

boostcamp AITECH NLP-09

지구코딩실

BEHIND THE RE



Yongwoo Song T4111
ywsong.dev@kakao.com

조금 다른 관점에서,

지구코딩실의 부캠에서 살아남기

1. 지구코딩실은 어떤 팀인가요?

2. 서비스 그리고 Creator

3. 지구코딩실이 개발하는 법

0. 앞으로의 여정

PART 1.

지구코딩실은 어떤 팀인가요?



지구코딩실은 어떤 팀인가요?

5명 모두 개성, MBTI, 전공도 다르지만,

ML를 활용한 End-to-End 서비스 개발이라는
하나의 목표를 향해 모인 팀입니다.

이를 통해 어떻게 세상에 가치 있는 서비스를 만들 수 있을까
다양한 고민을 하고 있습니다.

지구코딩실의 Core Value

1. 바보 같은 질문을 하자
2. 기억보다는 기록을
3. Follower 보다는 Creator

오늘 이야기 해볼 것

1. 바보같은 질문을 하자
2. 기억보다는 기록을
3. Follower 보다는 Creator

Creator의 관점으로 이번 대회를 접근해보자.

PART 2.

Follower 보다는 Creator, 그리고 서비스



지구코딩실이 이루고 싶은 목표

End-to-End ML 서비스를 만들어 보자!

그리고 이번 대회를 ML 서비스 개발의 관점으로 접근해보자

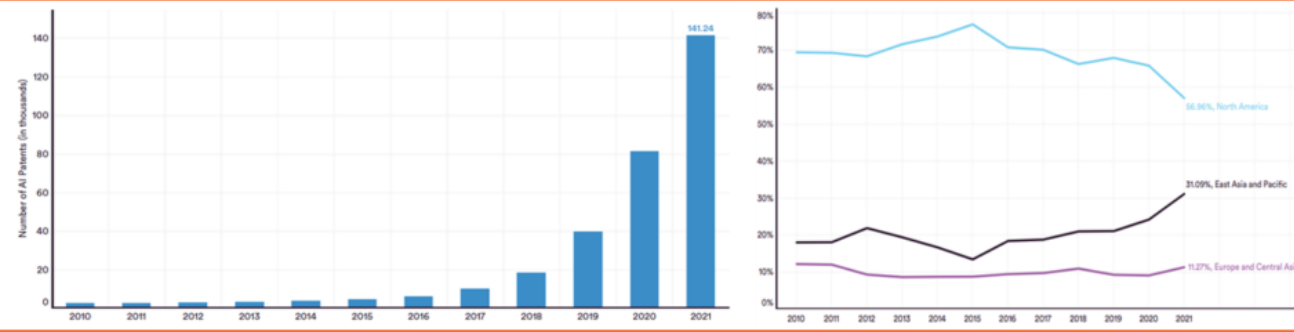
ML

서비스

시장은
우리를
기다려 주지
않는다.

지구코딩실이 해결하고자 하는 문제

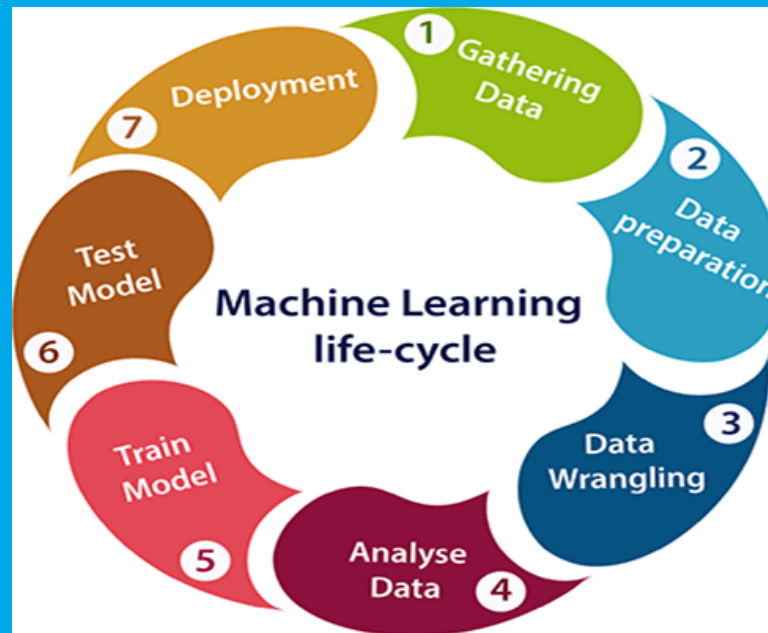
〈 연도별 AI 특허 출원(단위 : 천 개) 및 지역별 AI 특허 등록 비중 〉



