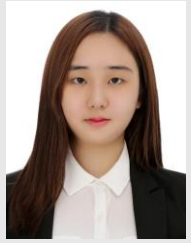
ELECTRA 기반 **영화평 감성분석**

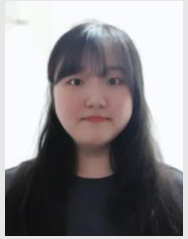
2nd implementation

<T6>

201611309 최지현



201611276 이규은



201612368 이지우

201611251 공민정

담당교수님 : 김학수 교수님

Contents

프로젝트 설명

데이터셋

시스템 테스트

추적성 분석

데모시나리오

프로젝트 설명

영화평 감정분석 시스템

학습된 데이터를 바탕으로 영화평을 긍/부정으로 감정분석하는 시스템

프로젝트 의의

1. 한국어 자연어처리

- 영어에 비해 문맥 파악과 토큰화 등이 어려움

=> 기존 영어 중심의 자연어처리 연구와 다른 접근 방식 요함.

2. 올해 초에 구글이 제안한 새로운 언어학습 모델인 ELECTRA를 사용

프로젝트 설명

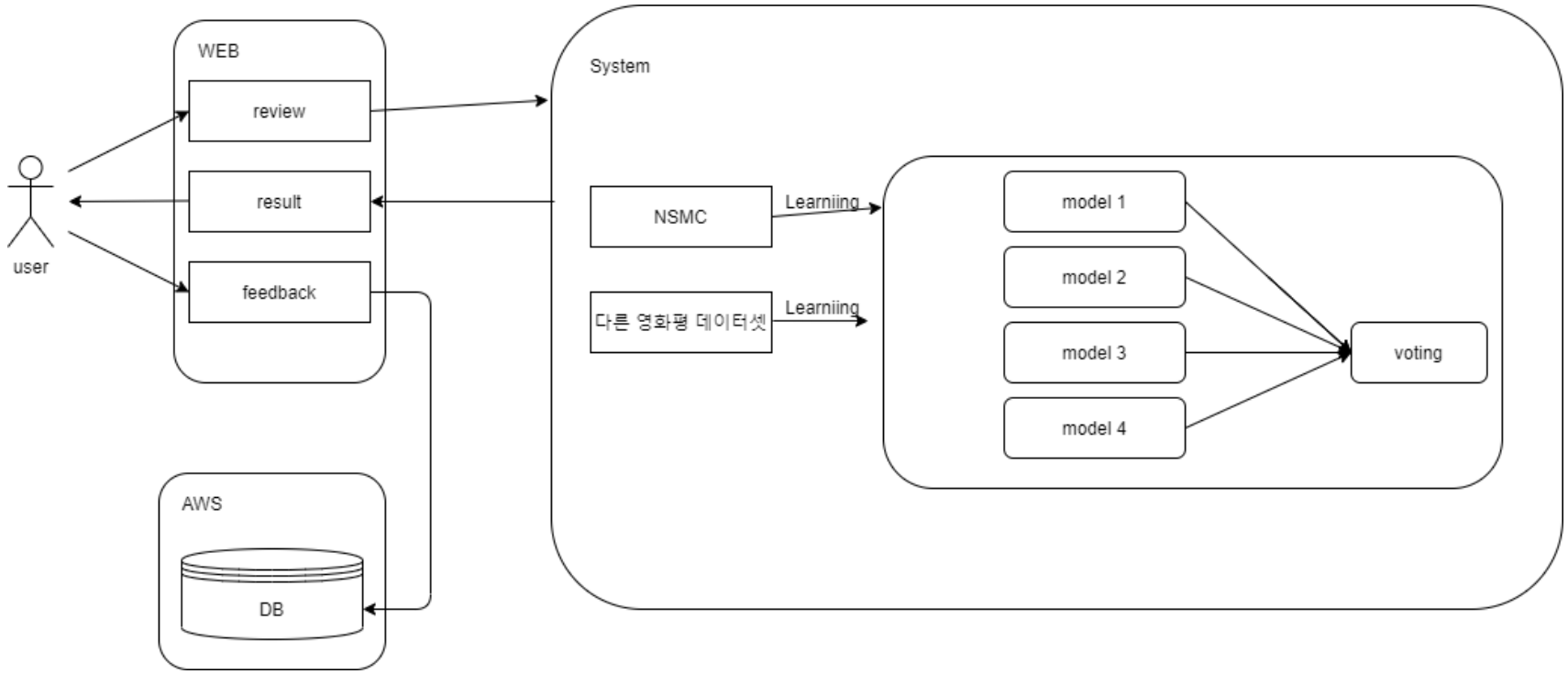
목적

단일 ELECTRA 모델 정확도인 **85% 보다 높은 정확도** 기대
(NSMC data로 Train 기준)

