리뷰에는 이미 별점이 있는데 별점을 또 예측해서 뭘 하나요

Cafe24 김준민

본 발표는 기술관련 내용이 없습니다.

본 슬라이드에 있는 데이터(리뷰)는
실제 데이터가 아닙니다.

발표자

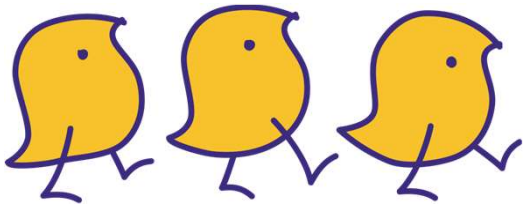
- 2019년
- 김준민
- 데이터 분석팀 기업전사
- 이케이케 잘 하기

프로젝트 당시 발표자

- 2017년
- 김준민
- 혼자 딥러닝 하는 DBA
- 이케이케 잘 하기

프로젝트 당시 cafe24

- 2017년
- Simplex Internet
- 쇼핑몰 대략 100만개
- 딥러닝 경험/인프라 없음



컴플레인

LangCon 2019

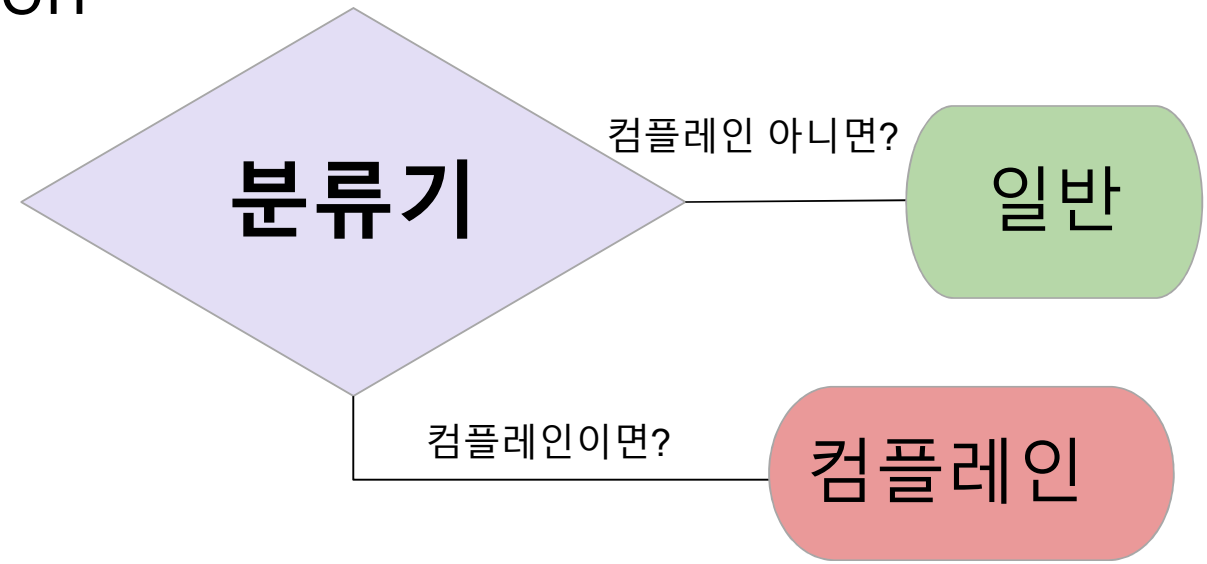
프로젝트 시작

- 좋은 리뷰에는 자동댓글이 달리는데
- 컴플레인에는 수동댓글이 달림
- 댓글다는 걸 자동화 해줄 수 없나?
- 유저들이 별점 5점 주고 불평하는데, 이런 거 찾아줄 수 없나?


