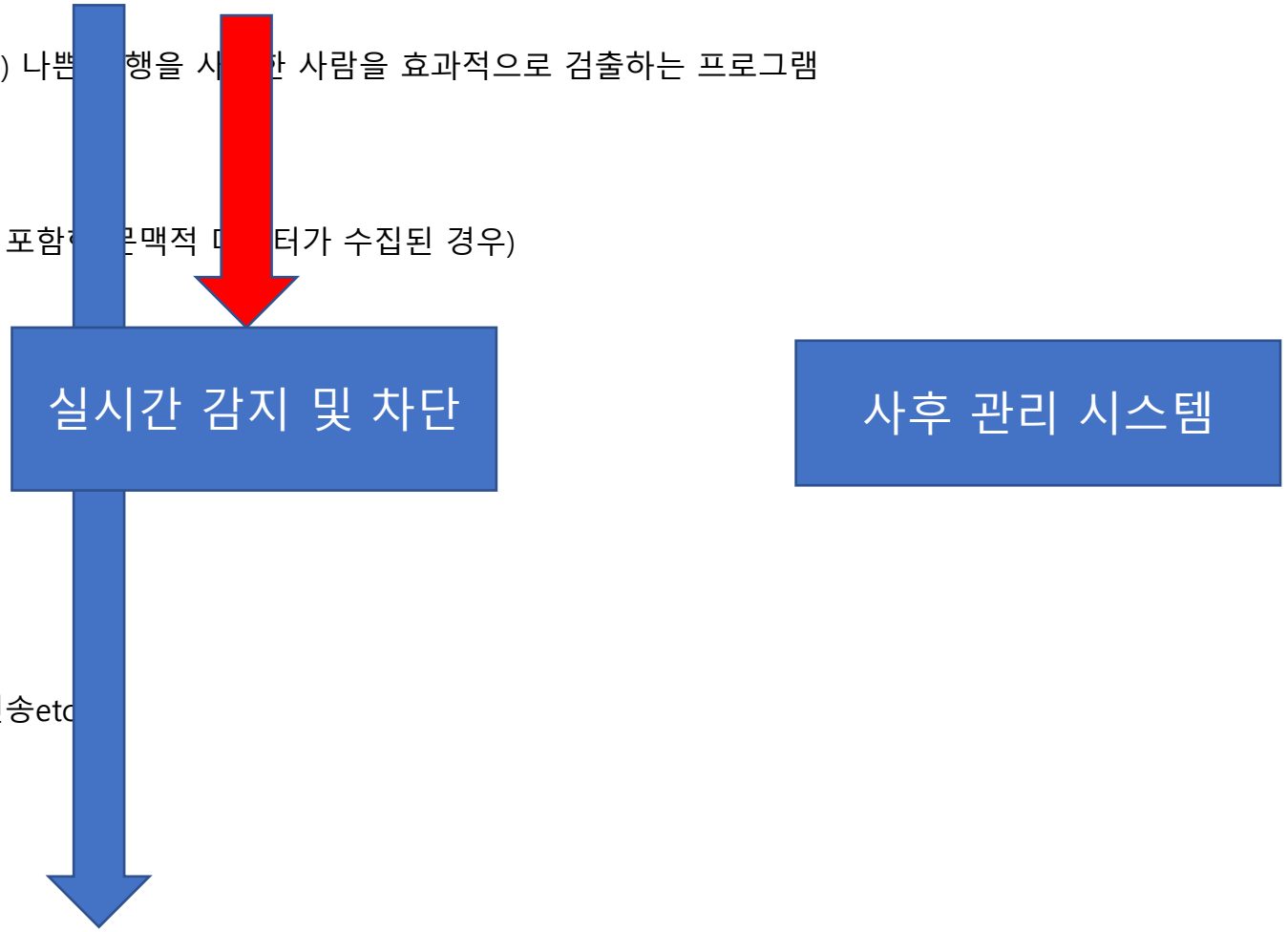
삽화 직접 그릴 예정

프로토타입

- 실시간으로 나쁜 언행을 막아주는 프로그램 vs (로그를 통해서) 나쁜 언행을 시[illegible]한 사람을 효과적으로 검출하는 프로그램

- 트래픽이 가장 많은 시점 혹은 채널의 채팅 로그 입력 (채팅을 포함[illegible]문맥적 [illegible]터가 수집된 경우)
- -> 데이터 취합하여 학습
- -> 서비스 내 적용
- -> 나쁜 언어를 사용하는 이용자 발생
- -> 이용자 계정 정보 추적
- -> 계정 정보를 채팅 관리 시스템에 전송
- -> 해당 이용자의 채팅 시스템 차단

