

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



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

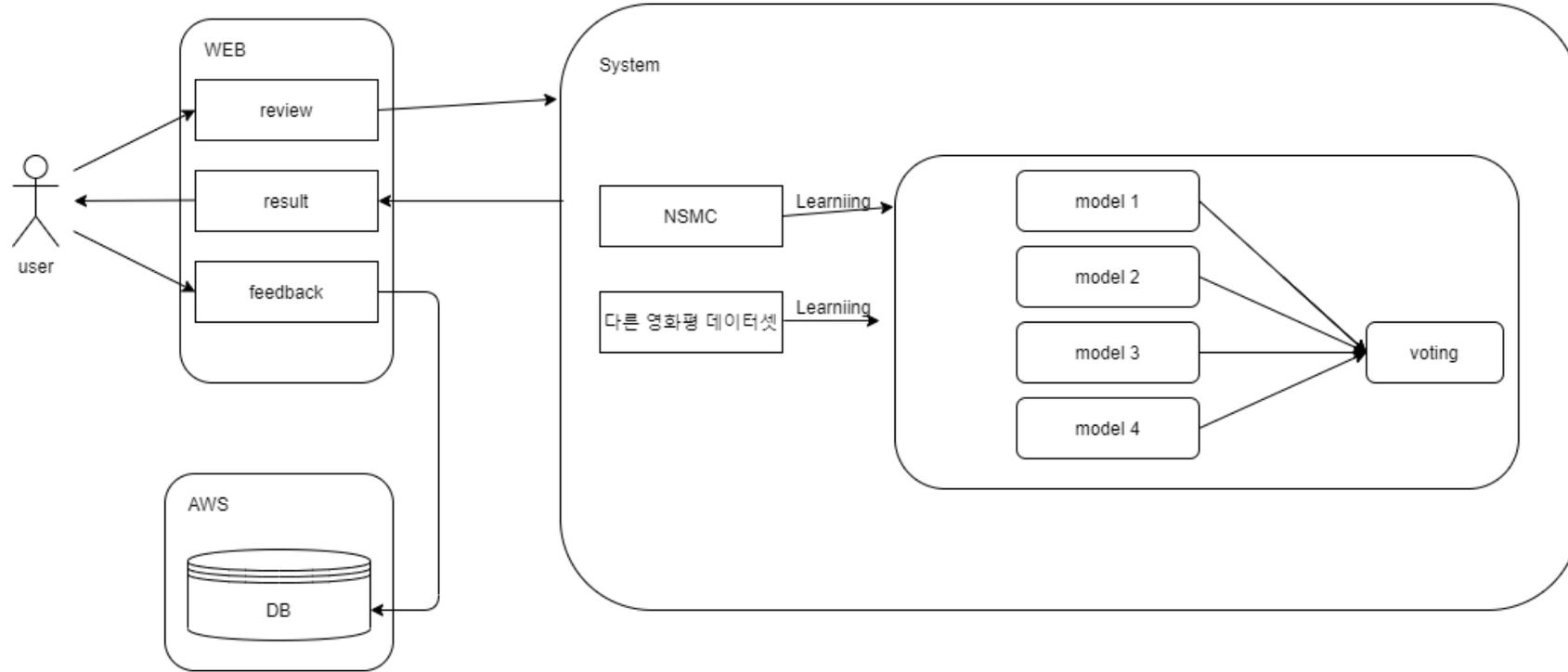
단일 ELECTRA 모델 정확도인 85% 보다 높은 정확도 기대
(NSMC data로 Train 기준)

