



01:19

너의 기분 이모지?

-트위터 데이터를 활용한 이모지 추천 서비스-

김지연 이명아 이혜원 최연식



START



목차



개요



주제 선정 배경



감정 기반 이모지 추천



단어 기반 이모지 추천





주제 선정 배경



기존 이모지 서비스 시스템의 문제점



➔ 텍스트 기반의 감정 이모지 추천기능이 존재하지 않음

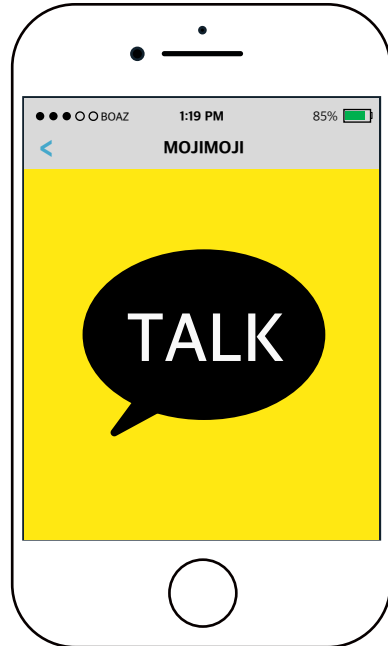




주제 선정 배경



기존 이모지 서비스 시스템의 문제점



카카오톡 이모지 서비스 현황

- 가장 최신에 쓴 것을 보여줌
- 원하는 이모지 페이지 찾는 데 힘이 든다



감정기반 이모지 추천 서비스










단어기반 이모지 추천 서비스





데이터 수집

 www.emojitracker.com

 2294752619	 1098743692	 937001068	 843743176	 700431186
 365326176	 338912449	 326568854	 318671093	 311386864
 208053781	 203601071	 201219432	 200967336	 200491981
 162658734	 161474852	 157025731	 152849543	 151141567
 129952037	 123593982	 122519437	 122116921	 119267902
 98939685	 98402165	 97404313	 95760098	 92982548
 87304996	 86267544	 83943210	 83714190	 80039932
 70546721	 70251198	 69952960	 69813945	 69237330



데이터 수집

emojitracker 란? www.emojitracker.com

세계에서 사용되는 Emoji의 개수를 실시간으로 집계

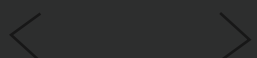
2294752619	1098743692	937001068	843743176	700431186
365326176	338912449	326568854	318671093	311386864
208053781	203601071	201219432	200967336	200491981
162058734	151474832	157023731	152849543	151141567
129532037	123553562	122519437	122170927	110207902
98939685	98402165	97404313	95760098	92982548
87304996	86267544	83943210	83714190	80039932
70546721	70251198	69952960	69813945	69237330

라벨 선정 기준

step1. emoji 공식 사이트에서 제공하는 분류 카테고리 참조

step2. 카테고리별로 지정된 이모지 개수에 따라 emojitracker에서 상위에 랭크된 이모지들을 배정한다.

총 30개 라벨 선정





Emoji Label 설정

smile	affection	skeptical	sleepy	unwell	concerned	tongue	hand	glasses	negative	Etc



데이터 크롤링

Twitter Scraper 이용 : 2010년 이후의 tweet 크롤링

Unicode	데이터 개수
1F44C	312221
1F44D	364179
1F601	750000
1F602	493643
1F600	220556
1F60A	374255
1F60F	359578

