영화 리뷰 분석 웹 서비스

텍스트마이닝을 활용한 영화 리뷰 분석

170269 김남수

172100 한규정

154008 천승민

목차

1. 프로젝트 주제 및 목표

1. 프로젝트 주제 및 목표

1. 기존 사용자의 평점은 **주관적**인 문제점

2. 긍정적 홍보를 위한 댓수 **조작**의 문제점!

홍보 위해 '좋아요' 누른다 VS 경쟁작의 평점 깎는다

한 온라인 마케팅 및 모니터링 관계자 ㄱ씨는 "지난해 개봉했던 한 영화에 '언더 바이럴'(온라인에서 매크로 프로그램 등을 활용해 영화를 포털 사이트 상단에 올려놓거나 댓글이나

한줄평 | 총 1,606건

★☆☆☆☆ 2 이상한데... 평점이 쫑네. 왜지. 내가 이상한 사람인가

주에천사(peac****) | 2020.06.06 13:00 | 신고

635 280

★★★★★ 10 오랜만에 극장에서 영화봐서 그런지 몰입감이 쩔었습니다 송지효 목소리가 이렇게 소름돋는지 그전에는 몰랐어요

asdfgh(mike****) | 2020.06.04 12:49 | 신고

494 172

★★★★★ 10 송지효가 스릴러 연기를 이렇게 잘하는구나 느껴짐 .. 특유의 중저음 목소리랑 분위기가 뒷받침되서 더 그런듯

L라(mase****) | 2020.06.04 11:53 | 신고

365 118

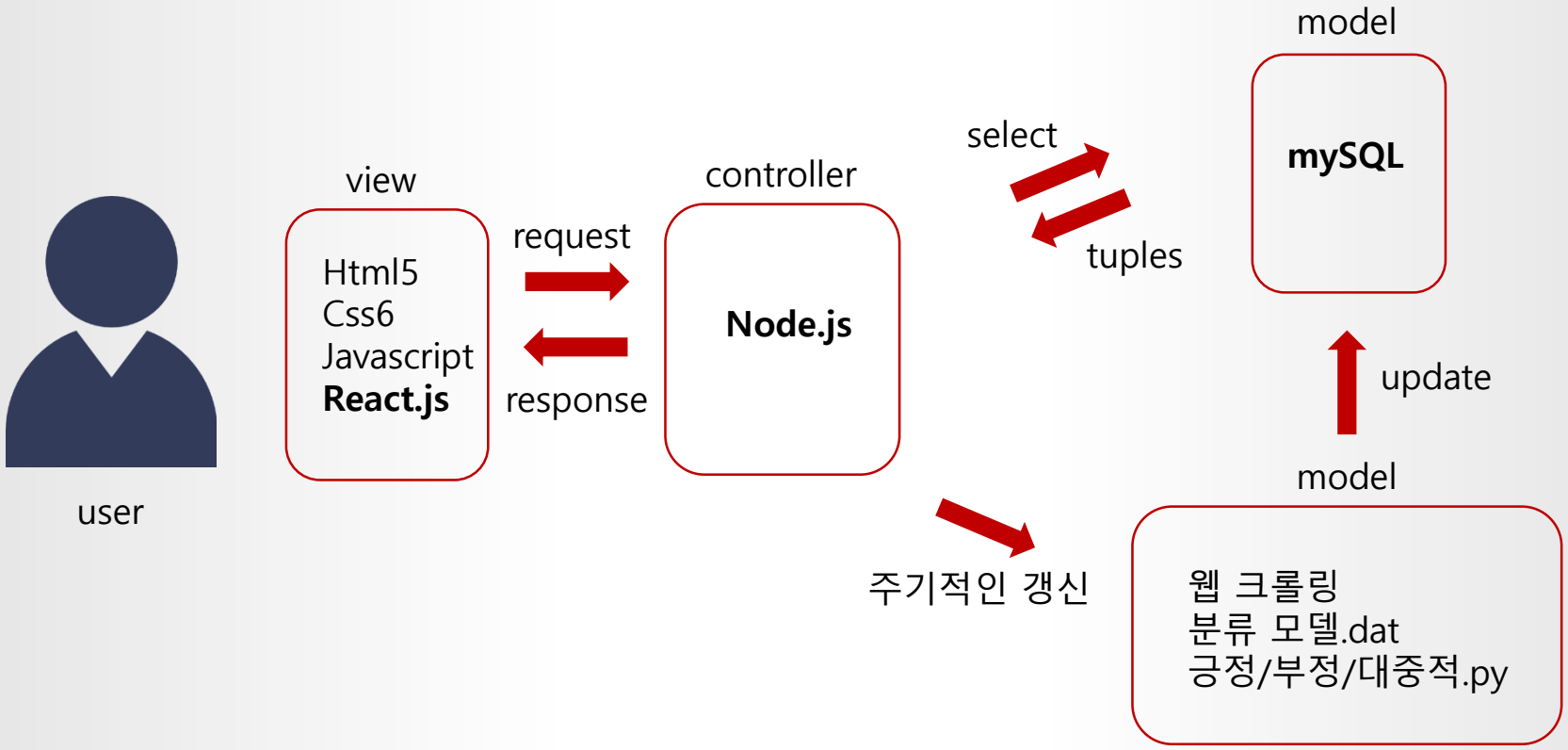
지 않는다"는 게 그의 설명이다.