사람의 감정이 첨가된 단문을 감정분석할 수 있는 모델 기대



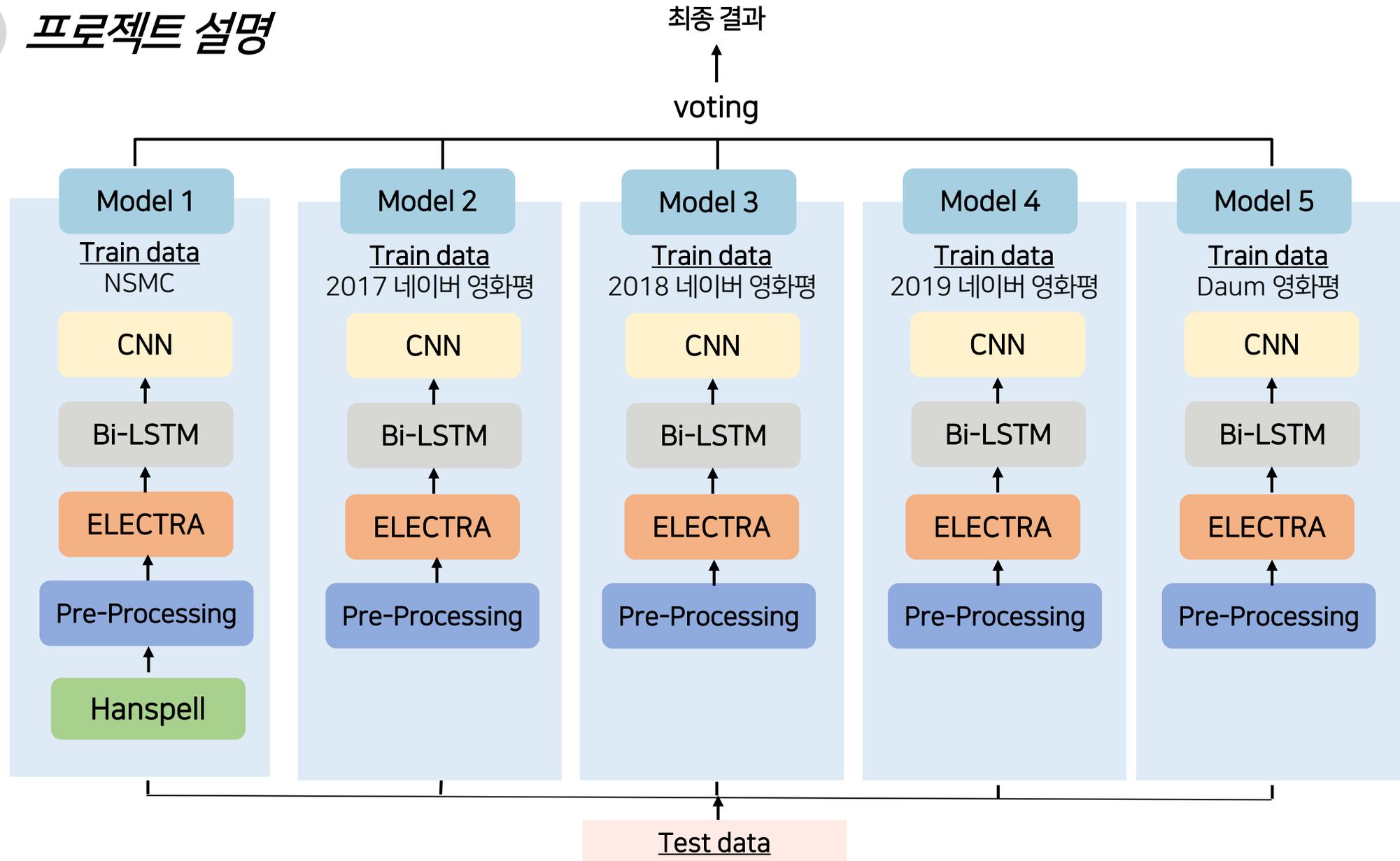
프로젝트 설명

시스템 구조





프로젝트 설명

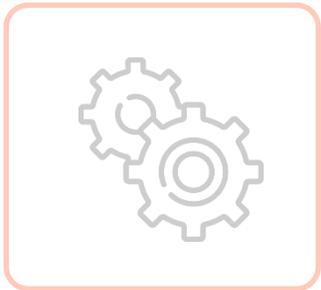




프로젝트 설명



데이터셋



시스템 테스트



추적성 분석



데모시나리오



NSMC 1

Naver Sentiment Movie Corpus의 준말로 github에 공개된 데이터셋

2016년 까지의 네이버 영화평

평점이 1~4점을 부정, 9~10을 긍정, 5~8은 제외함

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

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

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

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)