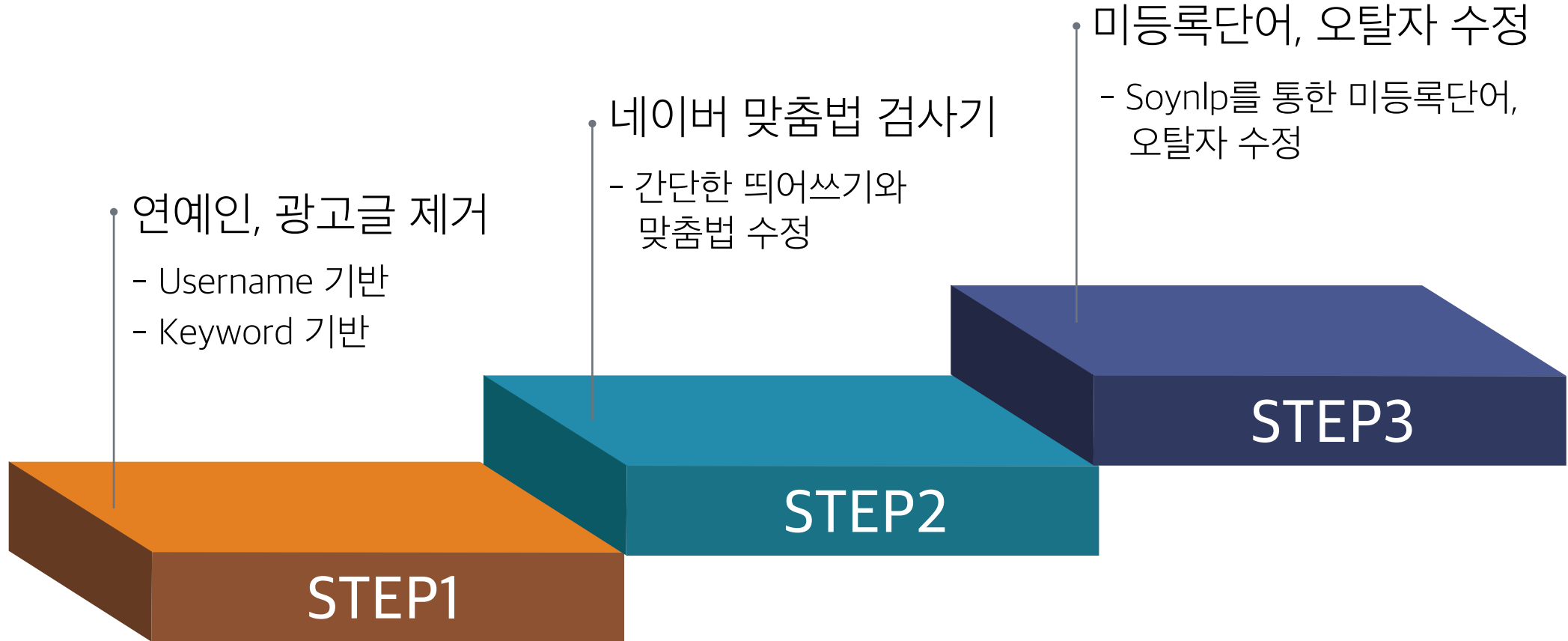
⋮

총 데이터 개수
약 **1103만개**





데이터 전처리





Step1. 연예인, 광고글 제거

1) 연예인 - Username 기반

- 연예인 관련 글의 경우 RT(리트윗)이 많아 반복되는 Username 존재
- 빈도순으로 정렬 후 상위 0.2%에 해당되는 Username이 들어간 트윗 삭제

User_name	Frequency
@BTS_twt	6609
@JYHeffect	429
@pledis_17	309
@CHA_NNNNN	276
@BAP_Daehyun	270
@BAP_Daehyun	262
@JUNGTW_LEO	220

2) 광고글 - Keyword 기반

- 광고글에 많이 나오는 단어를 기반으로 keyword 작성 후 해당 단어가 들어간 트윗 모두 제거

카카오톡
상담
주소
예약
문의
환불
링크
⋮

Step2. 네이버 맞춤법 검사기 - 간단한 띄어쓰기와 맞춤법 수정



Konlpy 와 Soynlp

Konlpy

: 기존에 품사가 적혀있는 데이터를 학습시켜 문장을 단어들로 분해

한계점

미등록 단어, 은어, 아직 남아있는 오타자가 있어 단어를 올바르게 인식 하는데 한계가 있음

Soynlp

: 통계적 패턴을 이용하여 단어를 찾아내줌(비지도학습)