<씨네 21 기사 중 일부>

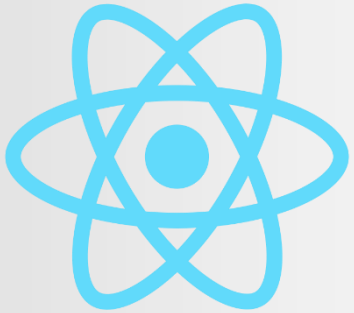
http://www.cine21.com/news/view/?idx=2&mag_id=90596

2. 프로젝트 내용

1. 시스템 아키텍처


model
mySQL
select
tuples
view
controller
request
response
Html5
Css6
Javascript
React.js
Node.js
update
user
model
주기적인 갱신
웹 크롤링
분류 모델.dat
긍정/부정/대중적.py

2. 사용 기술



React.js

Javascript API

가상 DOM tree를
메모리에 가지고 있어
사용자 반응에 대하여
기민하게 렌더링 가능함.

node.js

Javascript framework

Javascript로 서버를
구현할 수 있게 해줌.
포트포워딩을 통해 로컬
컴퓨터에 서버를 설치함.

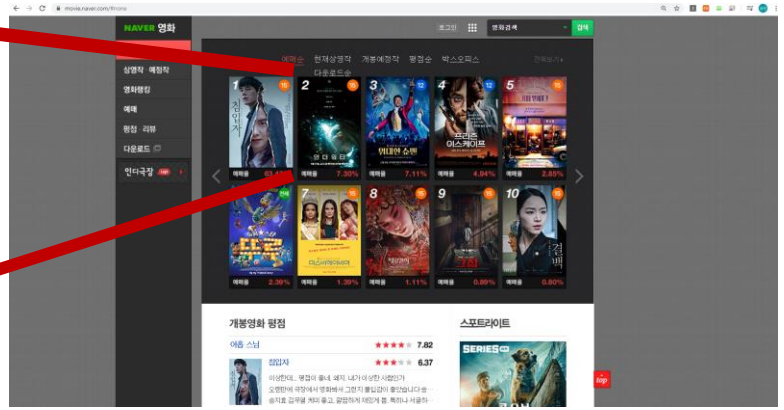
MySQL

관계형 모델.
스키마를 통해
정보들을 표현함.