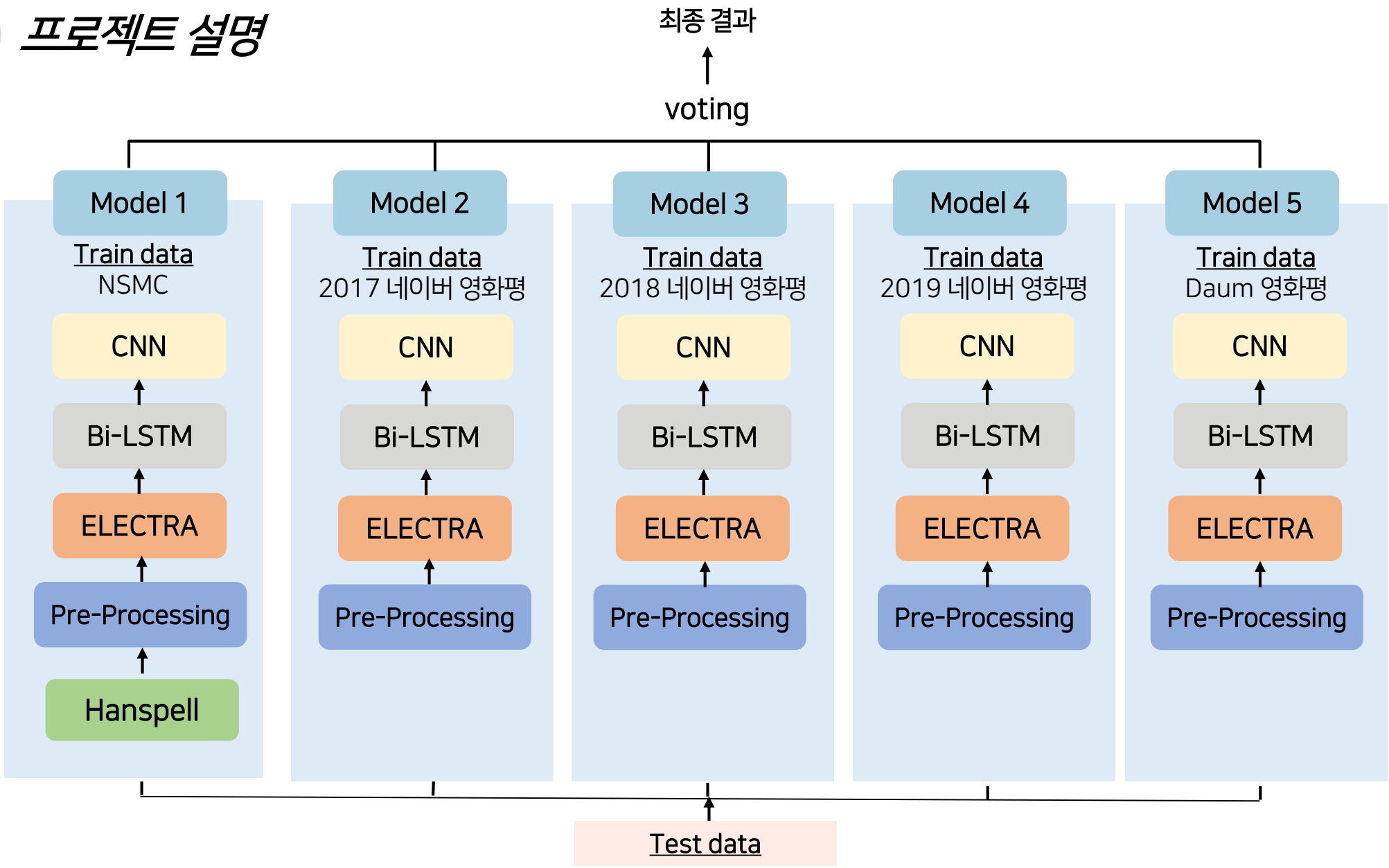
사람의 감정이 첨가된 단문을 감정분석할 수 있는 모델 기대

프로젝트 설명

시스템 구조



프로젝트 설명


최종 결과
voting
Model 1
Train data
NSMC
CNN
Bi-LSTM
ELECTRA
Pre-Processing
Hanspell
Model 2
Train data
2017 네이버 영화평
CNN
Bi-LSTM
ELECTRA
Pre-Processing
Model 3
Train data
2018 네이버 영화평
CNN
Bi-LSTM
ELECTRA
Pre-Processing
Model 4
Train data
2019 네이버 영화평
CNN
Bi-LSTM
ELECTRA
Pre-Processing
Model 5
Train data
Daum 영화평
CNN
Bi-LSTM
ELECTRA
Pre-Processing
Test data

프로젝트 설명

데이터셋

시스템 테스트

추적성 분석

데모시나리오

데이터셋

NSMC [1]

Naver Sentiment Movie Corpus의 준말로 github에 공개된 데이터셋

2016년 까지의 네이버 영화평

평점이 1~4점을 부정, 9~10을 긍정, 5~8은 제외함

Naver sentiment movie corpus v1.0

Characteristics

- All reviews are shorter than 140 characters
- Each sentiment class is sampled equally (i.e., random guess yields 50% accuracy)
  - 100K negative reviews (originally reviews of ratings 1-4)
  - 100K positive reviews (originally reviews of ratings 9-10)
  - Neutral reviews (originally reviews of ratings 5-8) are excluded

자체 구축 데이터셋

2017년 이후의 네이버 영화평 수집

평점이 1~4점을 부정, 8~10을 긍정으로 분류함

평점이 5~7점을 직접 긍/부정에 대한 라벨링을 진행함

생각보다 재미있었습니다신하균은 당연했고 디오도 연기잘해서 놀랐어요		1
정말 재미없습니다. 결말도 어이없고 흐지부지 끝나고 비추입니다.	부정적 →	0
	긍/부정 판단 불가능 → 삭제	
재밋게봤어용 ㅎㅎ시체가웃겼음...	긍정적 →	1
평점보고 봤는데 진짜 똥싸다가 덜싼기분저예산 영화도 아닌데 연출면에서 너무 아쉬움배우들 연기는 좋았지만 결국기억남는건 조선족 알바생 대사 중국사람입니다 이것뿐		0
재미없는거같아요 . 그냥 그래요 내영이해 안되요		0

1) https://github.com/e9t/nsmc

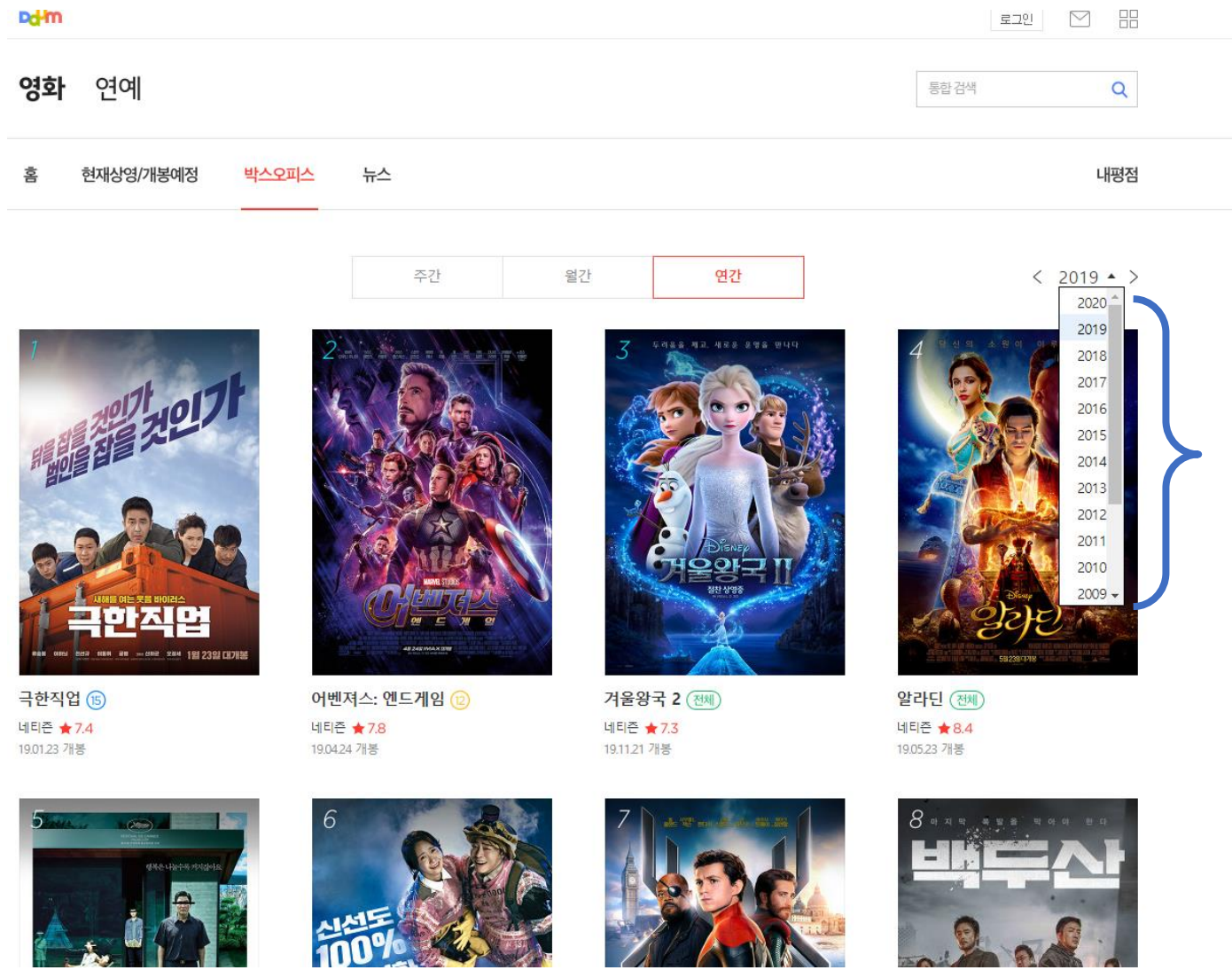
데이터셋

다음 영화평

2004~2019년 다음 영화평

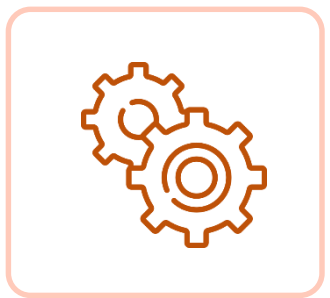
평점이 1~4점을 부정, 9~10을 긍정, 5~8은 제외함

(NSMC 기준 통일)