로그인



영화 연예

통합검색



홈

현재상영/개봉예정

박스오피스

뉴스

내평점

주간

월간

연간

< 2019 >

- 2020
- 2019
- 2018
- 2017
- 2016
- 2015
- 2014
- 2013
- 2012
- 2011
- 2010
- 2009



극한직업 15

네티즌 ★ 7.4

19.01.23 개봉



어벤져스: 엔드게임 12

네티즌 ★ 7.8

19.04.24 개봉



겨울왕국 2 전체

네티즌 ★ 7.3

19.11.21 개봉



알라딘 전체

네티즌 ★ 8.4

19.05.23 개봉

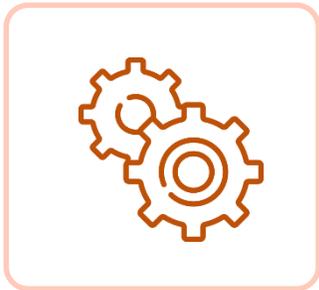




프로젝트 설명



데이터셋



시스템 테스트



추적성 분석



데모시나리오



System Test

No.	Test Case	입력상황	예상결과	P/F
1	remove email address	"재밋어요 aa@ex.com"	positive	P
2	trim jQuery Hash String	"재밋어요 jQuery2123"	positive	P
3	remove emoji not in unicode	"재밋어요 😊😊"	positive	P
4	padding if data is shorter than designated length	""	잘못된 입력	P
5	input right data through web	"너무 재밋어요"	positive	P
6	input blank data through web	""	잘못된 입력	P
7	input too long data through web	"너무 재밋어서 어찌구저찌구,,,"(140자 넘김)	잘못된 입력	P
8	input meaningless data through web	"12345678909876"	잘못된 입력	P
9	return result through web	"재밋어요"	<positive> 결과가 웹을 통해 반환된다.	P
10	web feedback button	피드백 버튼을 통해 피드백을 보낸다.	피드백이 보내진다.	P
11	input extremely positive data	"너무너무 재밋어요"	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	"재밋지만 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%