- -->> 각기 다른 적용 방법 (컨텐츠 차단, 이용자 차단, 로그 전송etc



삽화 직접 그릴 예정

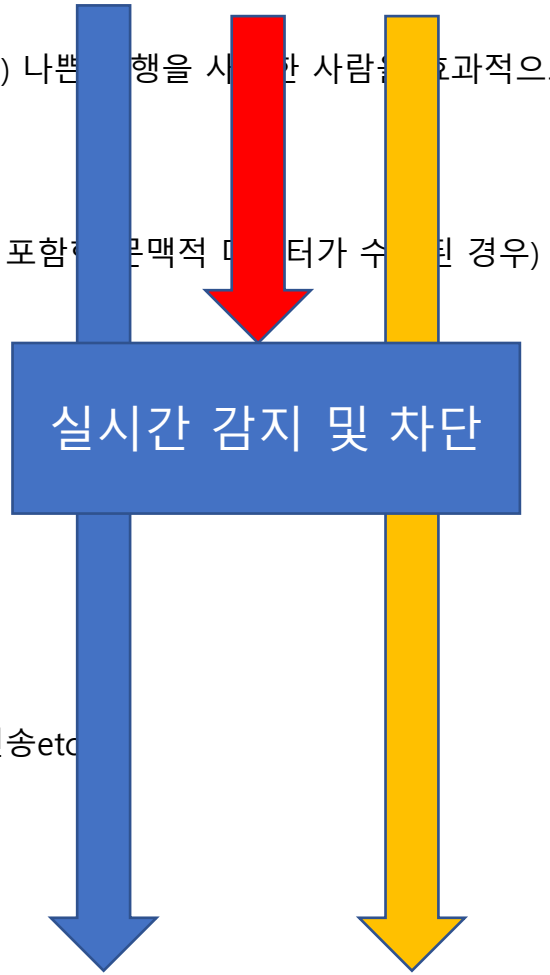
프로토타입

보이루 자이루

- 실시간으로 나쁜 언행을 막아주는 프로그램 vs (로그를 통해서) 나쁜 언행을 시 한 사람을 효과적으로 검출하는 프로그램

- 트래픽이 가장 많은 시점 혹은 채널의 채팅 로그 입력 (채팅을 포함 문맥적 데 터가 수 된 경우)
- -> 데이터 취합하여 학습
- -> 서비스 내 적용
- -> 나쁜 언어를 사용하는 이용자 발생
- -> 이용자 계정 정보 추적
- -> 계정 정보를 채팅 관리 시스템에 전송
- -> 해당 이용자의 채팅 시스템 차단