프로젝트 설명

데이터셋



시스템 테스트

추적성 분석

데모시나리오

System Test

No.	Test Case	입력상황	예상결과	P/F
1	remove email address	“재밌어요 aa@ex.com”	positive	P
2	trim jQuery Hash String	“재밌어요 jQuery2123”	positive	P
3	remove emoji not in unicode	“재밌어요 ☺☺”	positive	P
4	padding if data is shorter than designated length	“”	잘못된 입력	P
5	input right data through web	“너무 재밌어요”	positive	P
6	input blank data through web	“”	잘못된 입력	P
7	input too long data through web	“너무 재밌어서 어쩌구저쩌구,,,”(140자 넘김)	잘못된 입력	P
8	input meaningless data through web	“12345678909876”	잘못된 입력	P
9	return result through web	“재밌어요”	<positive> 결과가 웹을 통해 반환된다.	P
10	web feedback button	피드백 버튼을 통해 피드백을 보낸다.	피드백이 보내진다.	P
11	input extremely positive data	“너무너무 재밌어요”	positive	P
12	input extremely negative data	“이런 것도 영화라고 만들었냐? 감독 내려놓고 산에 들어가서 취직준비 해라”	negative	P

System Test

No.	Test Case	입력상황	예상결과	P/F
13	input neutral data	“재밌지만 두번볼 정도는 아니네요””	positive	P
14	importance of author sentiment than of others	“다른 사람들이 별로라고 해서 기대 안했는데 저는 재미있었어요”	positive	P
15	importance of present’s sentiment	“옛날엔 재미없었는데 얼마 전에 다시 보니 재미있네요”	positive	P
16	check if negated negative is positive	“노잼은 아니네요”	positive	P
17	check if negated of negative at the end is positive	“포스터만 보고 재미없을거라고 생각했는데 아니었어요”	positive	P
18	high accuracy	85% 이상의 정확도가 나온다	85% 이상의 정확도	P
19	short duration	결과가 5초 이내에 도출된다	결과가 5초 이내에 도출된다	P
20	additional learning	추가학습이 가능하다	추가학습이 가능하다.	P
21	model version control	모델 버전을 관리할 수 있다.	모델 버전을 관리할 수 있다.	P

※ Test Case 18 – Page7 Model1(NSMC 단일 모델) 정확도 89.07%

∴ 21/21 = 100%

프로젝트 설명

데이터셋

시스템 테스트

추적성 분석

데모시나리오

Traceability Matrix

color match
Requirement→

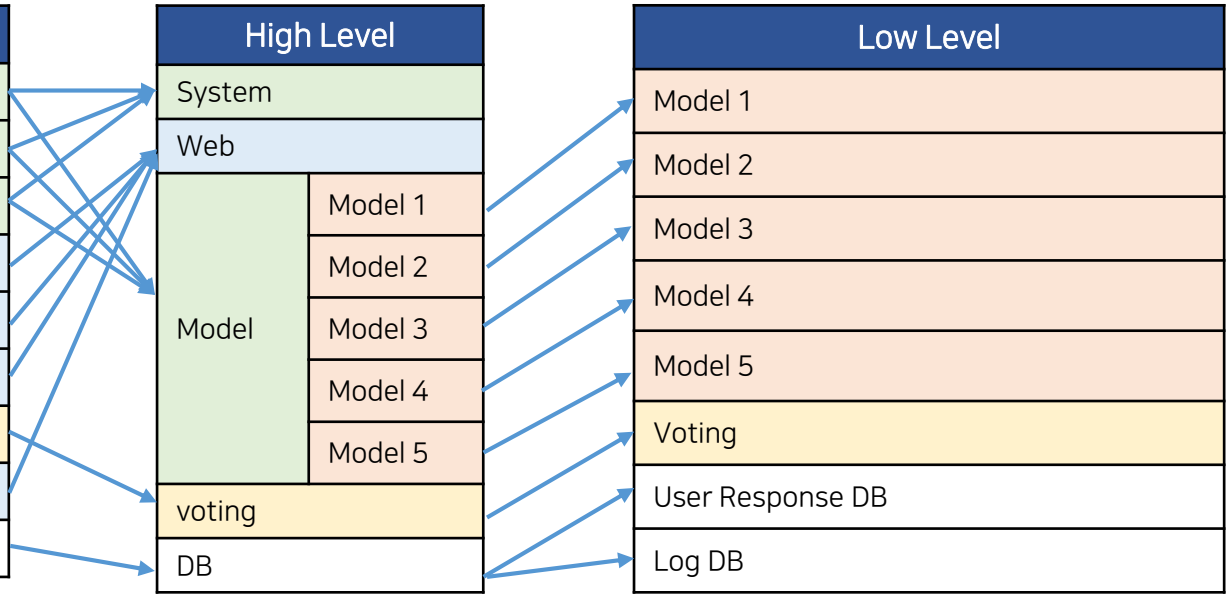
Requirement	R-No.
불필요한 데이터를 제거한다	FR 1.1
한글, 영어, 띄어쓰기, 일부 특수 문자를 제외하고 모두 제거한다.	FR 1.2
데이터가 같은 길이가 되도록 한다.	FR 1.3
사용자가 영화평을 입력한다.	FR 3.1
사용자가 입력한 영화평에 대해 감정 분석 결과를 반환한다.	FR 3.2
사용자가 피드백을 한다.	FR 3.3
기존 ELECTRA 단일 모델의 정확도보다 높은 정확도를 기대한다.	NFR 1.1
사용자가 대기할 수 있는 응답 시간을 기대한다.	NFR 2.1
DB를 통한 model quality 향상	NFR 3.1

R-No.	T-No.	Test Case
FR 1.1	1	remove email address
FR 1.1	2	trim jQuery Hash String
FR 1.2	3	remove emoji not in unicode
FR 1.3	4	padding if data is shorter than designated length
FR 3.1	5	input right data through web
FR 3.1	6	input black data through web
FR 3.1	7	input too long data through web
FR 3.1	8	input meaningless data through web
FR 3.2	9	return result through web
FR 3.3	10	web feedback button
NFR 1.1	11	input extremely positive data
NFR 1.1	12	input extremely negative data
NFR 1.1	13	input neutral data
NFR 1.1	14	importance of author sentiment than of others
NFR 1.1	15	importance of present's sentiment
NFR 1.1	16	check if negated negative is positive
NFR 1.1	17	check if negated of negative at the end is positive
NFR 1.1	18	high accuracy
NFR 2.1	19	short duration
NFR 3.1	20	additional learning
NFR 3.1	21	model version control

T-No.	Success Criteria
1, 2, 3, 4	한 영화평의 85% 이상이 어절별로 나누어진다.
1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17	긍/부정 값이 올바르게 도출된다.
	5초 이내에 결과가 도출된다.
18	긍/부정 값의 정확도가 85% 이상이다.
	감정 분석 결과의 정확도가 85% 이상이다.
	모델의 정확도가 85% 이상이다.
20, 21	추가 학습을 통해 Model의 정확도를 향상시킨다.
10, 20, 21	정확도가 향상된 Model을 DB에 저장한다.