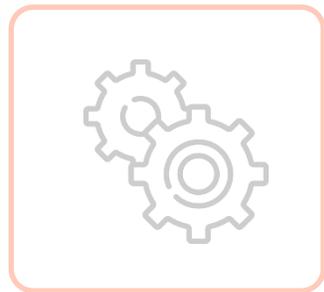
∴ 21/21 = 100%



프로젝트 설명



데이터셋



시스템 테스트



추적성 분석



데모시나리오



Traceability Matrix

color match Requirement→

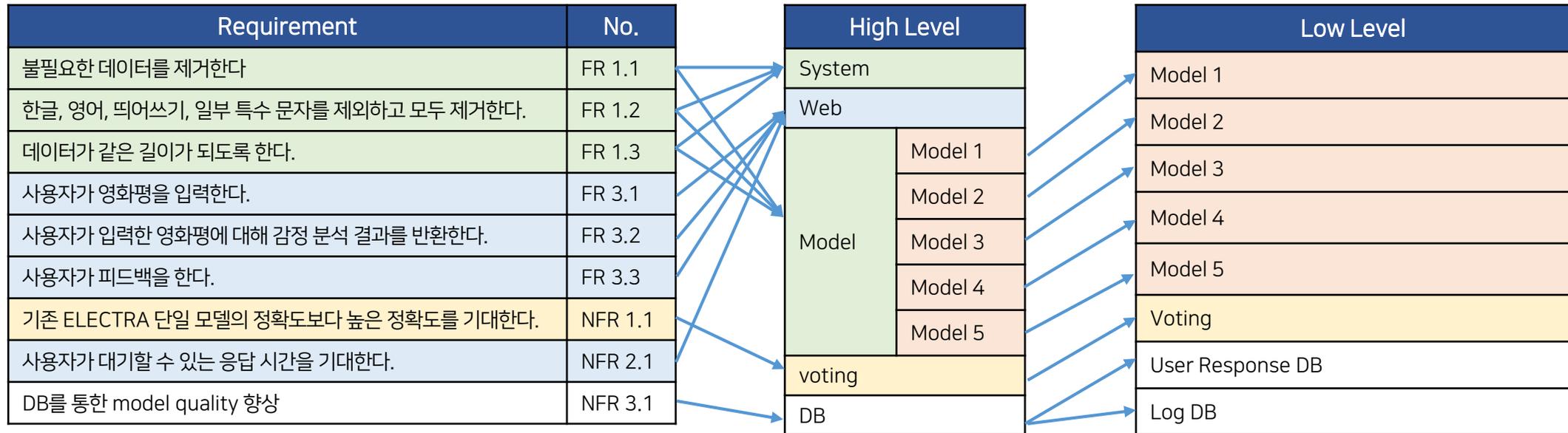
Requirement	R-No.
불필요한 데이터를 제거한다	FR 1.1
한글, 영어, 띄어쓰기, 일부 특수 문자를 제외하고 모두 제거한다.	FR 1.2
데이터가 같은 길이가 되도록 한다.	FR 1.3
사용자가 영화평을 입력한다.	FR 3.1
사용자가 입력한 영화평에 대해 감정 분석 결과를 반환한다.	FR 3.2
사용자가 피드백을 한다.	FR 3.3
기존 ELECTRA 단일 모델의 정확도보다 높은 정확도를 기대한다.	NFR 1.1
사용자가 대기할 수 있는 응답 시간을 기대한다.	NFR 2.1
DB를 통한 model quality 향상	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	금/부정 값이 올바르게 도출된다. 5초 이내에 결과가 도출된다.
18	금/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

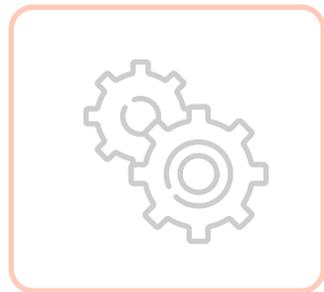




프로젝트 설명



데이터셋



시스템 테스트



추적성 분석



데모시나리오



데모시나리오

* 자세한 내용은 시연영상을 참고 바랍니다.

사용자 - 영화평 입력

RACCOON

영화평 입력을 통한 피드백 전송

INPUT REVIEW

모델 관리 및 업데이트

UPDATE MODEL

Team Raccoon
Copyright© 2020 All Right Reserved.

<웹 페이지 홈 화면>

INPUT REVIEW

긍정 / 부정 결과를 알고 싶은 영화평 입력

결과 보기

홈으로

공백만 입력

Team Raccoon
Copyright© 2020 All Right Reserved.

INPUT REVIEW

긍정 / 부정 결과를 알고 싶은 영화평 입력

1234567778

결과 보기

홈으로

숫자만 입력

Team Raccoon
Copyright© 2020 All Right Reserved.

INPUT REVIEW

긍정 / 부정 결과를 알고 싶은 영화평 입력

영화평을 입력해주세요...

다시 입력해 주세요

OK

다시 입력 받기

긍정적인 영화평 입력



INPUT REVIEW

긍정 / 부정 결과를 알고 싶은 영화평 입력

너무 재밌어요

결과 보기

홈으로

Team Raccoon
Copyright© 2020 All Right Reserved.



result

입력한 영화평

너무 재밌어요

예측한 결과

Positive한 영화평

내가 판단한 긍정/부정 선택

Positive Negative

피드백 보내기

▲ 예측 결과 -> 긍정

Team Raccoon
Copyright© 2020 All Right Reserved.