- -->> 각기 다른 적용 방법 (컨텐츠 차단, 이용자 차단, 로그 전송etc



사후 관리 시스템

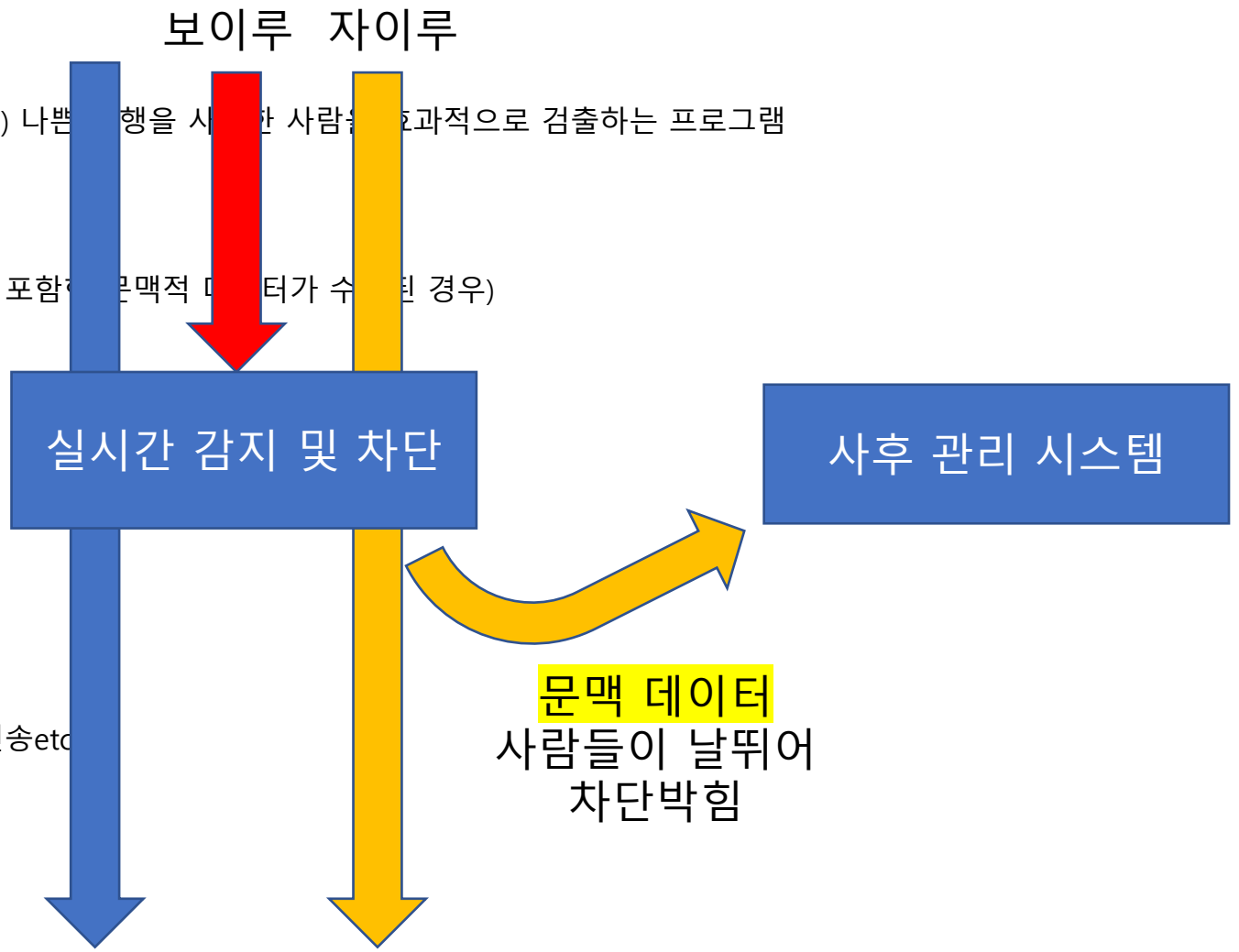
삽화 직접 그릴 예정

프로토타입

- 실시간으로 나쁜 언행을 막아주는 프로그램 vs (로그를 통해서) 나쁜 언행을 시[illegible]한 사람을 효과적으로 검출하는 프로그램

- 트래픽이 가장 많은 시점 혹은 채널의 채팅 로그 입력 (채팅을 포함[illegible]문맥적 데이터가 수[illegible]된 경우)
- -> 데이터 취합하여 학습
- -> 서비스 내 적용
- -> 나쁜 언어를 사용하는 이용자 발생
- -> 이용자 계정 정보 추적
- -> 계정 정보를 채팅 관리 시스템에 전송
- -> 해당 이용자의 채팅 시스템 차단