① Cohesion Score

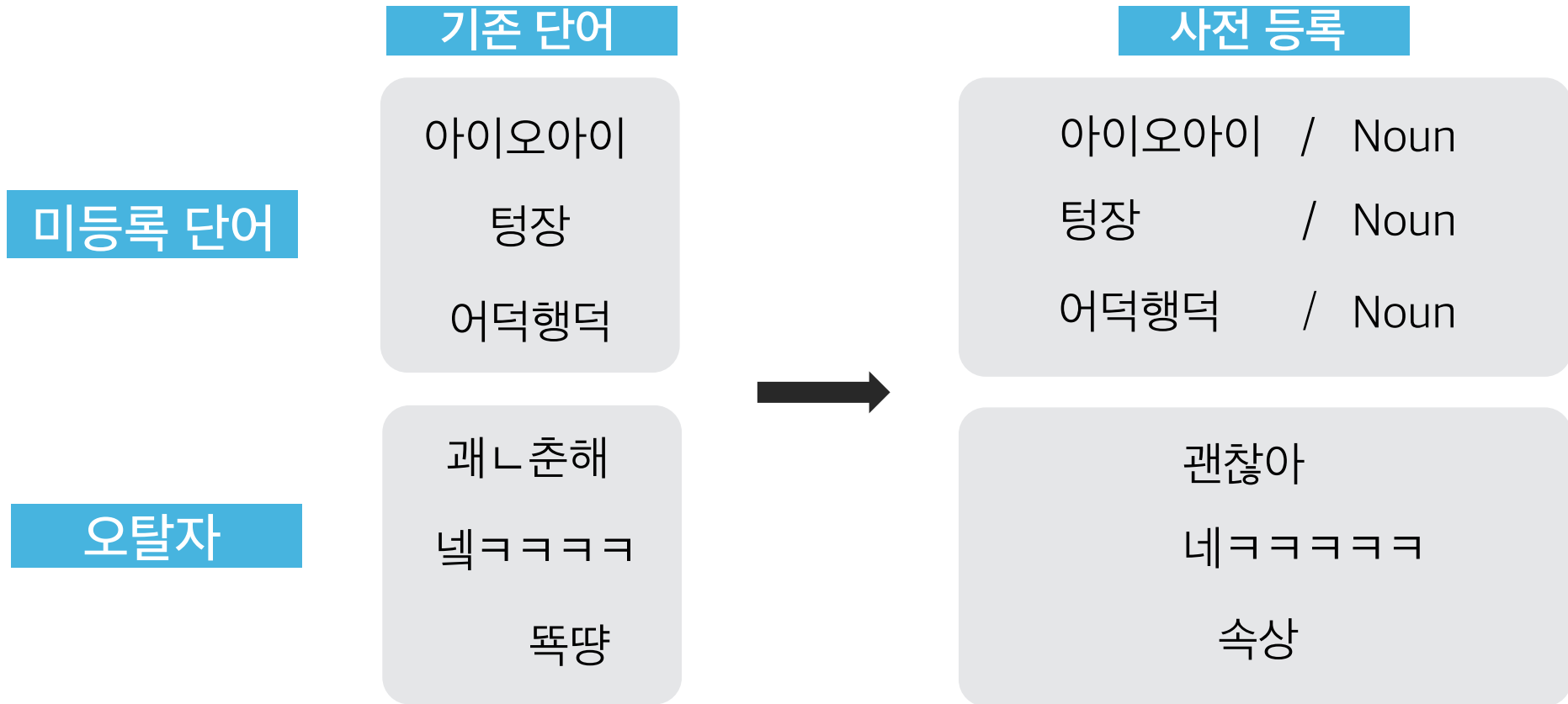
주어진 글자가 함께 자주 나타나는가

② right branching entropy

해당 단어의 우측에 다른 단어가 자주 등장하는가



Step3. 미등록 단어, 오타자 수정





학습데이터 Word Extraction - Soynlp
 : 미등록 단어 중 자주 등장하는 단어와
 오타자를 잡아냄

미등록 단어

오타자

기존 단어

아이오아이
 텅장
 어덕행덕

괘니춘해
 넉ㅋㅋㅋㅋ
 똑땅



사전 등록

아이오아이 / Noun
 텅장 / Noun
 어덕행덕 / Noun

관찰아
 네ㅋㅋㅋㅋ
 속상

Add dictionary & Replace - Konlpy
 : Word Extraction을 통해 추출한 단어들을
 사전에 등록해주고 교정해주는 작업





전처리 완료된 트윗

tweet

- 0 버스 안에서 다정하게 손잡은 연인 덕에 못 내릴뻔했어요 더럽게 고마워요 또 걸리면...
- 1 우와아 기능에 선택한 글을 읽어주는 기능이 있어요 이동하면서 액정 보기 곤란할 때...
- 2 이번 나꼼수 다운로드하면 홍준표 님을 배려해서 꼭 황금시간대에 들을랍니다
- 3 부모님과 동행할 때 절대 착장해서 안되는 아이템은 컨버스 운동화와 야상 재킷 이 두...
- 4 시간 안에 한 권 다 읽었다 오래간만에 집중 역시 역시 좋아
- 5 코 손대 닭갈비 먹고 싶다고 해서 난 쿨하게 오케이 했지 난 너무 착해 훗 여긴 우...
- 6 가을이면 가을답게 쌀쌀하기만 했음 좋겠다 내일 두고 보게 써
- 7 공부할 시간이 없다 고 핑계 대지 말자 운동할 시간이 없다 고 핑계 대지 말자
- 8 아 발표 끝 아우 아우 아우 이제 발표 두 개 남았다
- 9 반찬은 김치 깎두기 정도
- 10 서민 생각 산소 축내는 소리하고 자빠졌다 너 님이 영원히 푸 옥 쳐주무셔야 우리 서...

Tokenizing : Customized Konlpy의 Twitter 이용



전처리 완료된 트윗

tweet

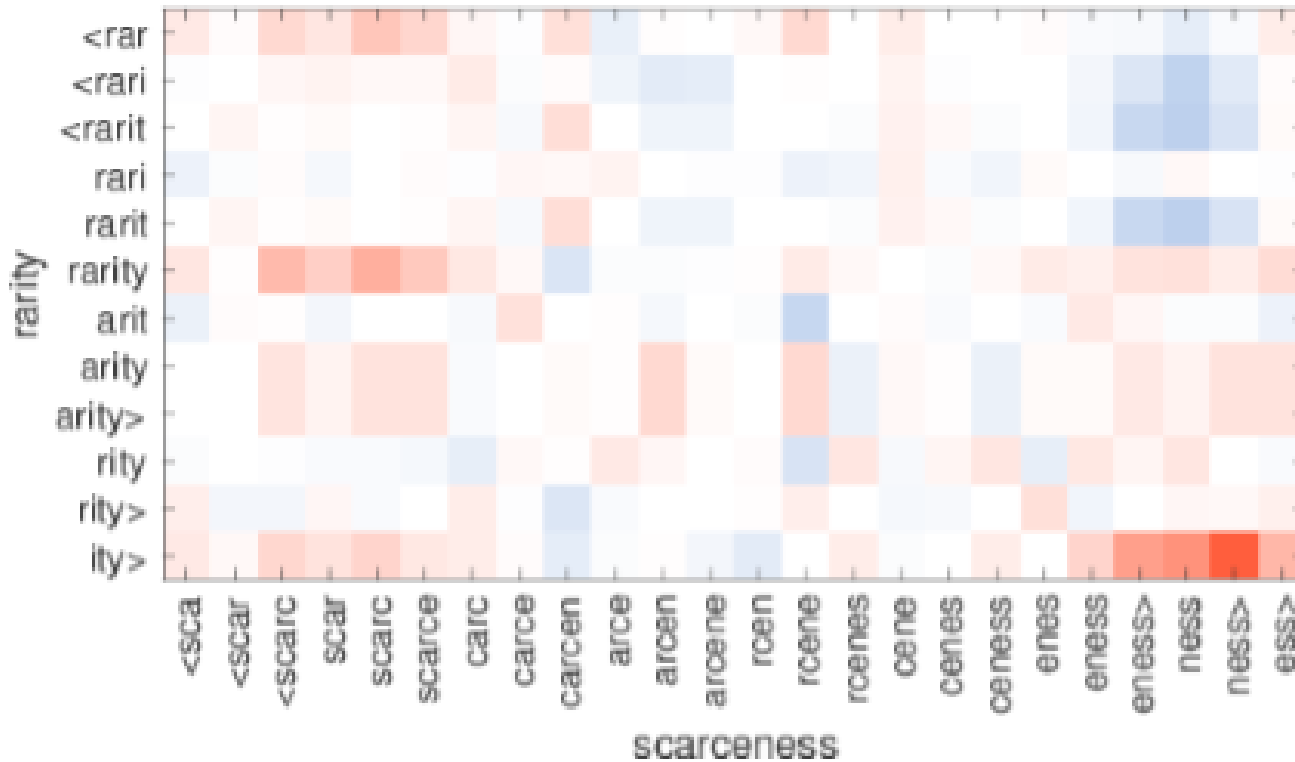
0 1. Upsampling & Downsampling을 통한
 1 데이터 불균형 해결
 2 이번 니킴수 다운로드하면 흥준표 님을 배려해서 꼭 황금시간대에 들을랍니다
 3 부모님과 동행할 때 절대 착장해서 안되는 아이템은 컨버스 운동화와 야상 재킷 이 두...
 4 2. 이모지당 데이터 개수 10만개
 5 총 데이터 300만개
 6 큰 손대 닭갈비 먹고 싶다고 해서 난 쿨하게 오케이 했지 난 너무 착해 훗 여긴 우...
 7 가을이면 가을답게 쌀쌀하기만 했음 좋겠다 내일 두고 보게 써
 8 공부할 시간이 없다 고 핑계 대지 말자 운동할 시간이 없다 고 핑계 대지 말자
 9 아 발표 끝 아우 아우 아우 이제 발표 두 개 남았다
 10 반찬은 김치 깍두기 정도
 서민 생각 산소 축내는 소리하고 자빠졌다 너 님이 영원히 푸 옥 쳐주무셔야 우리 서...