평점
아주 좋아요
키 159~161
상의사이즈 66
선택한 옵션 COLOR: 블랙
실밥 다 뜯어져 있어서 손으로 하나씩 다 제거했는데 뒤집어보니 박음질도 제대로 안 되어있어요. 검품은 하고 보내는건가요? 후기보고 믿고 구매했는데 이런물건이나 보내고 랭콘 정말 실망입니다 반품보낼테니 환불해주세요 다시는 여기서 구매하지 않을 듯
***
이 리뷰에 대해 아직 도움이 된다고 평가한 사람이 없습니다.
이 리뷰가 도움이 되었나요?
네
아니요
DETAIL

프로젝트 시작

- 일반리뷰 vs 컴플레인
- Binary Classification



파일럿 - 데이터

- 쇼핑몰들에서 코멘트 기준으로 리뷰를 수집
- 다른리뷰랑 거의 같은 코멘트가 써있으면 일반리뷰
- 거의 유니크한 코멘트가 써있으면 컴플레인 리뷰
- +수동 라벨링

- 클래스별 1000개씩 수집

파일럿 - 모델작성

- 맨날 하던거
  - konlpy
  - tfidf vectorizer
  - SVM, RF, NB, etc.
  - stratified 5 fold cv
- 83.7%의 정확도

파일럿 - 모델 확인

- 83.7%의 정확도
  - term frequancy 특성상
  - "실망하지 않네요"나
  - "박음질이 잘 되어있어요" 같은 건
  - 컴플레인

파일럿 - 모델 확인

N: 일반리뷰, C: 컴플레인

```
In [101]: small_grid.predict(vect2.transform(['기대를 너무 많이 했던터라 실망하면 어쩌지 했는데....마음에 들어요^^']))
Out[101]: array(['C'],
                dtype='<U1')

In [100]: small_grid.predict_proba(vect2.transform(['기대를 너무 많이 했던터라 실망하면 어쩌지 했는데....마음에 들어
Out[100]: array([[ 0.81000415,  0.18999585]])
```

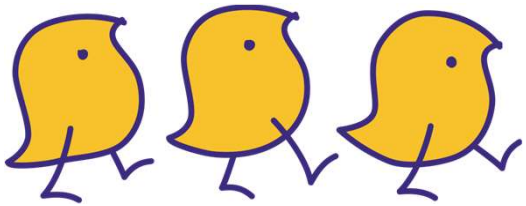
'박음질 좀 잘 해주세요' vs '박음질을 잘 해주셨네요'

1.데이터가 적어서 그렇다

- 추가 데이터 수집

2.문맥고려가 안 되어서 그렇다

- Term Frequency 말고 다른걸로 하기

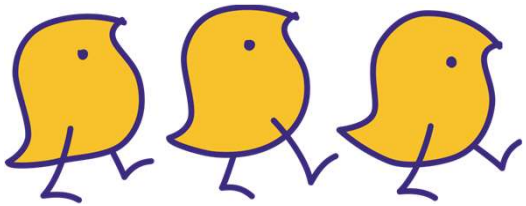


컴플레인 데이터

LangCon 2019

더 많은 불평 데이터가 필요한데

- 운영자 댓글 기준으로 판별하기엔 데이터가 부족함
- 데이터를 더 만들려면 수동라벨링 해야 할 것 같음
- 하기싫음
- (프로젝트 목적이 잘못된 거 아니야?)



딥러닝(생략)

LangCon 2019

Keras: The Python Deep Learning library

Keras

(=1.2.2)

You have just found Keras.

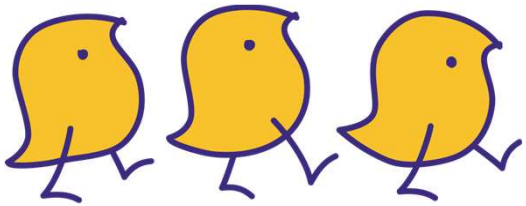
Keras

- you just found keras(=1.2.2) + konlpy
- 그럼 어떻게 구현하는지 남들이 하는걸 볼까?