부정적인 영화평 입력



INPUT REVIEW

긍정 / 부정 결과를 알고 싶은 영화평 입력

이런 것도 영화라고 만들었냐? 감독 내려놓고 산에 들어가서 취직준비 해라

결과 보기

홈으로

Team Raccoon
Copyright© 2020 All Right Reserved.



result

입력한 영화평

이런 것도 영화라고 만들었냐? 감독 내려놓고 산에 들어가서 취직준비 해라

예측한 결과

Negative한 영화평

내가 판단한 긍정/부정 선택

Positive Negative

피드백 보내기

▲ 예측 결과 -> 부정

Team Raccoon
Copyright© 2020 All Right Reserved.

result

입력한 영화평

이런 것도 영화라고 만들었나? 감독 내려놓고 산에 들어가서 취직준비 해라

예측한 결과

Negative한 영화평

내가 판단한 긍정/부정 선택

Positive Negative

피드백 보내기

피드백 보내지 않고 홈으로

피드백 보내기

feedback

피드백 전송 완료

피드백이 정상적으로 저장되었습니다.

홈으로

Team Raccoon
Copyright© 2020 All Right Reserved.

result

입력한 영화평

이런 것도 영화라고 만들었나? 감독 내려놓고 산에 들어가서 취직준비 해라

예측한 결과

Negative한 영화평

내가 판단한 긍정/부정 선택

Positive Negative

피드백 보내기

피드백 보내지 않고 홈으로

피드백 보내지 않기

RACCOON

영화평 입력을 통한 피드백 전송

INPUT REVIEW

모델 관리 및 업데이트

UPDATE MODEL

Team Raccoon
Copyright© 2020 All Right Reserved.

TRANSLATE REVIEW

DB에 저장되어 있는 영화평

id	review	label
3152	정말 재밌어요 ! ㅋㅋㅋㅋ최고예요 ㅋㅋㅋㅋ	1
3153	넘 호 재밌음 이견 꼭 봐야해~	1
3154	추석영화로 가족과 즐겁게 보면 괜찮아요.. 부모님도 즐거워 하시고, 일요일 추천함.. 아쉬운 건 제작비가 열악한지 좀 액션이 ... 소박하다는거 ^^	1
3155	옛날엔 재미없었는데 얼마 전에 다시 보니 재미있네요	1
3156	이런 것도 영화라고 만들었나? 감독 내려놓고 산에 들어가서 취직준비 해라	0

저장 이력

NO.	Start ID.	End ID.	File Name
1	1	1000	201014.txt
2	1001	2000	201022.txt

Team Raccoon
Copyright© 2020 All Right Reserved.

▲저장됨
DB 확인
▼저장안됨

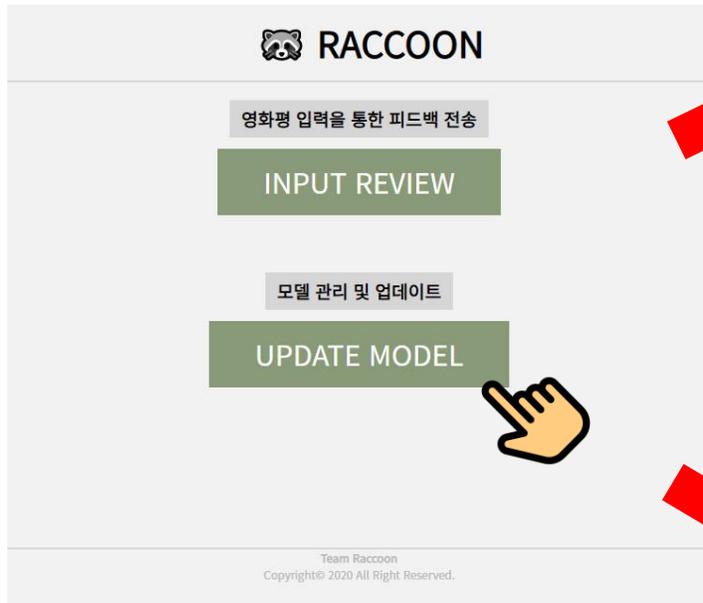
id	review	label
3150	몸에 꾸미지 않아도 빛나는 매력 또한 볼거리	1
3151	박보영으로 충분한 작품	1
3152	정말 재밌어요 ! ㅋㅋㅋㅋ최고예요 ㅋㅋㅋㅋ	1
3153	넘 호 재밌음 이견 꼭 봐야해~	1
3154	추석영화로 가족과 즐겁게 보면 괜찮아요.. 부모님도 즐거워 하시고, 일요일 추천함.. 아쉬운 건 제작비가 열악한지 좀 액션이 ... 소박하다는거 ^^	1
3155	옛날엔 재미없었는데 얼마 전에 다시 보니 재미있네요	1