Model 구축

FastText



word vector representation과 text classification을 도와주는 **오픈소스**

단어 안의 subword까지 고려하여 **word embedding - 추후 문장 분류모델로 이용**

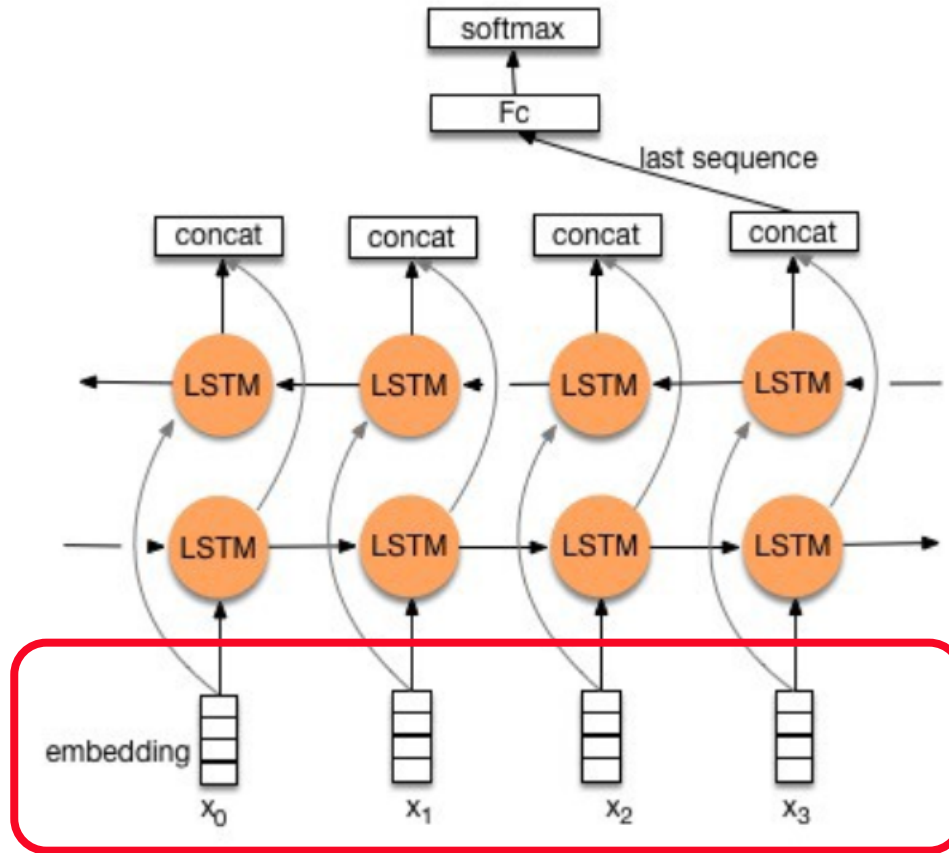
속도가 매우 빠르고 대용량 데이터를 처리할 수 있다는 장점이 있음



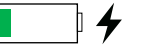
Model 구축

🔍 Bilstm with Attention Layer

input : 각 트윗의 문장