논문을 보니

- Online Review Sentiment Analysis
- 1,2,3점은 Neg, 4,5점은 Pos로 놓고 분석

Dataset	Classes	Train Samples	Test Samples	Epoch Size
AG's News	4	120,000	7,600	5,000
Sogou News	5	450,000	60,000	5,000
DBPedia	14	560,000	70,000	5,000
Yelp Review Polarity	2	560,000	38,000	5,000
Yelp Review Full	5	650,000	50,000	5,000
Yahoo! Answers	10	1,400,000	60,000	10,000
Amazon Review Full	5	3,000,000	650,000	30,000
Amazon Review Polarity	2	3,600,000	400,000	30,000



컴플레인 데이터

LangCon 2019

더 많은 불평 데이터가 필요한데

- 운영자 댓글 기준으로 판별하기엔 데이터가 부족함
- 데이터를 더 만들려면 수동라벨링 해야 할 것 같음
- 하기싫음

안 해

프로젝트 시작

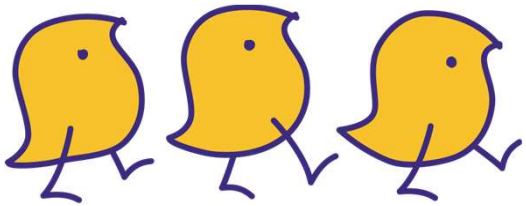
- 좋은 리뷰에는 자동댓글이 달리는데
- 컴플레인에는 수동댓글이 달림
- 댓글다는 걸 자동화 해줄 수 없나?
- 유저들이 별점 5점 주고 불평하는데, 이런 거 찾아줄 수 없나?

더 많은 불평이 필요해서

- 기본 별점이 5니까, 별점 5점 리뷰중에는 컴플레인 리뷰도 있음
- 최초목적: 별점 5점인 리뷰중에 컴플레인을 구분해내는 것
  - 4점리뷰를 일반리뷰 데이터로 사용
- 3점은 컴플레인인지 아닌지 모르겠음
  - 1,2점리뷰를 컴플레인 리뷰 데이터로 사용

더 많은 불평

- 1,2점 리뷰: 65,000개
- 4점 리뷰: 65,000개



프로젝트

LangCon 2019

파일럿과의 비교

텍스트 데이터로	텍스트 데이터로
운영자의 수동댓글 필요성을	리뷰 별점을
예측한다	예측한다
!	?

파일럿과의 비교

- 별점은 이미 있는데 이걸 예측한다니 무슨 소리요

파일럿과의 비교 - 프로젝트 개괄

텍스트 데이터로	텍스트 데이터로
운영자의 수동댓글 필요성을	유저가 주려던 별점을
예측한다	예측한다
!	!

파일럿과의 비교 - 예측모델의 기능(초심 잃음)

- 운영자의 수동댓글 필요여부를 예측한다
- 유저의 별점기입 미스를 예측한다

파일럿과의 비교 - 프로젝트 목적(초심 있음)

- 운영자의 수동댓글 필요여부를 예측한다
- 유저의 별점기입 미스 검출를 통해 수동댓글 필요여부를 예측한다

어차피 별점 낮은 리뷰는 불평일 것임

파일럿과의 비교 - 데이터

리뷰데이터	리뷰데이터
수동댓글 작성된 리뷰	별점 1,2점 리뷰
그 외의 리뷰	별점 4점 리뷰
2K	많이

파일럿과의 비교 - 시스템

Konlpy(twitter)	Konlpy(twitter)
TF-IDF Vectorize	Word Embedding
SVM	LSTM
acc 83%	acc 83%?