1) 웹 크롤링

css selector

서버 ⇄ 네이버 영화 웹 페이지




```
selecter = await page.$("div.flick-container > ul#flick0 > li.item" + i + " > div.obj_off.tab4 > a > img")

movie_image = await page.evaluate(element => {

    return element.src

}, selecter);
```

2) 모델

긍정/부정 분류 모델의 경우

1) 데이터 수집
- 웹 크롤링
- 130,000개의 리뷰, 평점

2) 전처리
- 텍스트 처리
- 평점 처리 (3점 이하 부정, 8점 이상 긍정)

3) 형태소 분석

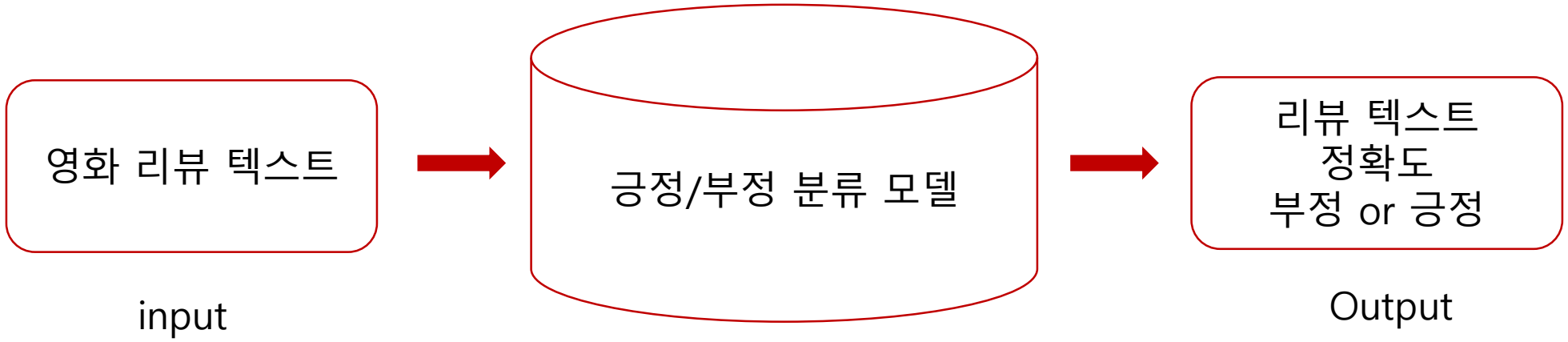
- KoNLPy 라이브러리
- Okt 형태소 분석기 사용
- 형태소 단위 토크나이징