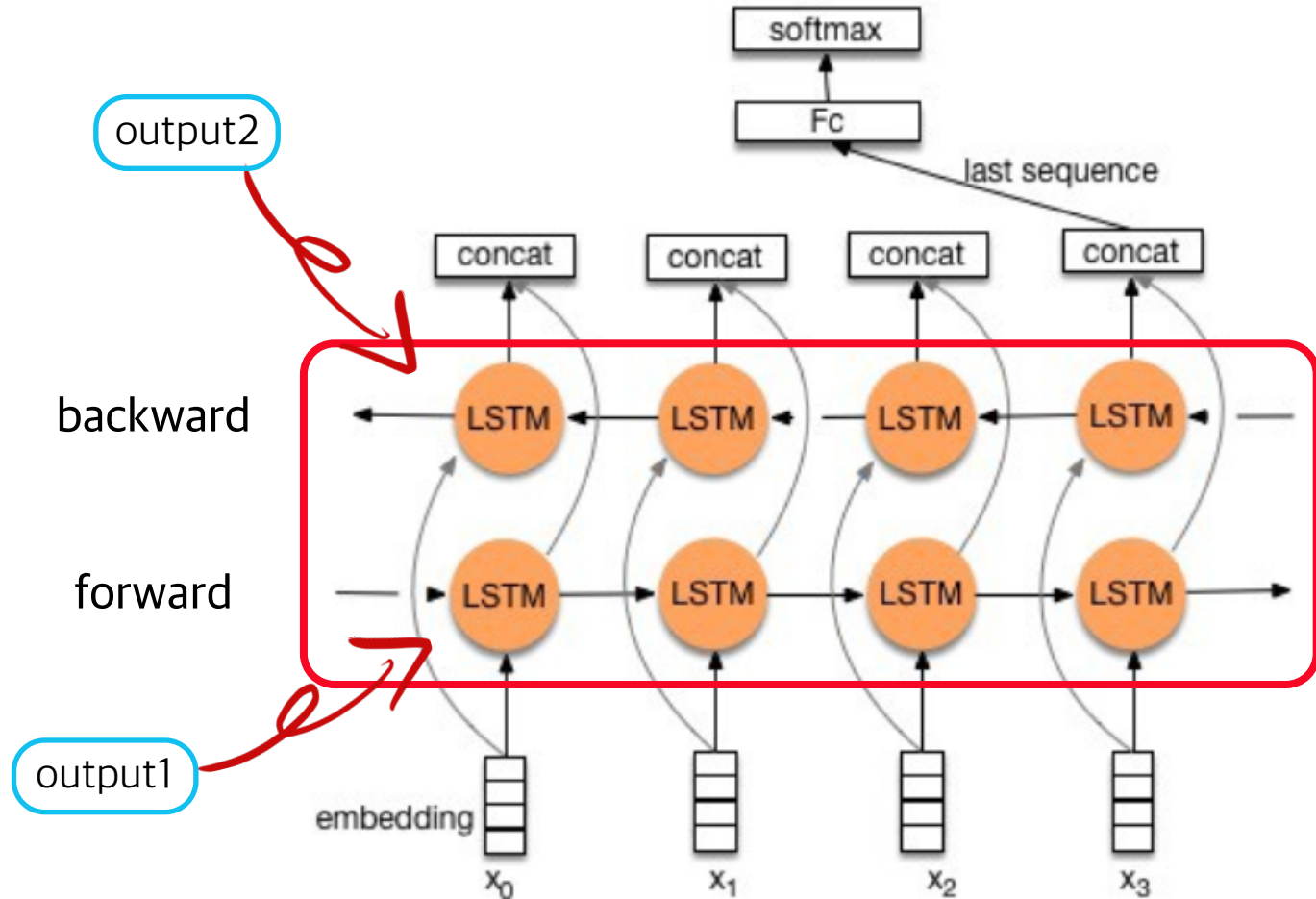


돈 벌기 참 힘들다 : 하나의 셀에 각각의 단어를 input으로 받음



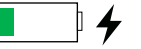
Model 구축

🔍 Bilstm with Attention Layer



bidirectional lstm : 양방향의 lstm 존재
 각 cell마다 2개의 output 산출

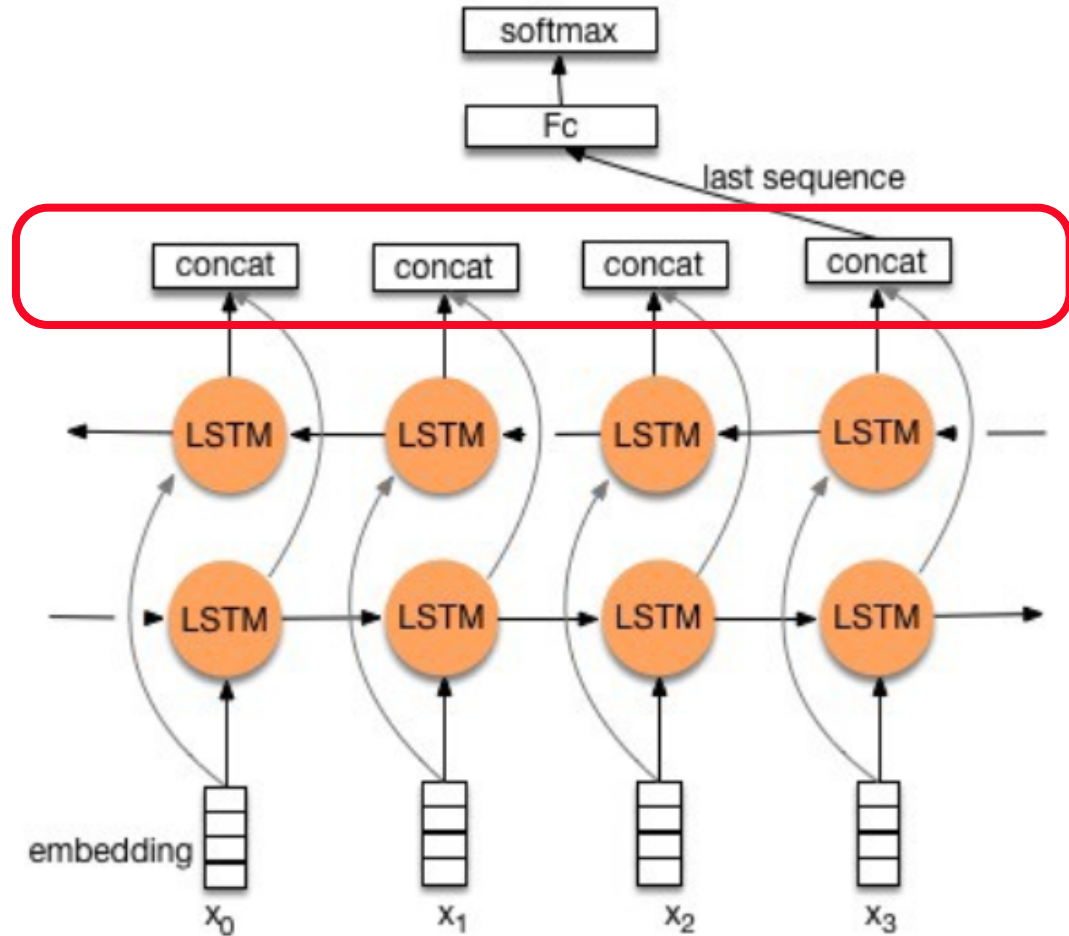
문장의 순서대로 학습
 각 cell의 input 단어와 앞단의 정보를
 보존하면서 학습을 진행