Traceability Matrix

Requirement	No.
불필요한 데이터를 제거한다	FR 1.1
한글, 영어, 띄어쓰기, 일부 특수 문자를 제외하고 모두 제거한다.	FR 1.2
데이터가 같은 길이가 되도록 한다.	FR 1.3
사용자가 영화평을 입력한다.	FR 3.1
사용자가 입력한 영화평에 대해 감정 분석 결과를 반환한다.	FR 3.2
사용자가 피드백을 한다.	FR 3.3
기존 ELECTRA 단일 모델의 정확도보다 높은 정확도를 기대한다.	NFR 1.1
사용자가 대기할 수 있는 응답 시간을 기대한다.	NFR 2.1
DB를 통한 model quality 향상	NFR 3.1



프로젝트 설명

데이터셋

시스템 테스트

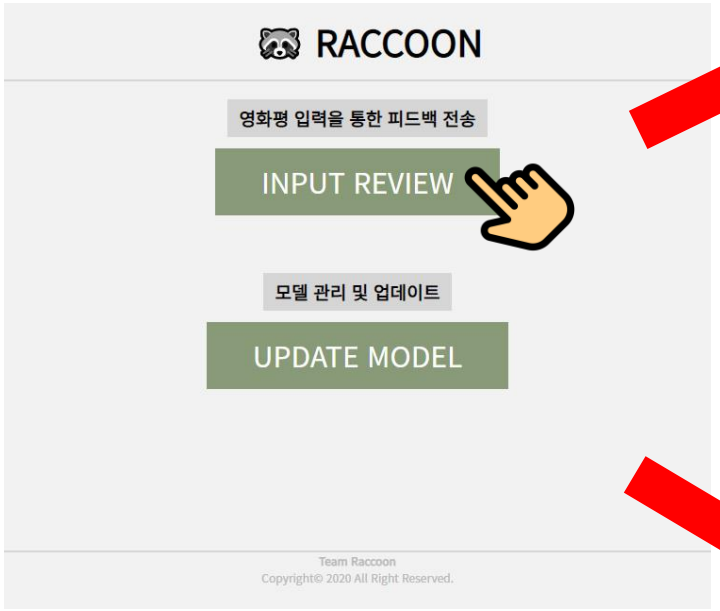
추적성 분석

데모시나리오

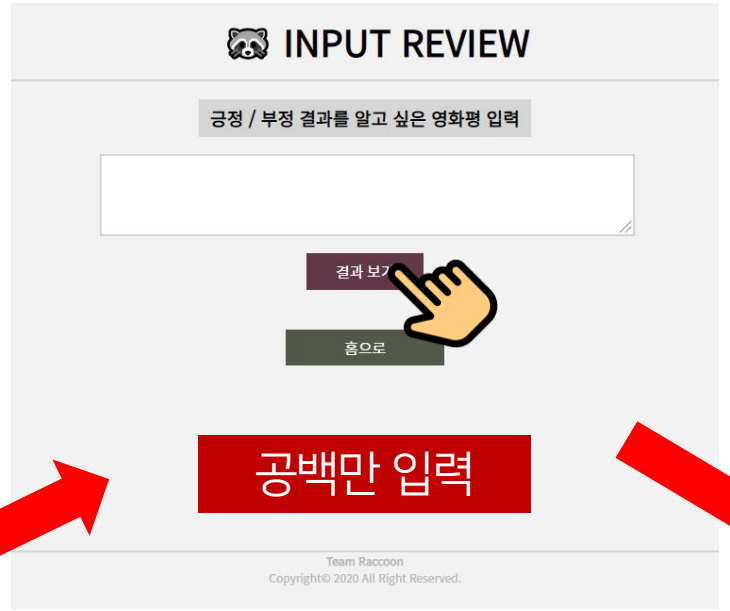
데모시나리오

* 자세한 내용은 시연영상을 참고 바랍니다.