자료 : Stanford HAI, Artificial Intelligence Index Report 2022

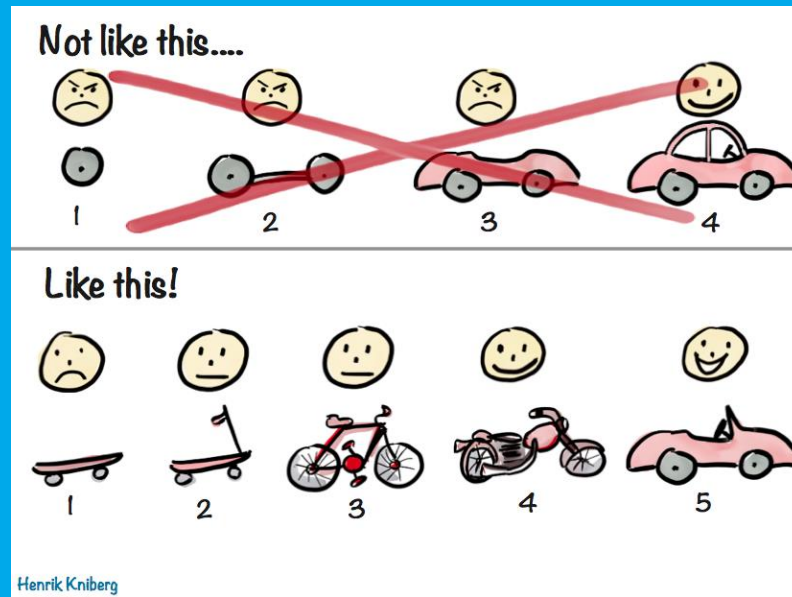
최근 하루가 멀다하고
수많은 논문, 모델, 데이터, 니즈가
쏟아지는 시대에 살고 있다.

지구 코딩실이 해결하고자 하는 문제



그렇다면 고객의 피드백에 대해
빠르게 대응하는 것이 무엇보다 중요

지구 코딩실이 해결하고자 하는 문제



차근차근 처음부터 개발하면 시장의 속도를 따라잡을 수 없다
필수요소를 빠르게 개발하고, 지속적으로 발전시켜야 한다

지구 코딩실이 해결하고자 하는 문제

”개발자는 늘 최악의 경우를 고려해야 한다”

ML 엔지니어에게 최악의 경우는?

“몇 개월, 몇 년 동안 피땀 들여 개발한 서비스가 시장에서 필요 없어진 경우”

우리는 이 변화와 불확실성에 대응할 수 있어야 한다

질문!

**지금까지 다룬 개념을
관통하는 하나의 키워드는?**

애자일

Agile

지코실 스타일 애자일

지구코딩실 Core Value

1. 바보같은 질문을 하자
2. 기억보다는 기록을
3. Follower 보다는 Creator

애자일에 필요한 요소들을 Creator로써 직접 개발해보자!
그리고 RE 대회는 이를 증명할 첫번째 장!

PART 3.