- -->> 각기 다른 적용 방법 (컨텐츠 차단, 이용자 차단, 로그 전송etc



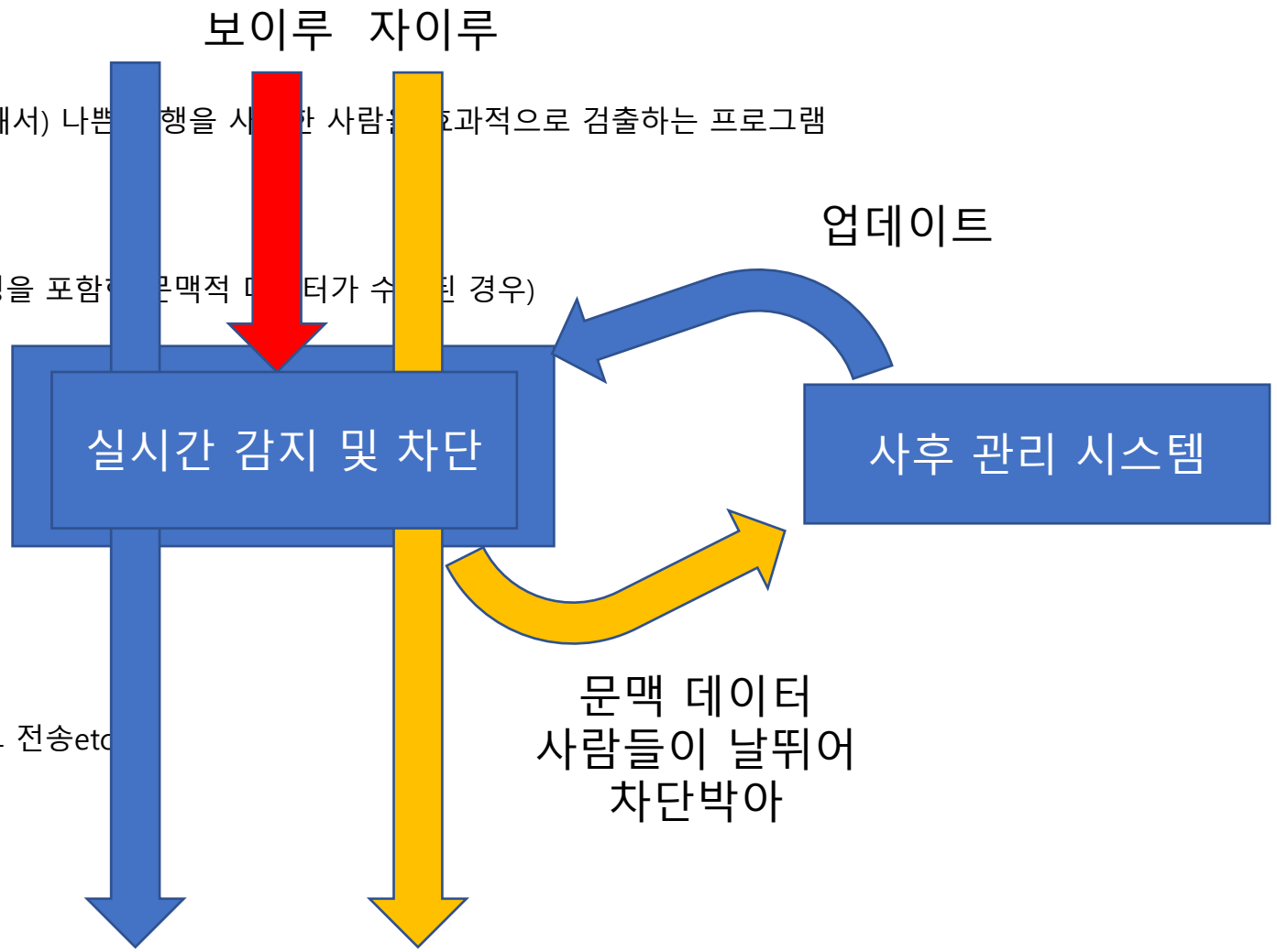
삽화 직접 그릴 예정

프로토타입

- 실시간으로 나쁜 언행을 막아주는 프로그램 vs (로그를 통해서) 나쁜 언행을 사후 한 사람을 효과적으로 검출하는 프로그램

- 트래픽이 가장 많은 시점 혹은 채널의 채팅 로그 입력 (채팅을 포함한 문맥적 데이터가 수집된 경우)
- -> 데이터 취합하여 학습
- -> 서비스 내 적용
- -> 나쁜 언어를 사용하는 이용자 발생
- -> 이용자 계정 정보 추적
- -> 계정 정보를 채팅 관리 시스템에 전송
- -> 해당 이용자의 채팅 시스템 차단