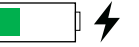
Model 구축

🔍 Bilstm with Attention Layer

output

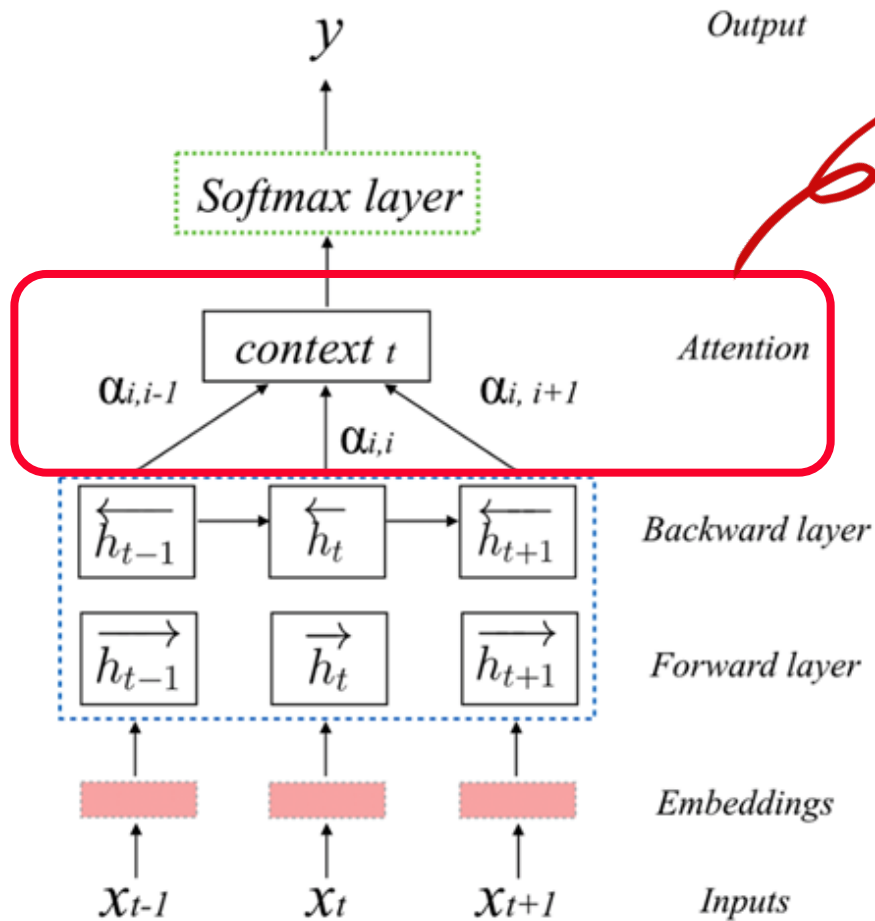


앞단에 나온 2개의 결과물을 concatenate.
Bilstm의 최종 output이 됨



Model 구축

🔍 Bilstm with Attention Layer



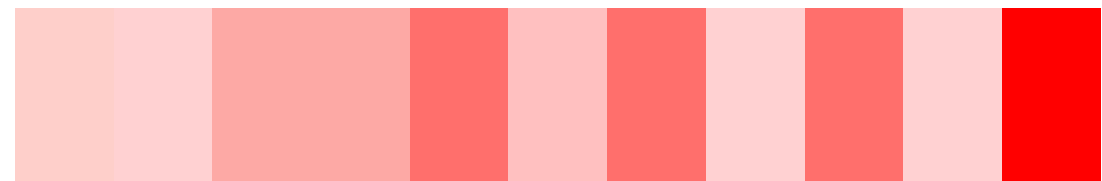
Bilstm 결과물에 **Attention Mechanism** 적용

: 각 cell의 output과 attention weight(alpha)의 곱을 합한 weighted sum 을 구한 후 softmax layer를 통과시켜 class 분류 진행

*attention weight는 학습 과정 중 분류를 잘 하는 방향으로 update 진행



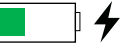
내 맘은 이리 울적한 데 말하다 사람이



없다 나다 가끔 활짝 웃다 싶다 결엔 아무도 없다

색이 진한 단어 '울적', '없다' 는 attention weight가 가장 큰 단어

Class 분류에 중요한 영향을 미친 단어일수록 attention weight(or score)가 큼



Accuracy 비교

label 20개

	top1 Accuracy	top5 Accuracy
FastText	21.4%	53.3%
text CNN	24%	63.2%
Att Bilstm	25%	64%

VS

label 30개

	top1 Accuracy	top5 Accuracy
FastText	19.0%	45%
text CNN	16%	46.9%
Att Bilstm	19.1%	50.1%