파일럿과의 비교를 하고 싶은데

- 정확도가 83%라니 이게 무슨 소리요

파일럿보다 분명 더 잘 나올 것 같은데

- Hyper Parameter Tuning
- LSTM & CNN
- Batch Norm
- Regularization
- Text Pre-Processing
- Dictionary
- Word2Vec
- FastText

원인은

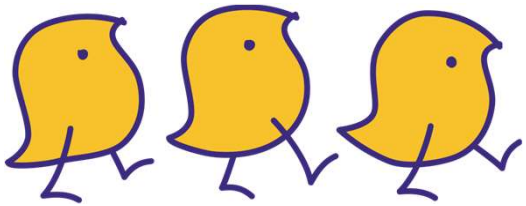
- 텍스트데이터를 수집할 때, 텍스트 첫 줄만 수집된 것ㅋ
- 줄바꿈 뒤로는 수집이 안 됨..
- 왜 몰랐냐면

Out[34] :

	0
0	실밥 다 뜯어져 있어서 손으로 하나씩 다 제거했는데 뒤집어보니 박음질도 제대로 안 ...
1	역시 랭콘 실망시키지 않네요. 여기서 구매한 거 다 대만족이에요.
2	사진에는 겨자색인데 배송받아서 보니까 완전 레몬색이에요. 다른 구매후기들 보니까
3	산골짜기 다람쥐 아기 다람쥐 도토리 점심 가지고 소풍을 간다 다람쥐야 다람쥐야 재주...
4	모델 착샷이랑 완전 똑같아요! 머리도 작아보이고 제가 찾던 딱 그 스탈! 근데

데이터를 다시 수집했더니

Konlpy(twitter)	Konlpy(Mecab)
TF-IDF Vectorize	Word Embedding
SVM	LSTM
acc 83%	acc 90%!



모델 배포

LangCon 2019

모델을 배포하면 어떻게 되지요

- 1개, 혹은 1개 이상의 데이터를 입력받아
- 모델에 따른 예측을 수행하여
- 예측값을 출력하는 웹서비스

모델을 배포하면 어떻게 되지요

웹서비스

Previously...

- 2017년
- 혼자 딥러닝 하는 DBA
- 쇼핑몰 대략 100만개
- 회사는 딥러닝 경험/인프라 없음
- 그런데 모델은 keras를 사용한 딥러닝 시스템

회사에서 신규 서비스를 런칭하려면

- 서버 성능에 따른 쓰루풋을 계산해서
- 현재와 미래의 리퀘스트 수를 고려해서
- GPU를 골라서 서버를 신청 한 다음
- 서버 정담당자/부담당자를 지정하고(둘 다 내가?)
- 서버신청서를 쓰고...
- 서버가 나오면
- 모니터링/알람 기준을 설정하고
- 백업스케쥴 정하고
- 비상연락망에 이름도 올리고
- 서버에다가 CUDA설치하고
- 패키지 업데이트 정책 정하고
- 보안설정도 내가 해야 하나..?
- 선생님 저는 안 하고 싶어요

귀찮음

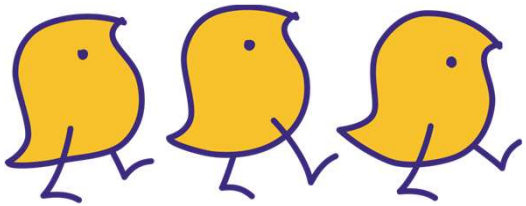
어떻게 하면 내가 일을 안 할 수 있을까요?

- 남이 해주면 됩니다.

어떻게 하면 내가 일을 안 할 수 있을까요?