- -->> 각기 다른 적용 방법 (컨텐츠 차단, 이용자 차단, 로그 전송etc



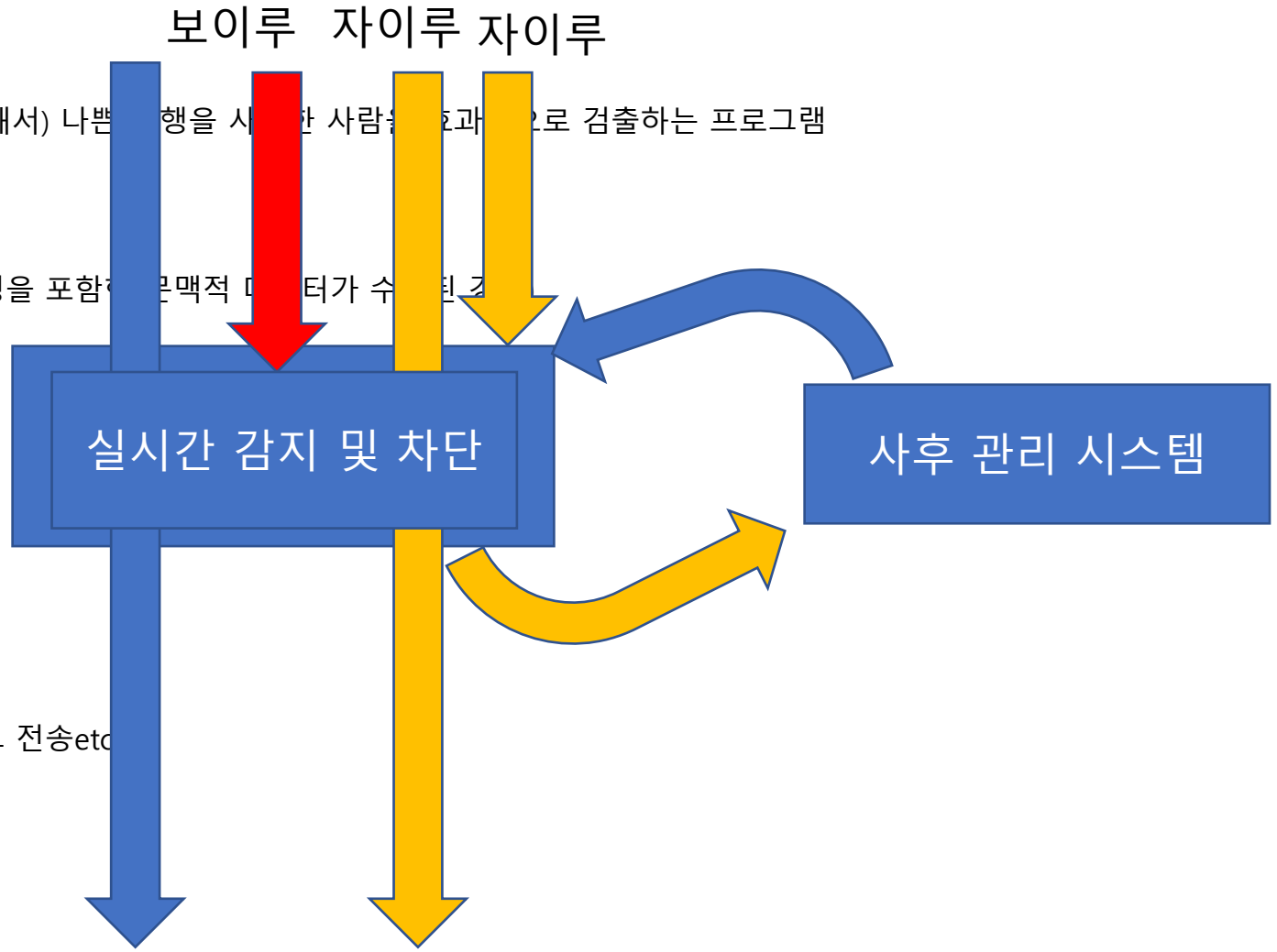
삽화 직접 그릴 예정

프로토타입

- 실시간으로 나쁜 언행을 막아주는 프로그램 vs (로그를 통해서) 나쁜 언행을 사후 관리한 사람을 효과적으로 검출하는 프로그램

- 트래픽이 가장 많은 시점 혹은 채널의 채팅 로그 입력 (채팅을 포함한 문맥적 데이터가 수집된 경우)
- -> 데이터 취합하여 학습
- -> 서비스 내 적용
- -> 나쁜 언어를 사용하는 이용자 발생
- -> 이용자 계정 정보 추적
- -> 계정 정보를 채팅 관리 시스템에 전송
- -> 해당 이용자의 채팅 시스템 차단