사용자 - 영화평 입력

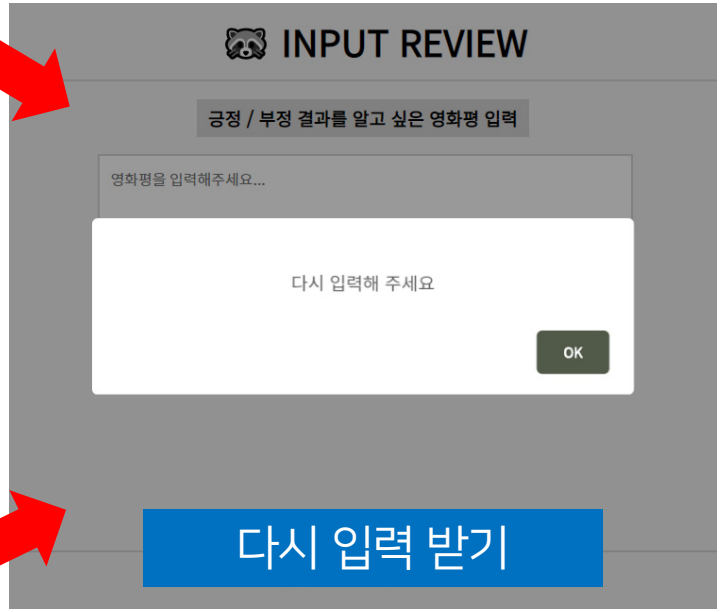


<웹 페이지 홈 화면>



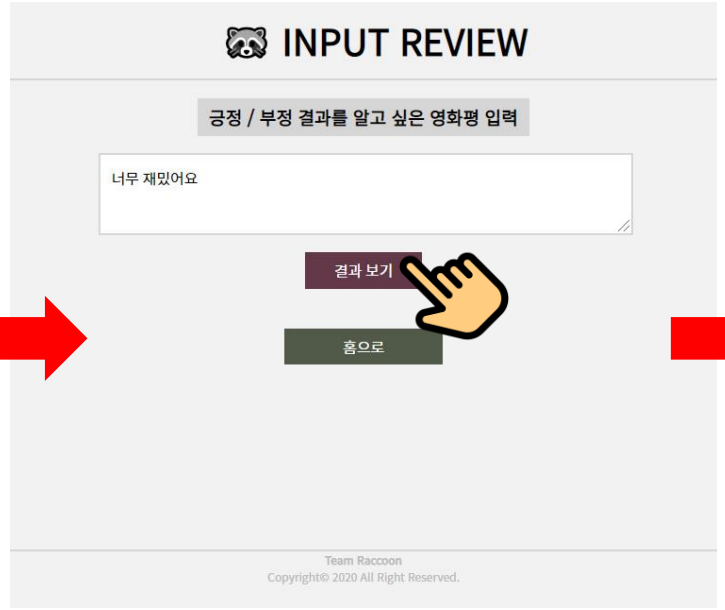
공백만 입력

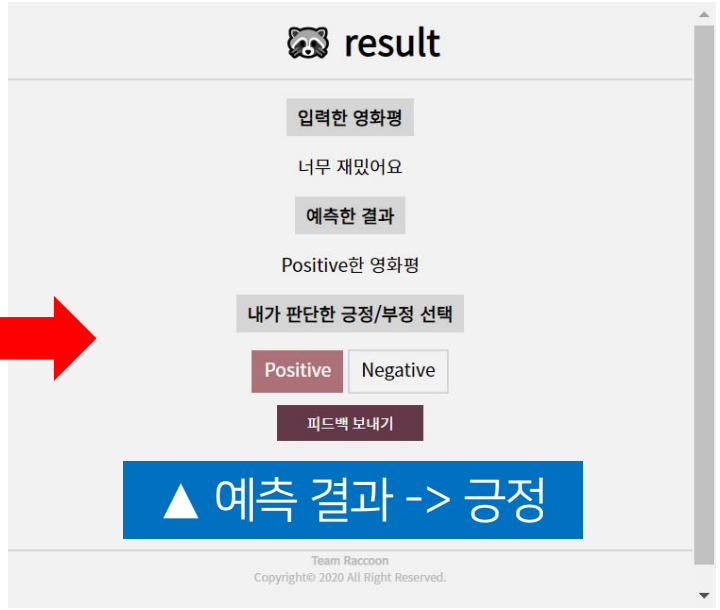
숫자만 입력



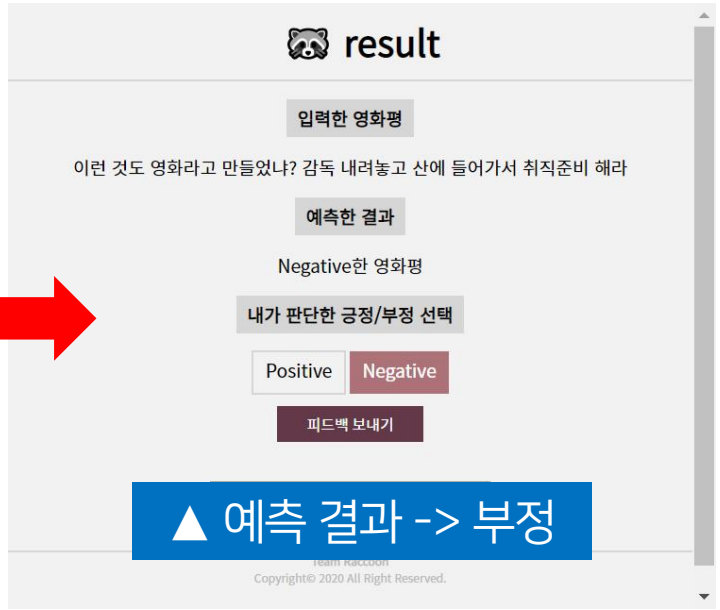
다시 입력 받기

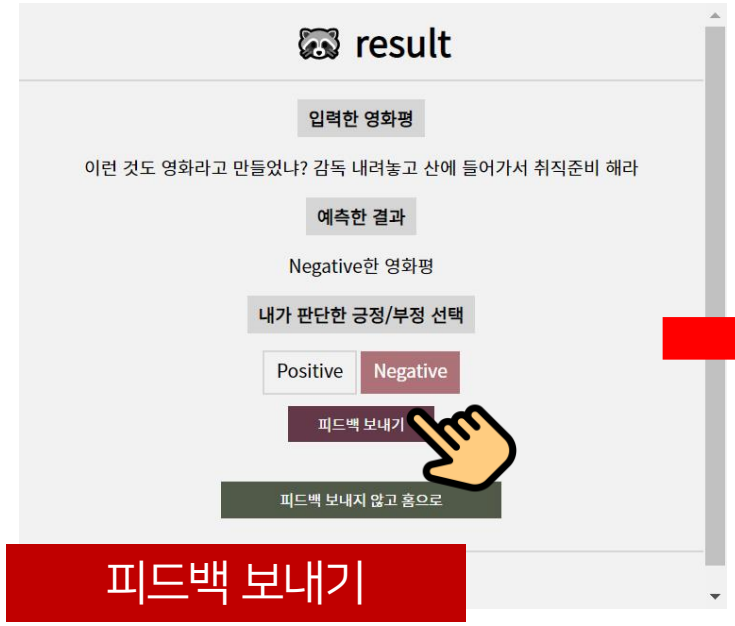
긍정적인 영화평 입력