2017 당시 파이썬을 지원하는 서비스는

2.7



Lambda에 모델을 올려봅시다

LangCon 2019

AWS Lambda Deployment Limits

Item	Default Limit
Lambda function deployment package size (compressed .zip/.jar file)	50 MB
Total size of all the deployment packages that can be uploaded per region	75 GB
Size of code/dependencies that you can zip into a deployment package (uncompressed .zip/.jar size). **Note** Each Lambda function receives an additional 500MB of non-persistent disk space in its own `/tmp` directory. The `/tmp` directory can be used for loading additional resources like dependency libraries or data sets during function initialization.	250 MB

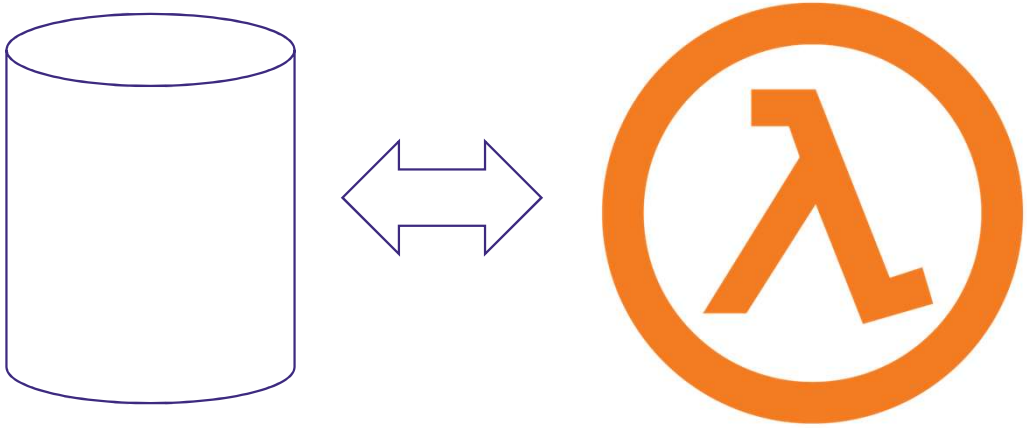
50MB 패키지를 만들어봅시다

- 그런데 keras+theano를 압축하면 70메가
- 테스트코드 삭제하면 겨우겨우 50메가 아래로 내려감
- 모델은 s3에서 /tmp에 받아 쓰면 된대!
- 아 근데 konlpy를 넣을 공간이 없네...

사실은 옛날에도

호출 페이로드(요청 및 응답)	6 MB(동기식) 256 KB(비동기식)
배포 패키지 크기	50 MB(직접 업로드용 압축 파일) 250 MB(계층을 포함해 압축 해제됨) 3 MB(콘솔 편집기)
테스트 이벤트(콘솔 편집기)	10
/tmp 디렉터리 스토리지	512 MB