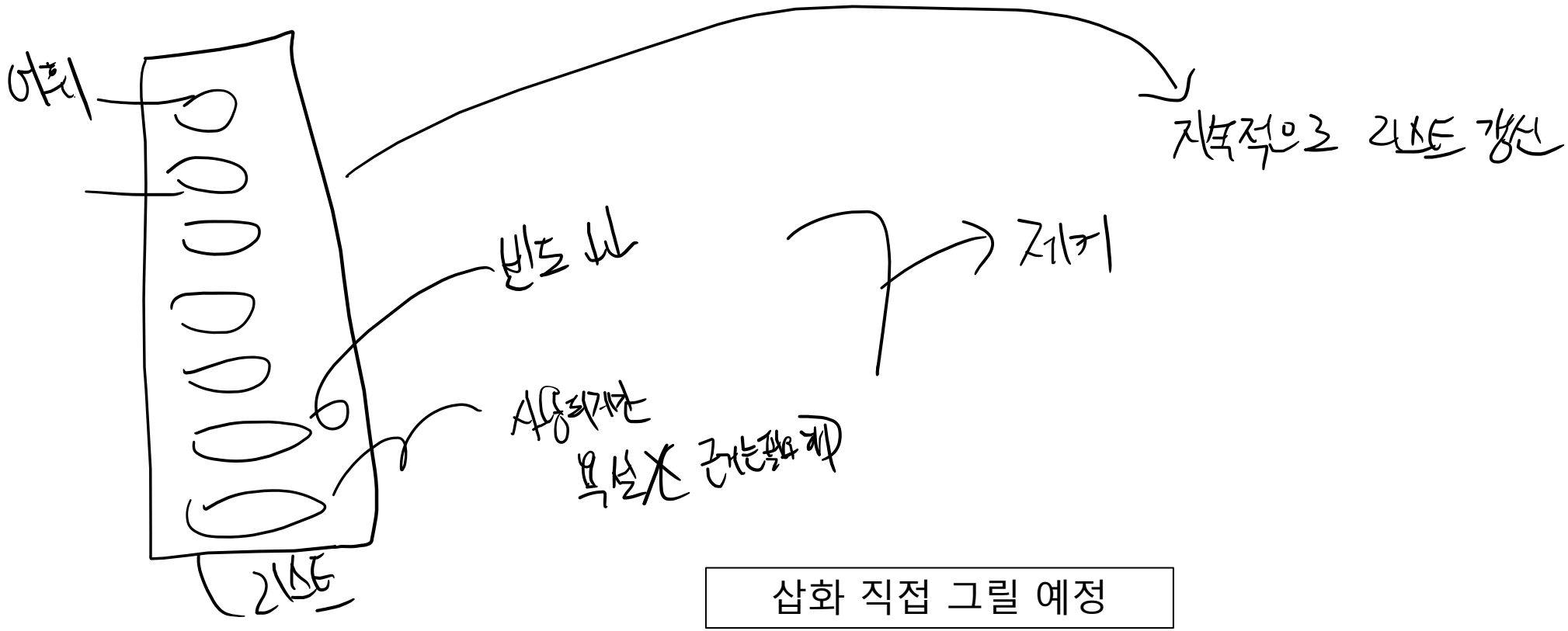
- -->> 각기 다른 적용 방법 (컨텐츠 차단, 이용자 차단, 로그 전송etc



삽화 직접 그릴 예정

운영 전략

- Case1 -> 차단, 신고, 등의 로그가 채팅 로그에 함께 출력되도록 프로그램 코드를 수정할 필요가 있다.
- Case2 -> 채팅 간격 (현행 시스템에서도 사용 가능(채팅 시간 이용))
- ~~Case3 -> 게임 내적인 데이터를 함께 이용하자. (어려워)~~



삽화 직접 그릴 예정

사업화 가능성

- 다른 서비스에 얹어서 사용할 수 있는 프로그램
- 이식성이 높은 프로그램(지향점)

- 키즈 프로그램
- 무균 돼지같은 무균 프로그램
- 무해한 시스템 환경 만들기

https://d2.naver.com/helloworld/7753273