저장 이력

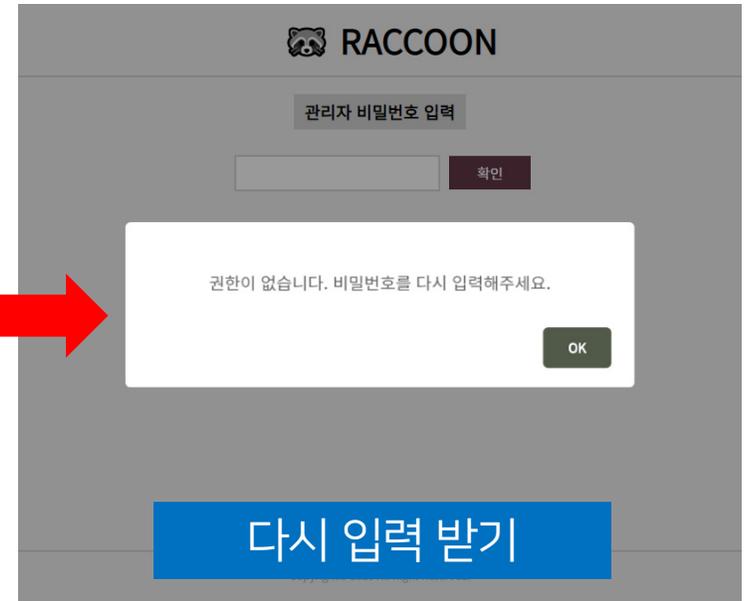
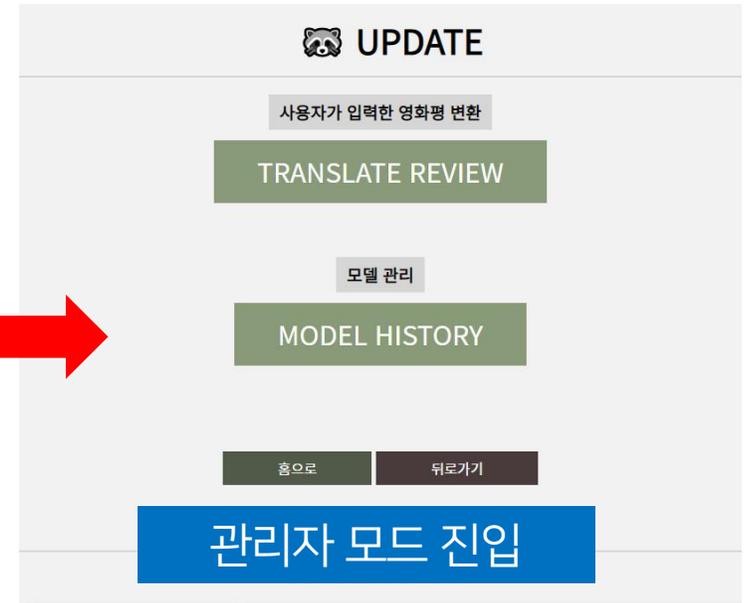
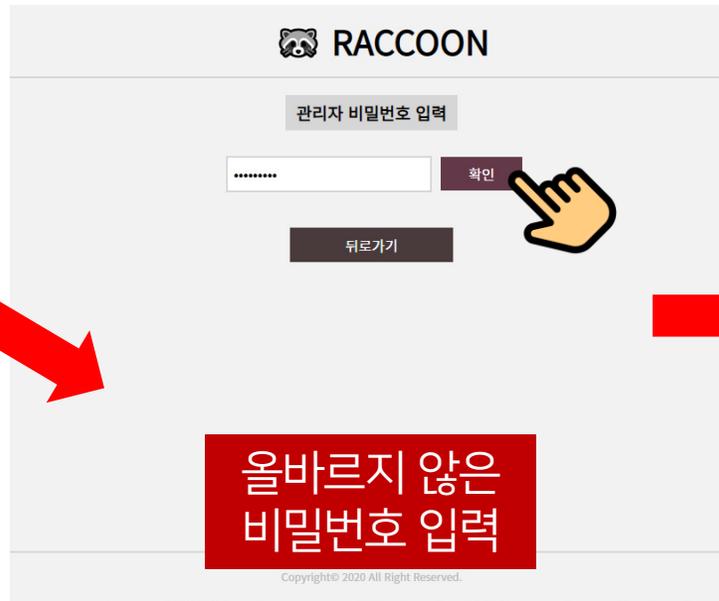