Accuracy 비교

label 20개

label 30개

최종 모델

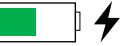
Label 30개 & Att Bilstm 선택

Model	top1 Accuracy	top5 Accuracy
FastText	21.4%	53.3%
text CNN	24%	63.2%
Att Bilstm	25%	64%

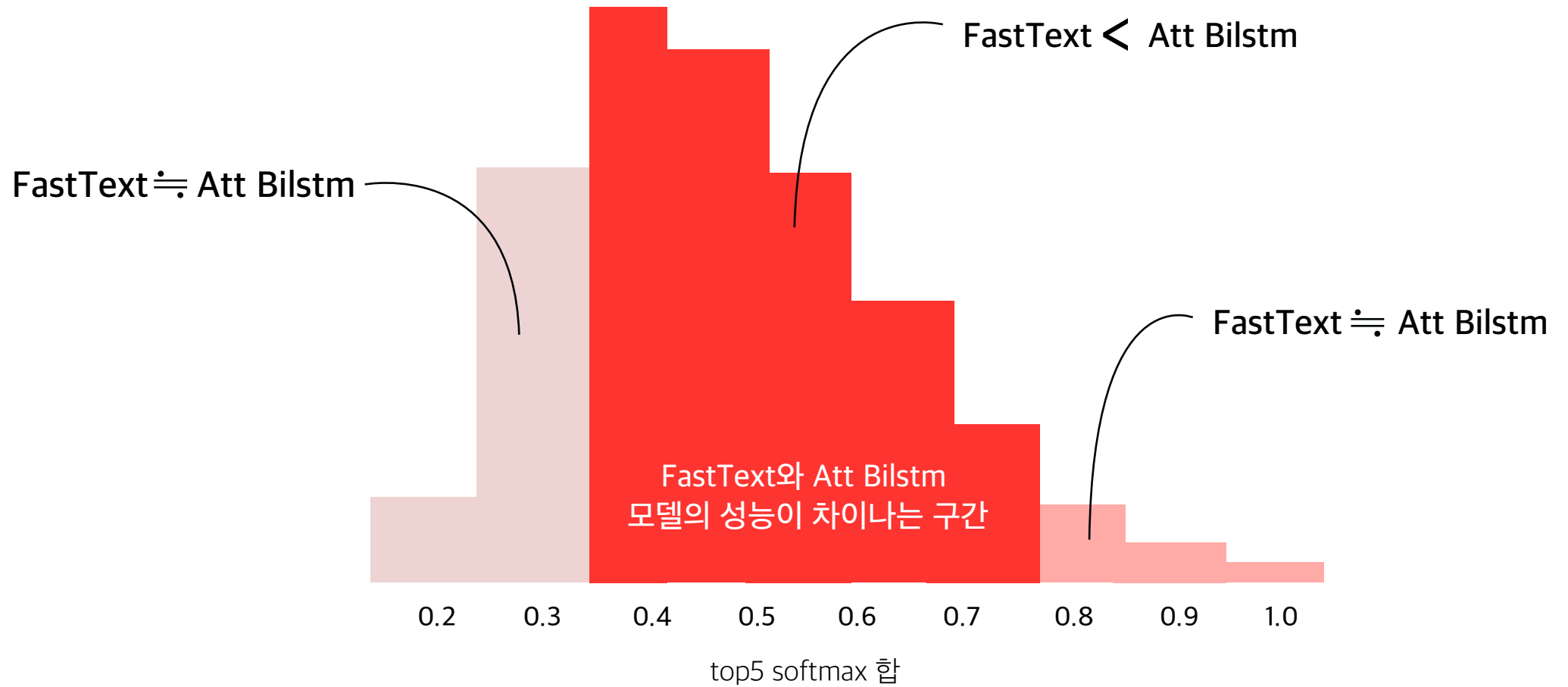
VS

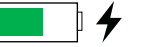
Model	top1 Accuracy	top5 Accuracy
FastText	19.0%	45%
text CNN	16%	46.9%
Att Bilstm	19.1%	50.1%








모델 성능 비교 - FastText와 Att Bilstm 의 차이



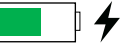


softmax 합 0.8~1.0 사이

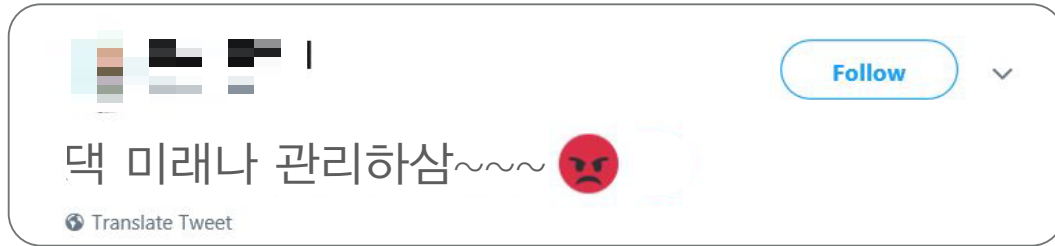
이모지와 대응되는 **정확한 토큰** 존재

- ex) 여러분들 **독감** 조심해요 $\pi\pi$ 죽겠어요 진짜 : 
- 약속이 없으니 대신 **오케이**입니다. : 
- 으익 누구야 언니 **화나게** 한 사람 : 