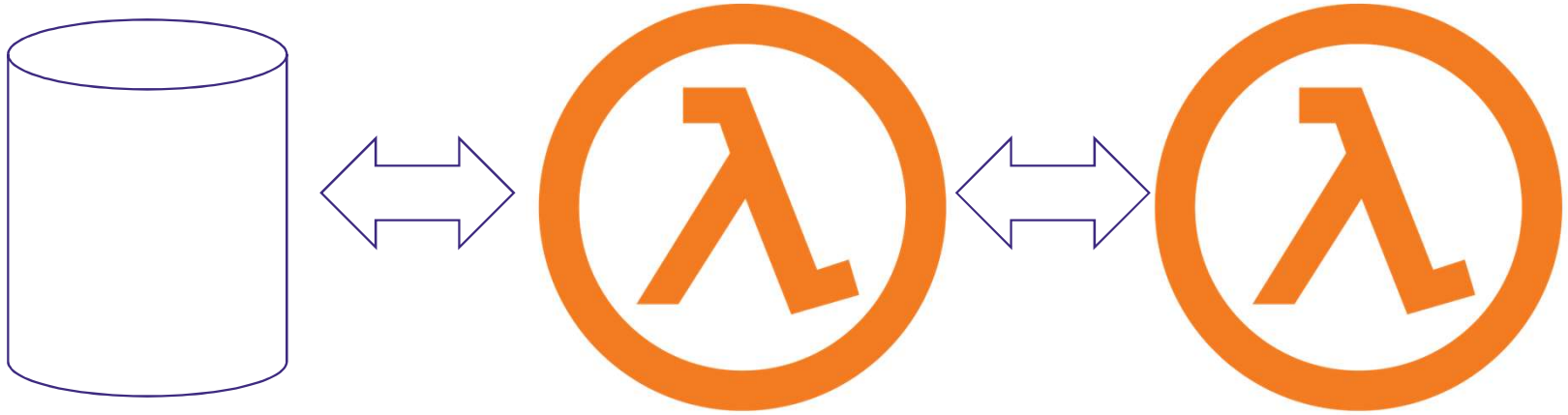
그런데 konlpy를 올릴 공간이 없어요

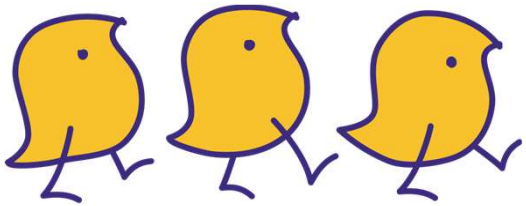


SHOP(>1k+)

keras

이렇게.....?





자연어 처리를 하기 위해 자연어 처리를 덜 해봅시다

LangCon 2019

자연어 처리를 덜 해봅시다

Character-level Convolutional Networks for Text Classification*

Xiang Zhang **Junbo Zhao** **Yann LeCun**
Courant Institute of Mathematical Sciences, New York University
719 Broadway, 12th Floor, New York, NY 10003
{xiang, junbo.zhao, yann}@cs.nyu.edu

Abstract

자연어 처리를 덜 해봅시다

Character-Aware Neural Language Models

Yoon Kim[†] **Yacine Jernite**[*] **David Sontag**[*] **Alexander M. Rush**[†]

[†]School of Engineering and Applied Sciences
Harvard University
{yoonkim,srush}@seas.harvard.edu

[*]Courant Institute of Mathematical Sciences
New York University
{jernite,dsontag}@cs.nyu.edu

No Konlpy

Abstract

We describe a simple neural language model that relies only on character-level inputs. Predictions are still made at the word-level. Our model employs a convolutional neural network (CNN) and a highway network over characters, whose output is given to a long short-term memory (LSTM) recurrent neural network language model (RNN-LM). On the English Penn Treebank the model is on par with the existing state-of-the-art despite having 60% fewer parameters.

While NLMs have been shown to outperform count-based n-gram language models (Mikolov et al. 2011), they are blind to subword information (e.g. morphemes). For example, they do not know, a priori, that *eventful*, *eventfully*, *uneventful*, and *uneventfully* should have structurally related embeddings in the vector space. Embeddings of rare words can thus be poorly estimated, leading to high perplexities for rare words (and words surrounding them). This is especially problematic in morphologically rich languages with long-tailed frequency distributions or domains with dynamic

자연어 처리를 덜 했더니 모델 성능이 덜 좋아

데이터 모으기가 힘들어요?

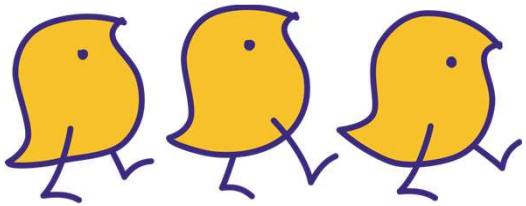
있는 데이터를 써봅시다

있는 데이터 써보기

Dataset	Classes	Train Samples	Test S
AG's News	4	120,000	7,
Sogou News	5	450,000	60
DBPedia	14	560,000	70
Yelp Review Polarity	2	560,000	38
Yelp Review Full	5	650,000	50
Yahoo! Answers	10	1,400,000	60
Amazon Review Full	5	3,000,000	650
Amazon Review Polarity	2	3,600,000	400