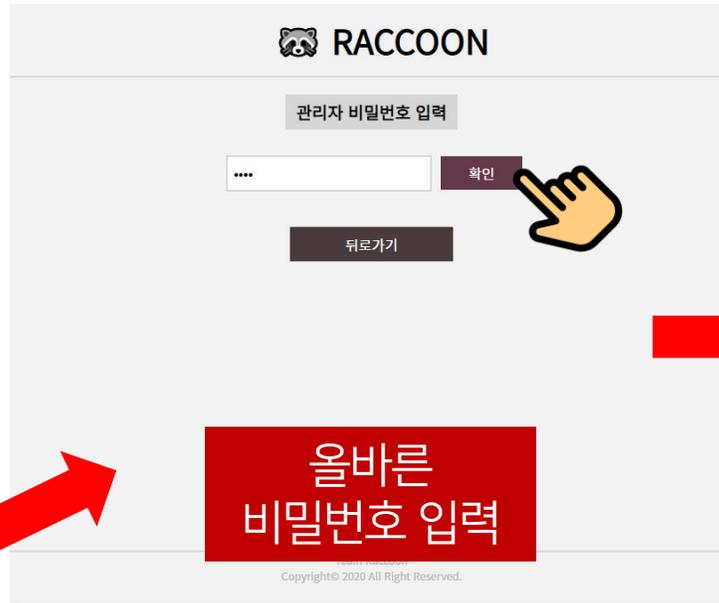
NO.	Start ID.	End ID.	File Name
1	1	1000	201014.txt
2	1001	2000	201022.txt