지코실이 애자일하게 개발하는 법



지코실 스타일 애자일을 위해

베이스라인 아키텍처 개선
지코실 스타일 Git-Flow
결과분석 대시보드
코드 리뷰
코드 테스트
Custom MLFlow
개발의 문서화
github action 활용
데일리 스크럼 룰
노션 템플릿 제작
커밋, PR 컨벤션
주별 TODO 정리

·
·
·

지코실 스타일 애자일을 위해

베이스라인 아키텍처 개선

지코실 스타일 Git-Flow

결과분석 대시보드

코드 리뷰

코드 테스트

Custom MLFlow

개발의 문서화

github action 활용

데일리 스크럼 룰

노션 템플릿 제작

커밋, PR 컨벤션

주별 TODO 정리

·
·
·

베이스라인 아키텍처 개선

용우님 이번에 새로 구현한 전처리 함수 어디에 있나요?

아마 train.py에 있을걸요..?

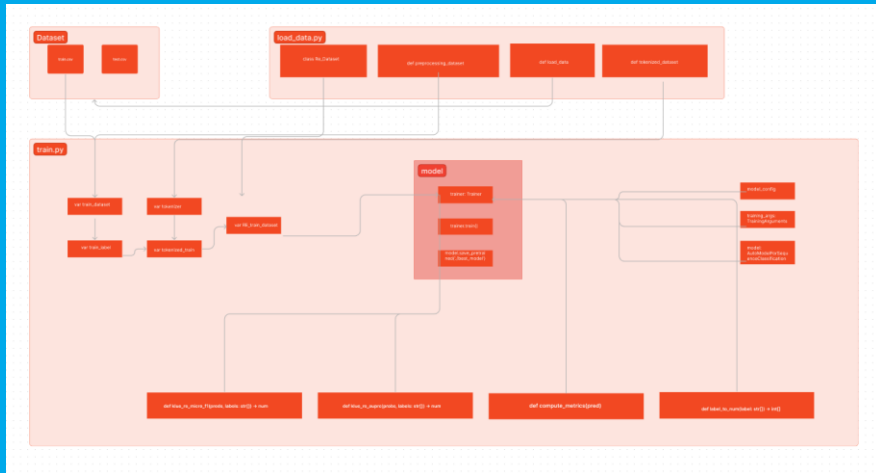
한번 확인해보고 답변 드릴게요!

넵넵 ☺

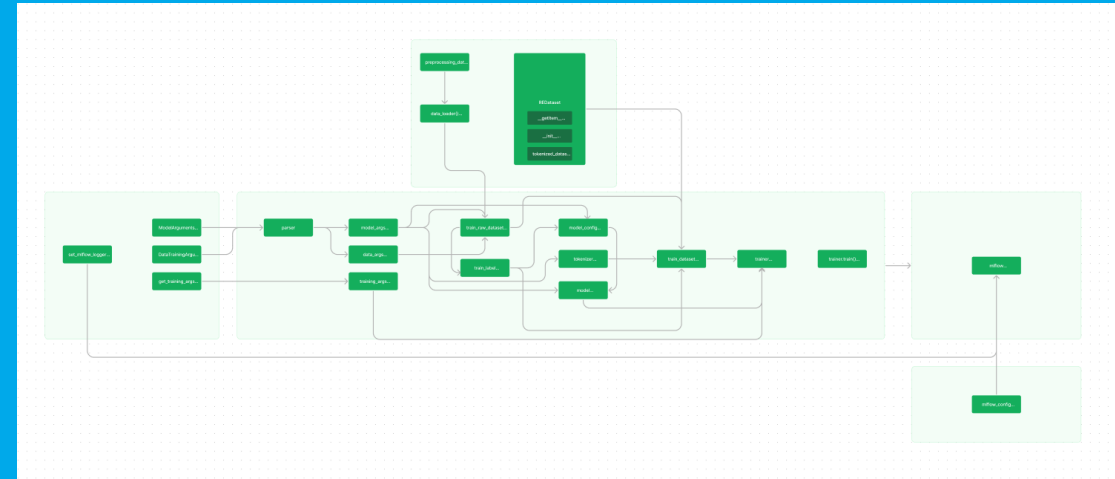
지금 확인해보니 model.py 742번째 줄에 있어요!