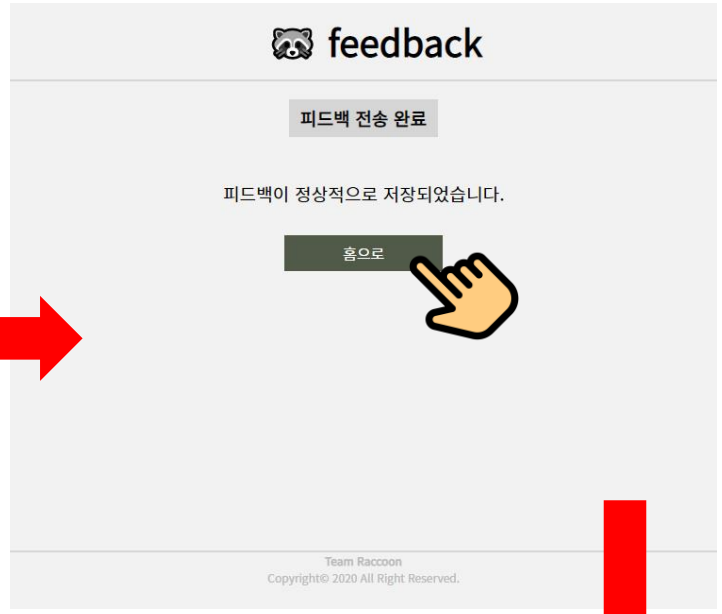


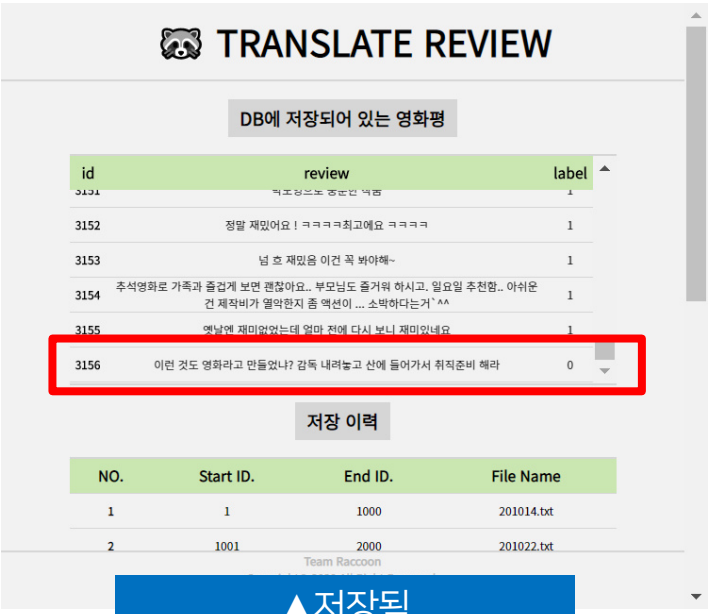
부정적인 영화평 입력

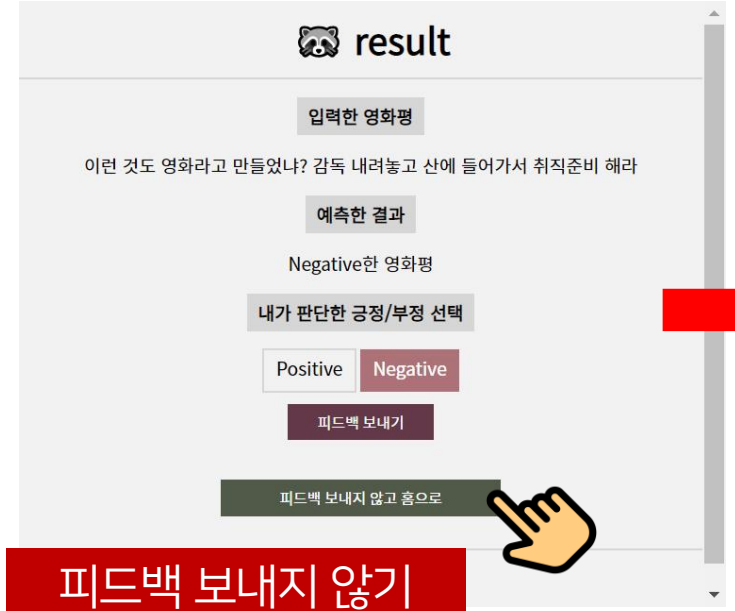




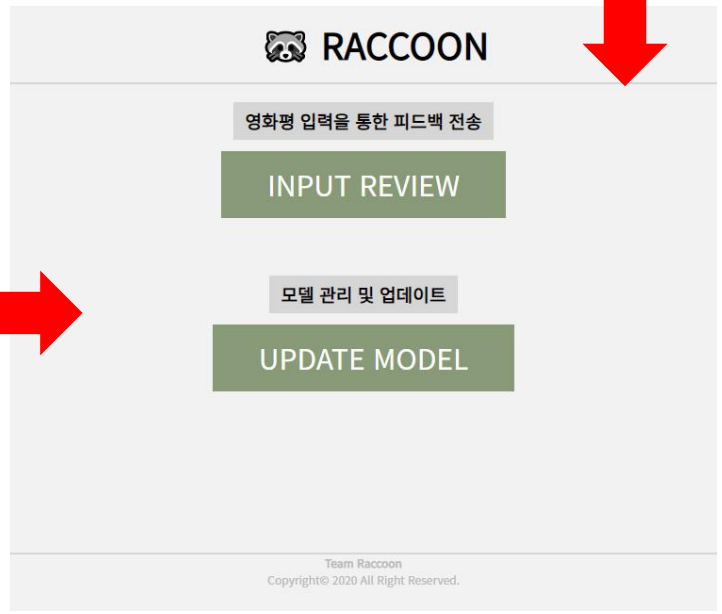
피드백 보내기



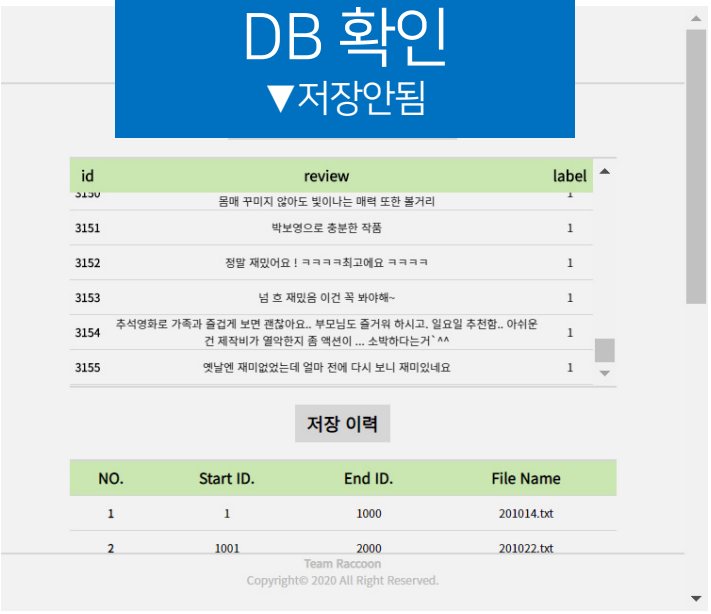




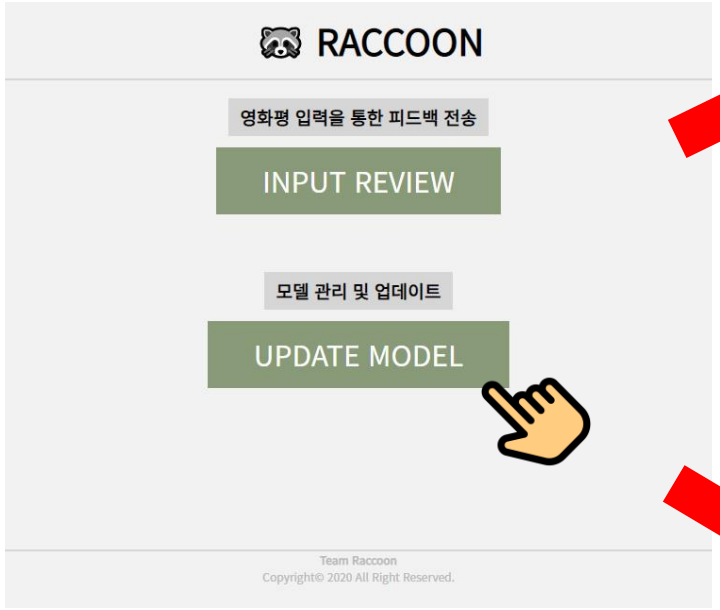
피드백 보내지 않기



▲저장됨
DB 확인
▼저장안됨