Team Raccoon
Copyright© 2020 All Right Reserved.

관리자 모드 진입



<웹 페이지 홈 화면>



관리자 모드 - 영화평 DB 확인 및 변환

UPDATE

사용자가 입력한 영화평 변환

TRANSLATE REVIEW

모델 관리

MODEL HISTORY

홈으로 | 뒤로가기

Team Raccoon
Copyright© 2020 All Right Reserved.

<관리자 페이지 화면>

피드백으로 저장된
영화평 DB 확인 가능

TRANSLATE REVIEW

DB에 저장되어 있는 영화평

id	review	label
1	처음부터.. 끝나기 5분전까진 정말 쟁있게 봤습니다.. 끝 5분은 대략 난감 ^^;;	1
2	고장석 부분빼고 다 노잼	0
3	템덴스라니 처음 보는 스타일의 영화라서 기대도 많이 되고 배우들도 너무 좋아해서 영화 너무 즐겁게 볼 것 같아요!! 이 영화로 힐링도 하고 그 당시 사람들의 아픔도 느끼며 마음 따뜻해지는 시간을 보내고 싶습니다. 너무너무 기대하고 있어요	1
4	여기 평점 이상하네요객관적으로 정말 안웃기고 억지연기 보기도 민망하고미안하지만 솔직히 문주고 보긴 아까워요	0
5	기대된다 남북 바뀐거 같아서 흥미롭고	1

저장 이력

NO.	Start ID.	End ID.	File Name
1	1	1000	201014.txt
2	1001	2000	201022.txt
3	2001	3000	201031.txt
4	3001	3100	201105.txt

TXT파일로 변환

Start ID. 1

End ID. 2000

Train-Set 201107.txt

저장 하기

홈으로 | 뒤로가기

Team Raccoon
Copyright© 2020 All Right Reserved.

변환할 영화평 ID와
파일 이름 입력

TRANSLATE REVIEW

저장 이력

NO.	Start ID.	End ID.	File Name
1	1	1000	201014.txt
2	1001	2000	201022.txt
3	2001	3000	201031.txt
4	3001	3100	201105.txt
5	1	2000	201107.txt

TXT파일로 변환

Start ID. 시작 ID

End ID. 종료 ID

Train-Set 저장할 File 이름

저장 하기

▲ TXT파일 변환 확인 가능

관리자 모드 - Model 관리

UPDATE

사용자가 입력한 영화평 변환

TRANSLATE REVIEW

모델 관리

MODEL HISTORY

홈으로 뒤로가기

Team Raccoon
Copyright© 2020 All Right Reserved.

<관리자 페이지 화면>

여러 모델 정보 확인 가능

MODEL HISTORY

Model History

ID	NAME	Parameter	Train-set	Test-set	Accuracy
1	ELECTRA+(CNN+LSTM)	<CNN> 필터 16 / 커널 3,4,5	nsmc	daum-trim	0.8741
2	ELECTRA+(CNN+LSTM)	<CNN> 필터 16 / 커널 3,4,5	2019	daum-trim	0.9278
3	ELECTRA+(CNN+LSTM)	<CNN> 필터 16 / 커널 3,4,5	nsmc	nsmc	0.8822
4	ELECTRA+(CNN+LSTM)	<CNN> 필터 16 / 커널 3,4,5	2019	2019	0.9525
5	ELECTRA-LSTM-CNN	<CNN> 필터 16 / 커널 3,4,5	nsmc	nsmc	0.8819

Model 저장

NAME: PreProcessing-ELECTRA-CNN-LSTM

Parameter: <CNN> 필터 16 / 커널 3,4

Train-Set: nsmc

Test-Set: nsmc

Accuracy: 0.8788

저장 하기

홈으로 뒤로가기

저장할 모델 정보 입력

MODEL HISTORY

Model History

ID	NAME	Parameter	Train-set	Test-set	Accuracy
13	PreProcessing-ELECTRA-CNN-LSTM	<CNN> 필터 32 / 커널 3,4,5	nsmc	nsmc	0.8778
14	PreProcessing-ELECTRA-CNN-LSTM	<CNN> 필터 08 / 커널 3,4,5	nsmc	nsmc	0.8784
15	PreProcessing-ELECTRA-CNN-LSTM	<CNN> 필터 16 / 커널 2,3,4,5	nsmc	nsmc	0.8785
16	PreProcessing-ELECTRA-CNN-LSTM	<CNN> 필터 16 / 커널 4,5	nsmc	nsmc	0.8788
17	PreProcessing-ELECTRA-CNN-LSTM	<CNN> 필터 16 / 커널 3,4	nsmc	nsmc	0.8788

모델 정보 저장 확인 가능

NAME: Name

Parameter: Parameter

Train-Set: Train-Set

Test-Set: Test-Set

Accuracy: Accuracy

저장 하기

홈으로 뒤로가기

Team Raccoon
Copyright© 2020 All Right Reserved.

감사합니다!