자연어 처리를 덜 했더니 모델 성능이 덜 좋아

기존	Neg	Neg	-	Pos	-
바뀜	1	2	3	4	5

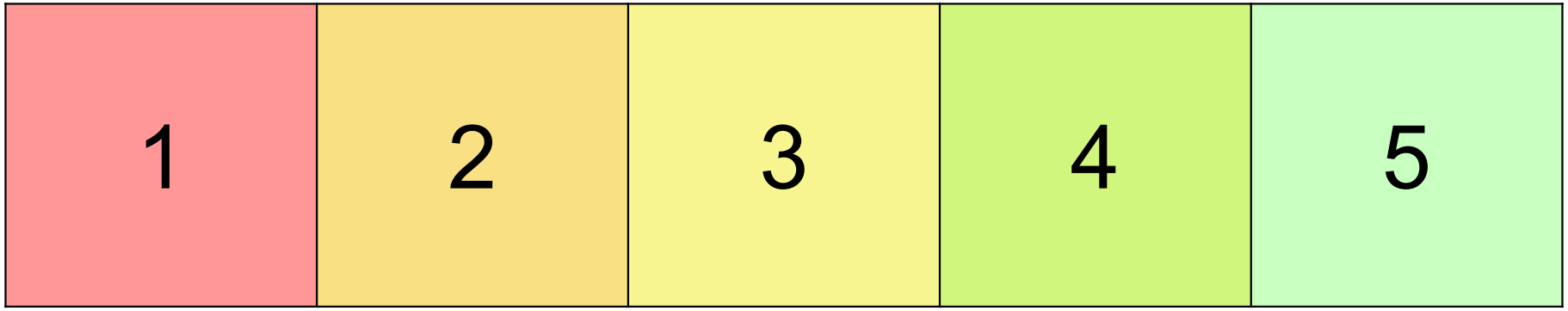


별점이란 무엇인가 물어라

LangCon 2019

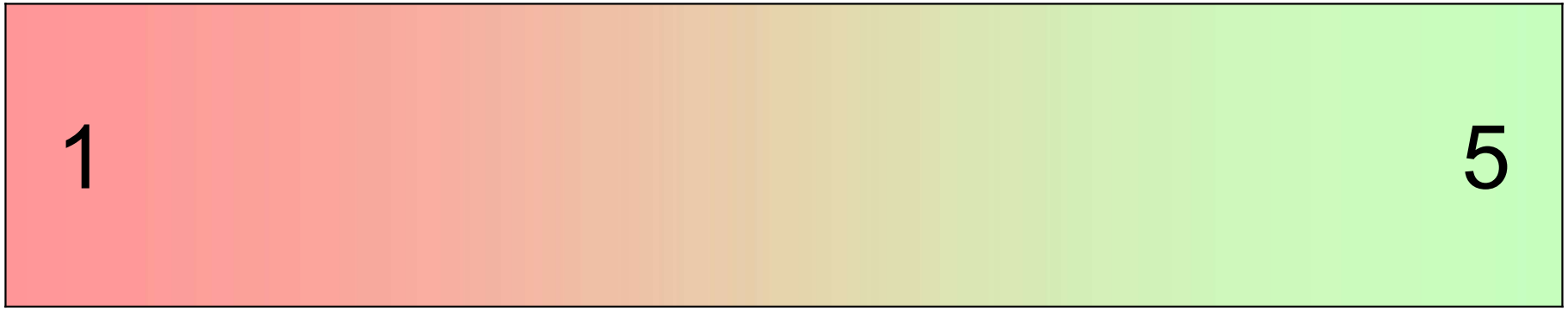
어라 근데... 이거..... 분류.....

- ["1점", "2점", "3점", "4점", "5점"]
- Class???



어라 근데... 이거..... 분류.....?