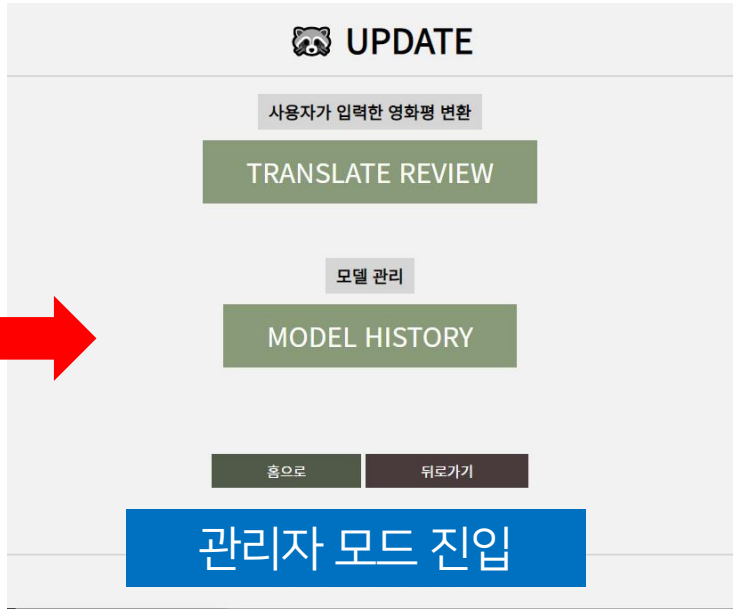
관리자 모드 진입



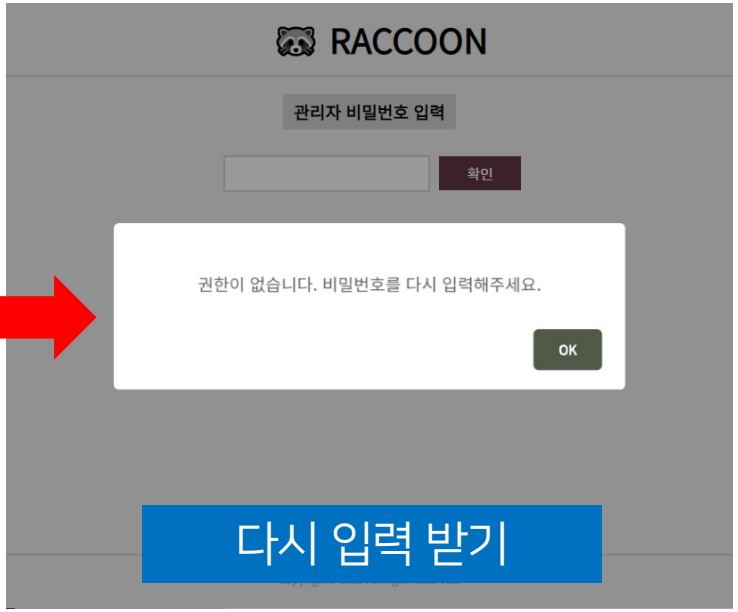
<웹 페이지 홈 화면>

올바른 비밀번호 입력

올바르지 않은 비밀번호 입력

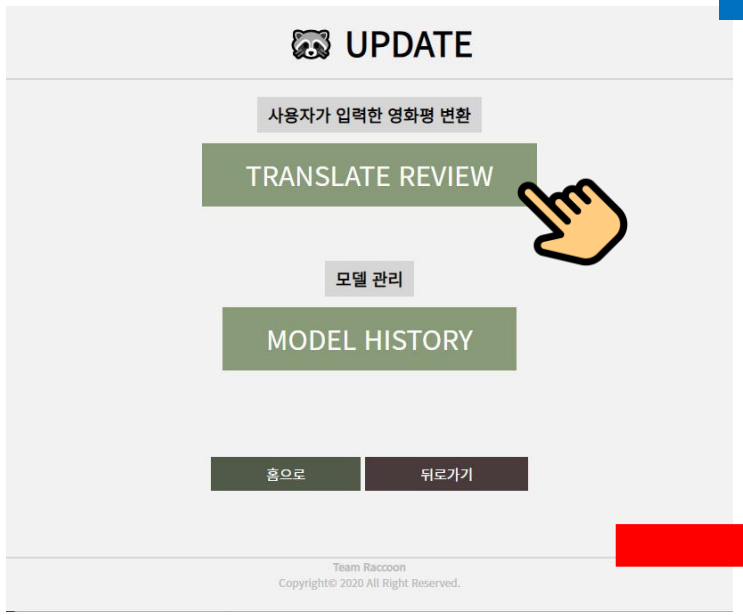


관리자 모드 진입



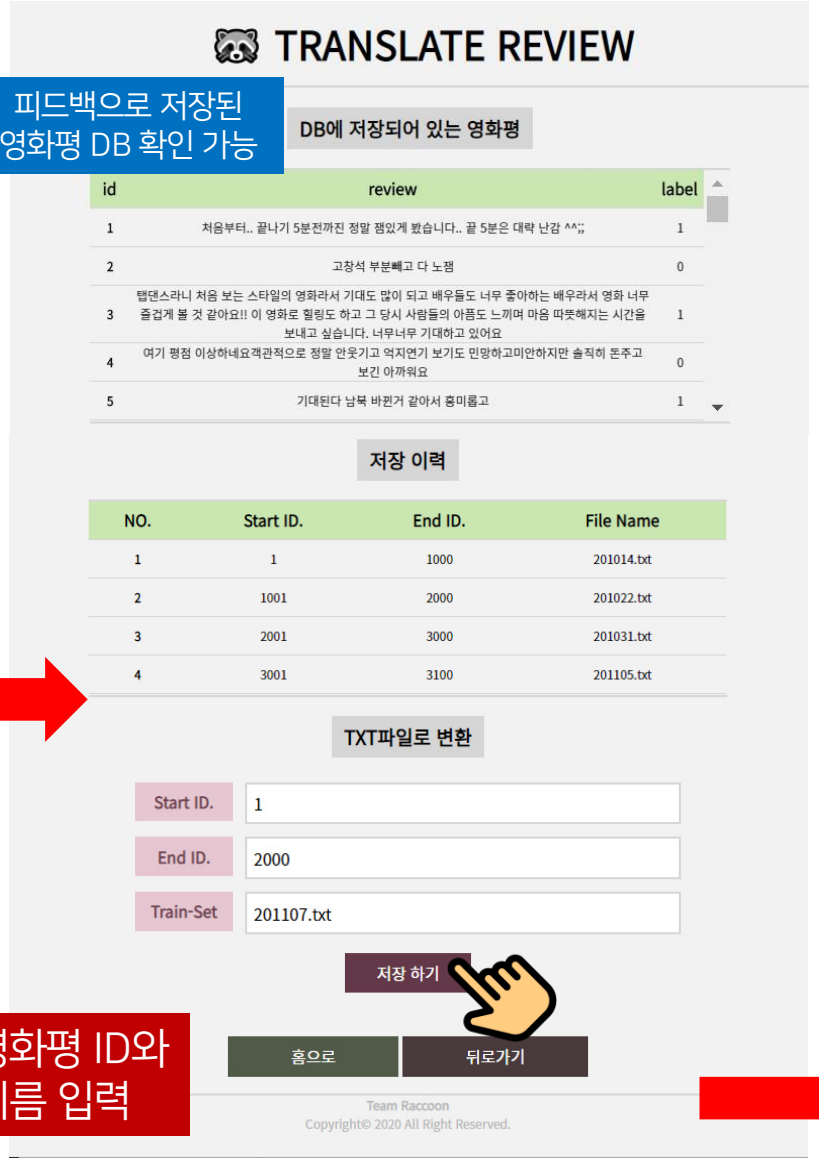
다시 입력 받기

관리자 모드 – 영화평 DB 확인 및 변환

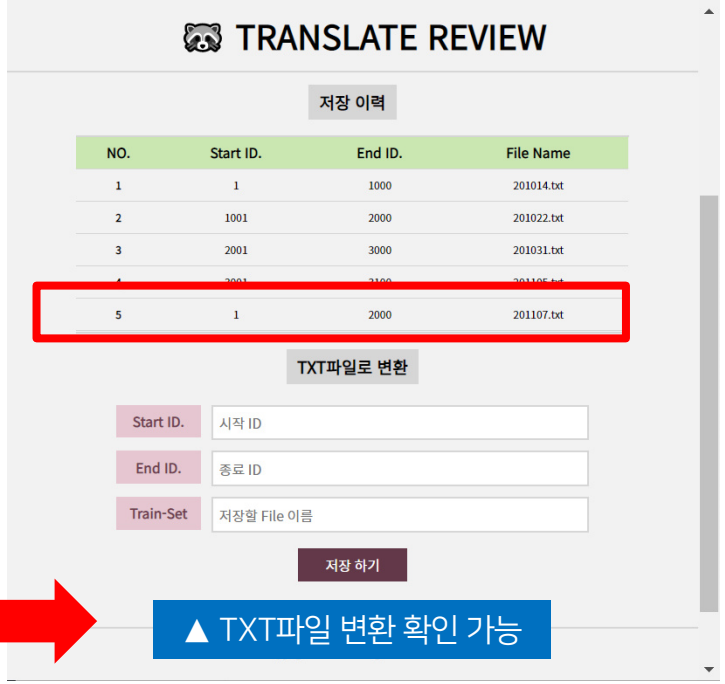