FastText와 Att Bilstm 둘 다 accuracy 비슷

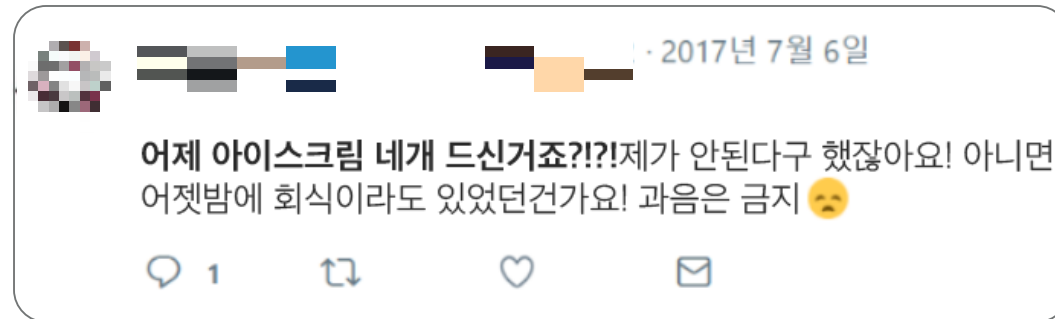


softmax 합 0.4~0.8 사이



FastText : 😎 😊 😱 😊 😞

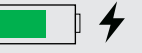
Att Bilstm : 😐 😞 😞 😞 😞



FastText : 😊 😞 😡 😊 😞

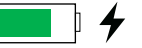
Att Bilstm : 😐 😞 😞 😞 😞

문맥을 고려해야 하는 경우에서 Att Bilstm 이 FastText 보다 Accuracy **7~8%** 높게 나타남



단어기반 이미지 추천 서비스



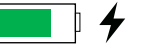


데이터 수집



www.emojitracker.com

2294752619	1098743692	937001068	843743176	700431186
365326176	338912449	326568854	318671093	311386864
208053781	203601071	201219432	200967336	200491981
162658734	161474852	157025731	152849543	151141567
129952037	123593982	122519437	122116921	119267902
98939685	98402165	97404313	95760098	92982548
87304996	86267544	83943210	83714190	80039932
70546721	70251198	69952960	69813945	69237330



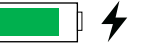
데이터 수집



www.emojitracker.com

사물이모지 198개 크롤링

2294752619	1098743692	937001068	843743176	700431186
365326176	338912449	326568854	318671093	311386864
208053781	203601071	201219432	200967336	200491981
162658734	161474852	157025731	152849543	151141567
129952037	123593982	122519437	122116921	119267902
98939685	98402165	97404313	95760098	92982548
87304996	86267544	83943210	83714190	80039932
70546721	70251198	69952960	69813945	69237330



단어 사전 구축

step2. Soynlp의 Word Extraction 활용하여 Wordlist 추출

step3. 각 이모지에 쓰이는 은어, 관용어, 유사어 사전 구축

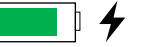


'피자', '음식', '핏짜', '피맥', '피짜', '피자', '핏자', 'PIZZA', 'pizza'



flight', '비행기', '출국', '비행', '입국', '승승', '뱅기', '다녀와', '공항', 'airport', '떠나', '여행





MOJIMOJI

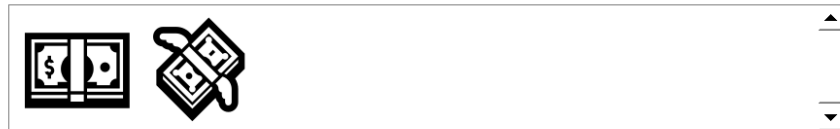
예시 1)

오늘도 내 통장은 텅장

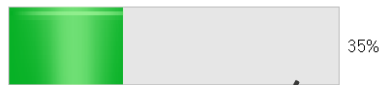
submit



감정기반 emoji



사물기반 emoji



35%

예시 2)

생일 축하드려요 좋은 하루 보내세요

submit



감정기반 emoji



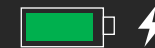
사물기반 emoji



74%



ProgressBar : top5 emoji의 softmax값의 합을 보여줌



한계점 및 보완점

데이터 중복 라벨링을 고려하지 않음

데이터 클렌징의 한계

개인의 발화 & 이모지 사용 특성을 고려하지 못함





BOAZ



지금부터

새로운 이모지 추천 서비스

MOJIMOJI를 시연하겠습니다.





MOJIMOJI 시연 1



mojimoji

Artificial emotional intelligence

발표자님 너무 아름다우시네요

submit

오늘도 내 통장은 텅장



왜 내 카톡을 안 봐 너무 똑딱해

안 봐서 모르지만 본인이 그렇다고 하시니

Clear

보아즈 컨퍼런스에 오신 여러분 환영합니다



48%





MOJIMOJI 시연 2



mojimoji

Artificial emotional intelligence

발표자님 지금 많이 떨리시나요?

submit

오늘도 내 통장은 텅장

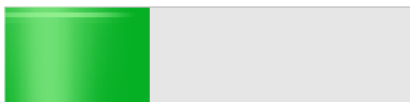


왜 내 카톡을 안 봐 너무 똑땡해

안 봐서 모르지만 본인이 그렇다고 하시니

Clear

보아즈 컨퍼런스에 오신 여러분 환영합니다



36%





MOJIMOJI 시연 3



mojimoji

Artificial emotional intelligence

너나 잘하세요ㅋ

submit

오늘도 내 통장은 텅장



왜 내 카톡을 안 봐 너무 똑딱해

안 봐서 모르지만 본인이 그렇다고 하시니

Clear

보아즈 컨퍼런스에 오신 여러분 환영합니다



43%



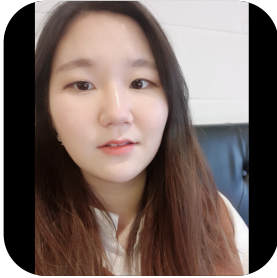





🔍 emoji 팀 소개

MOJIMOJI 제안

Team emoji

			
김지연	이명아	이혜원	최연식