4. 모델 생성

- TfidVectorizer
- Logistic Regression 모델
- 이진 분류 모델 생성

5) 결과

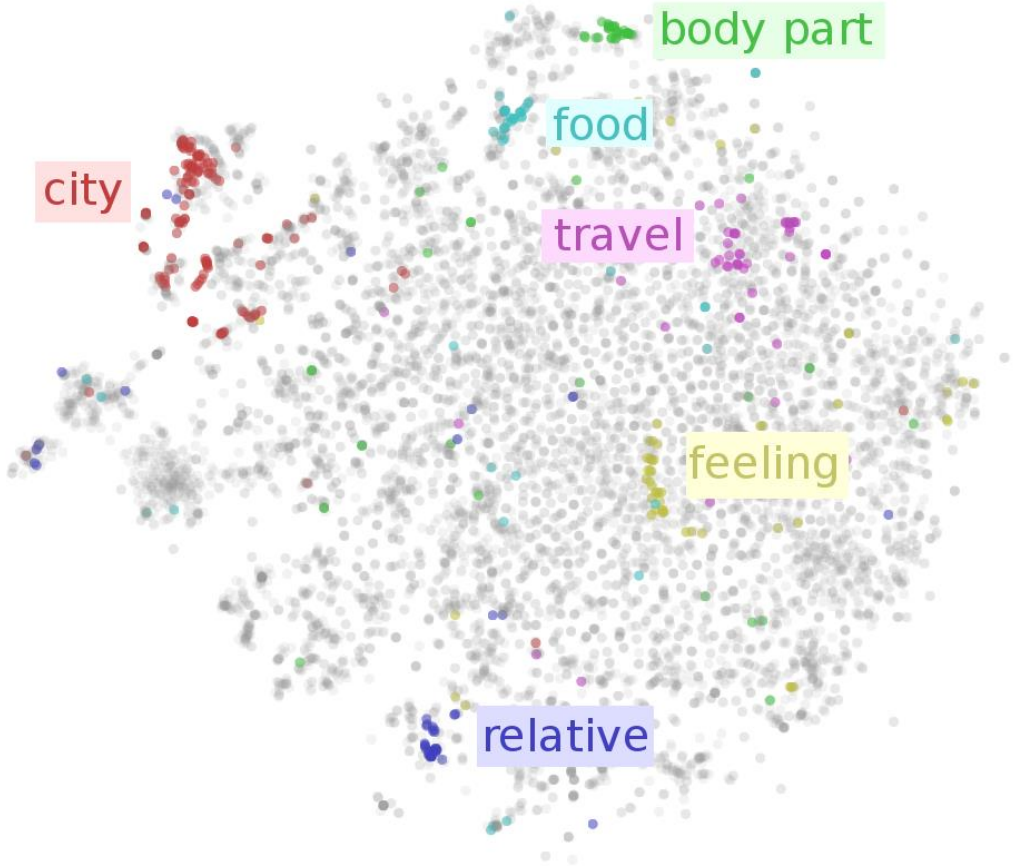
- 모델에 적용하여 긍정/부정 분류



많은 사람들이 공통적으로 느낀 리뷰

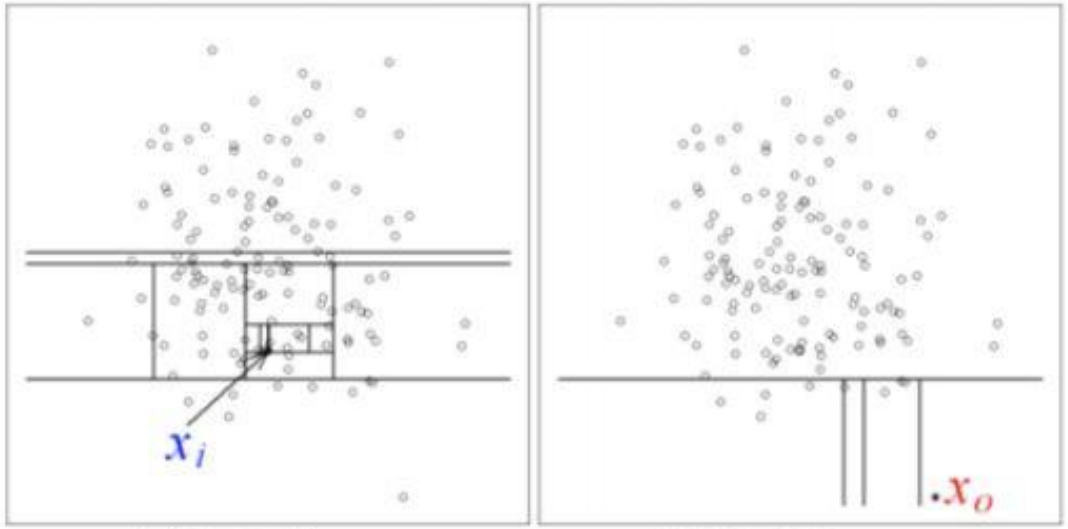
1) 워드 임베딩(Word Embedding)