<관리자 페이지 화면>

피드백으로 저장된 영화평 DB 확인 가능

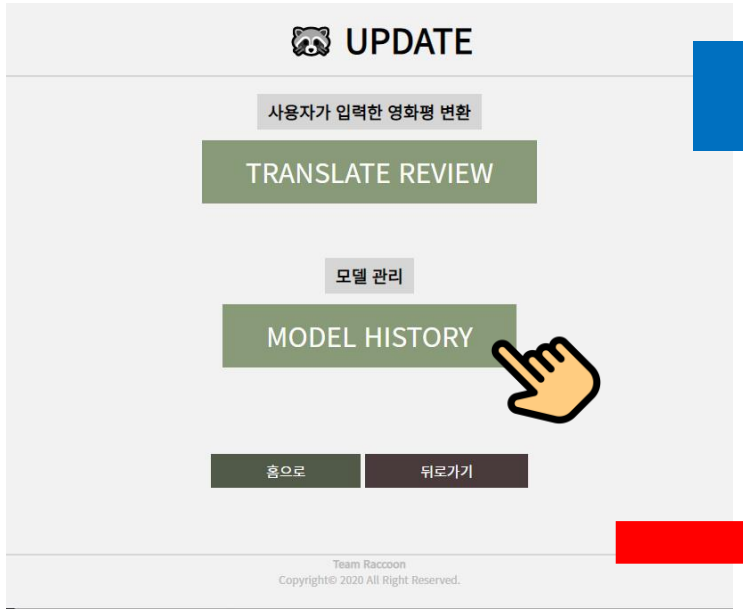


변환할 영화평 ID와 파일 이름 입력



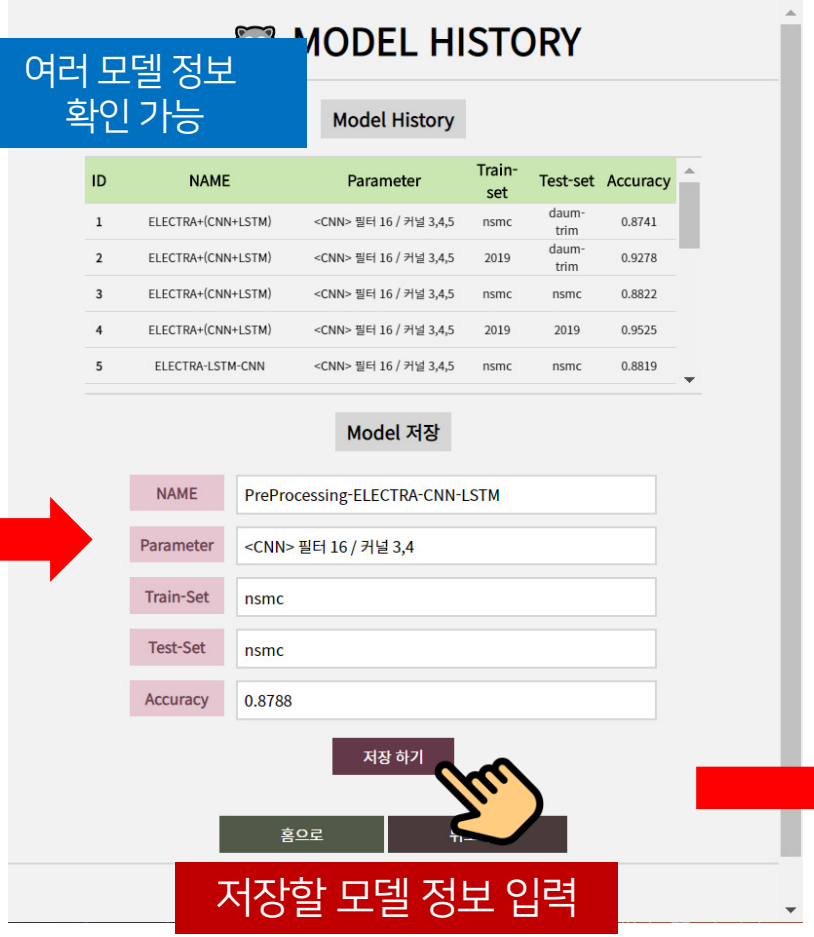
▲ TXT파일 변환 확인 가능

관리자 모드 – Model 관리

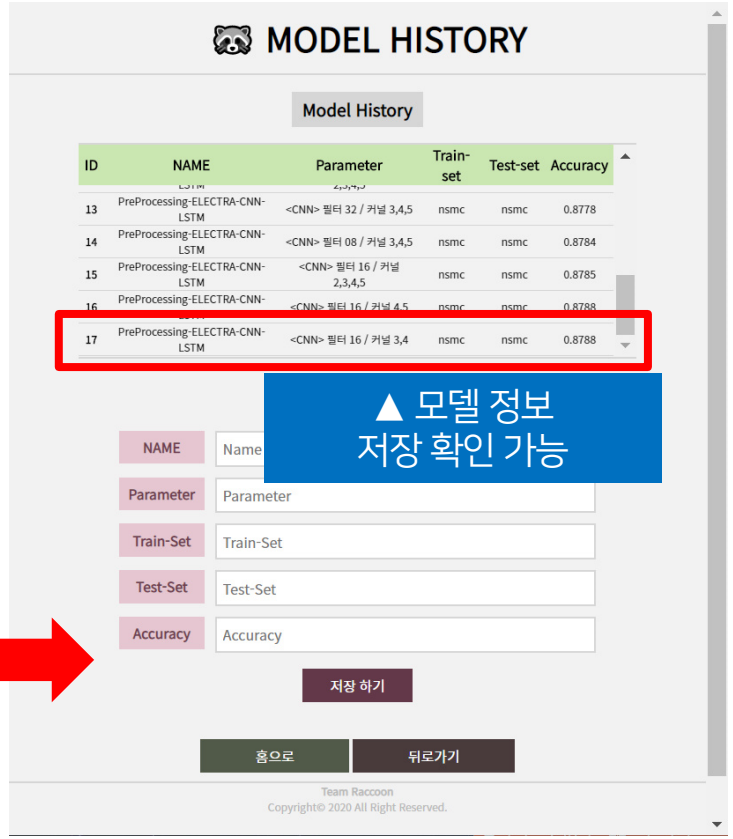


<관리자 페이지 화면>

여러 모델 정보 확인 가능



저장할 모델 정보 입력



▲ 모델 정보 저장 확인 가능

감사합니다!