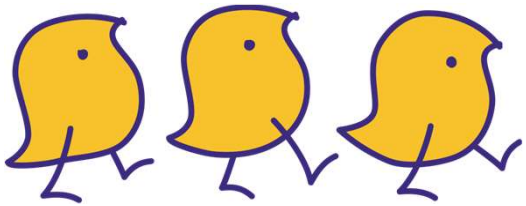
- 1-5
- Continuous!



프로젝트 목적(초심은 예전에 잃음)

별점기입 미스 검출을 통해 수동댓글 필요여부를 예측한다

- 유저의
- 유저의
- 리뷰에
- 점수를 준다



오늘밤에도 별점이 리뷰에 스치운다

LangCon 2019

머신러닝 모델의 변화

- 수동댓글이 필요한 리뷰를 분류하는 모델
- 텍스트만 봐서는 별점이 낮을 것 같은 리뷰를 분류하는 모델
- 텍스트만 보고 점수를 예측해주는 모델

머신러닝 모델의 변화(상세)

- Twitter pos — TF-IDF — SVM
- Twitter pos — Keras(LSTM)
- Preprocessing — Mecab pos(dict) — w2v — Keras(LSTM)
- Preprocessing — Keras(CNN)

프로젝트 진행 목적의 변화

수동댓글 대상을 알려주자(나머지는 자동으로 처리하게)

리뷰들 중에서 뭐가 컴플레인인지 알려주자
별점 5점인 리뷰들 중에서도
뭐가 진짜 좋은 리뷰인지 알려주자

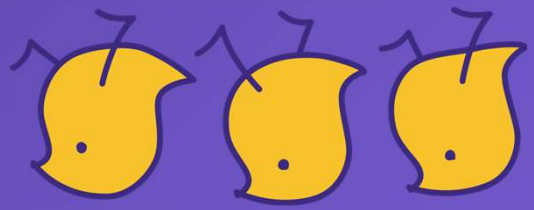
데이터의 변화

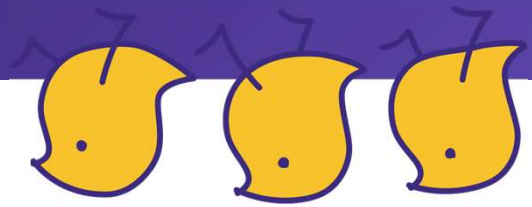
- 반 수동으로 수집한, 컴플레인리뷰와 일반리뷰의 데이터
- 1,2점 리뷰와 4점 리뷰의 데이터
- 리뷰 전체 데이터

기타

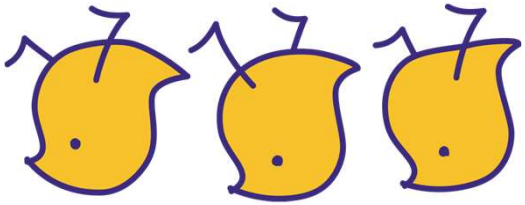
- 처음에 Binary Classification으로 진행했더니,
  그렇게 할 필요가 없어져서도 그 모델로 계속 진행하게 됨
- 서버관리하기 귀찮아서 serverless로 했음
- 그러다보니 word-embedding대신 char-embedding을 씀

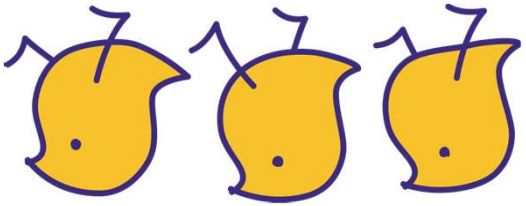
Q

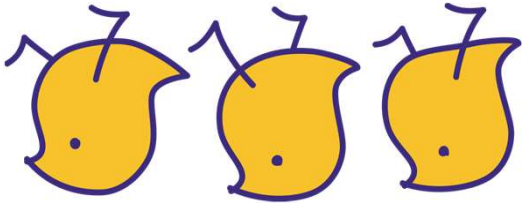




A







LangCon 2019