ㅇ 단어를 벡터값으로 표현



- 다차원 상의 벡터값으로 표현
- 비슷한 의미를 가질수록 벡터상의 거리가 가까움
- 단어 사이의 유사도를 판단

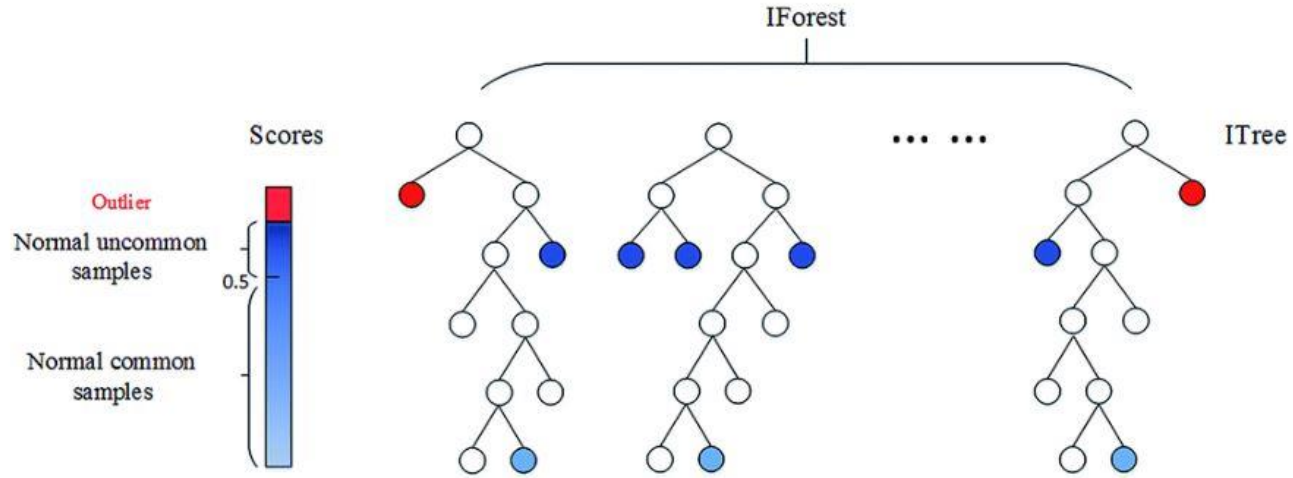
2) 모델 생성



(a) Isolating x_i (b) Isolating x_o

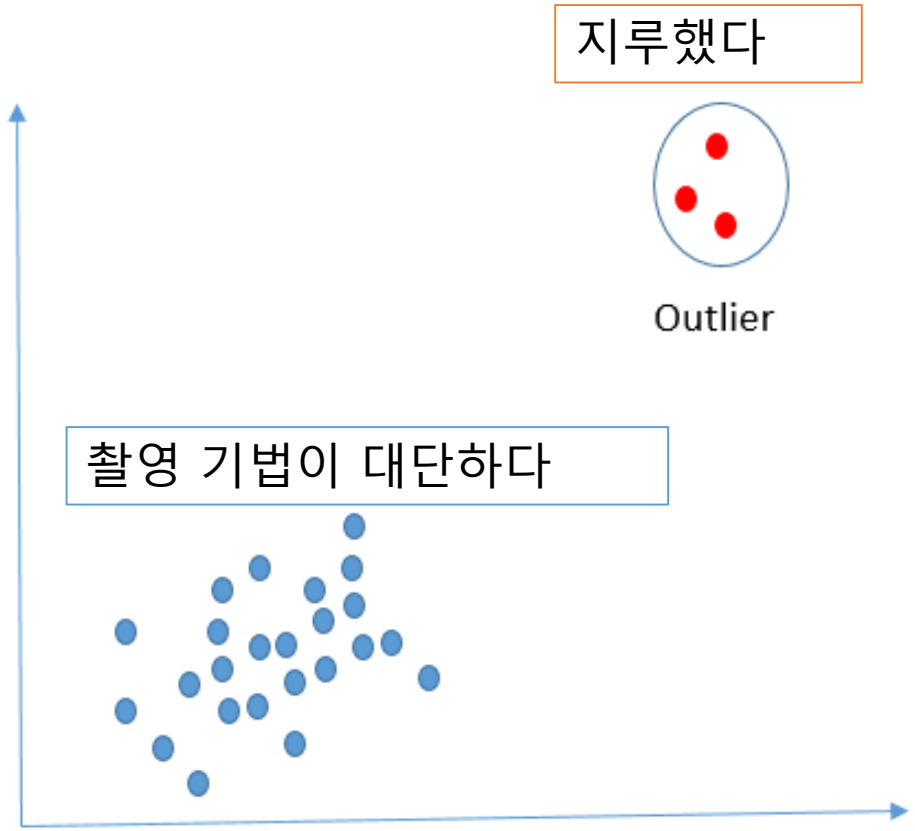
- Isolation Forest 모델
- Regression tree 기반의 split
- 모든 데이터 관측치를 고립시키는 방법

3) 이상치 점수(Outlier score)



- 특정 한 개체가 isolation 되는 leaf 노드(terminal node)까지의 거리를 outlier score로 정의

- 그 평균거리(depth)가 짧을 수록 outlier score는 높아짐



- 가장 중심에 있는 리뷰
  - Outlier score가 가장 낮은 리뷰

- Input Data
  - 100차원으로 임베딩된 리뷰

- Output Data
  - 리뷰의 Outlier Score

구분	기술	설명
S/W	웹 크롤링	- 네이버 영화 사이트에서 리뷰 & 평점을 크롤링.
	형태소 분석	- KoNLPy 형태소 분석기(Okt)
	Word Embedding	- 리뷰 내용을 처리하기 위해 기계학습 라이브러리를 활용 (Python) - Word Embedding (100차원) & Sentence Embedding
	Clustering(클러스터링)	- 기계학습 모델에 적용하여 벡터 사이 거리를 판단하기 위한 Outlier Score 분석

3. 기대효과

3. 기대효과

기존의 **주관적인 0-10 평점의 시스템**이

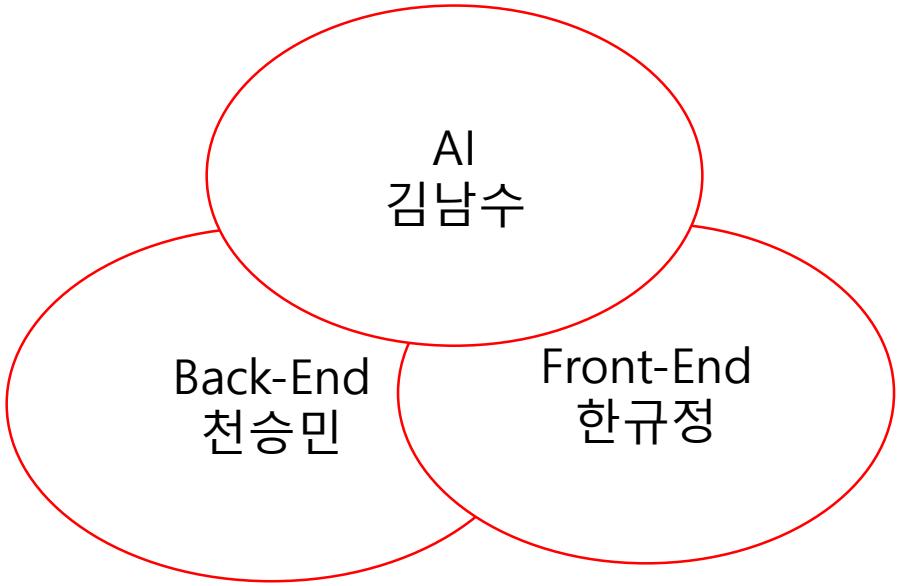
학습된 모델을 통하여
직관적인 긍정/부정/대중적인 리뷰 시스템으로 개선

이를 통해

사용자의 영화 예매 선택에 도움을 준다.

구분	추진내용	수행일정										
		4월 1주차	4월 2주차	4월 2주차	4월 3주차	4월 4주차	5월 1주차	5월 2주차	5월 3주차	5월 4주차	6월 1주차	6월 10일
계획	문제파악, 계획서 작성, 역할분담											
분석 및 설계	필요기술 분석 후 개발환경 구축											
개발	서버 구축											
	get, post 요청 응답 구축											
	db와 서버 연동											
	index 페이지 만들기											
	메인 페이지 만들기											
	영화 소개 페이지만들기											
	웹 디자인											
	DB에서 영화 목록을 받아와 엘리멘트 생성											
	웹 크롤링											
	데이터 전처리											
	Tokenize											
	정규화											
	word embedding											
	sentence embedding											
	모델 적용											
	outlier score 분석											
	DB설계											
	DB구축											
	DB관리											
테스트	오류 수정											

이름	색
김남수	
한규정	
천승민	
모두	



Q & A