베이스라인 아키텍처 개선

좋은 코드란
당연히 있어야 할 위치에서
당연히 해야 할 일을 한다.



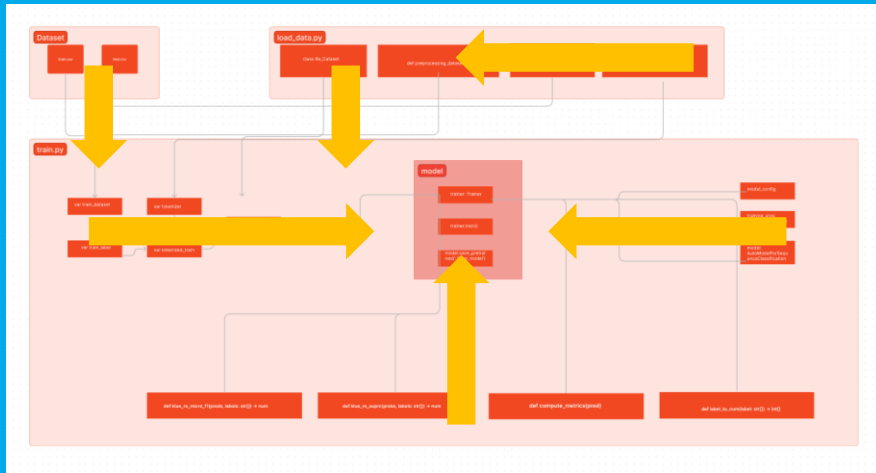
이전 베이스라인 아키텍처



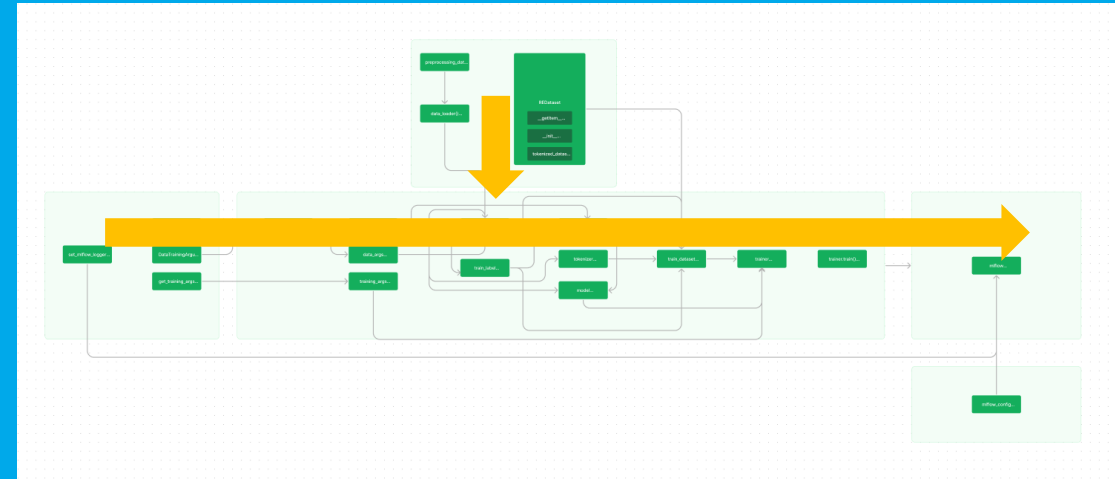
새로 개선한 아키텍처

베이스라인 아키텍처 개선

좋은 코드란
당연히 있어야 할 위치에서
당연히 해야 할 일을 한다.



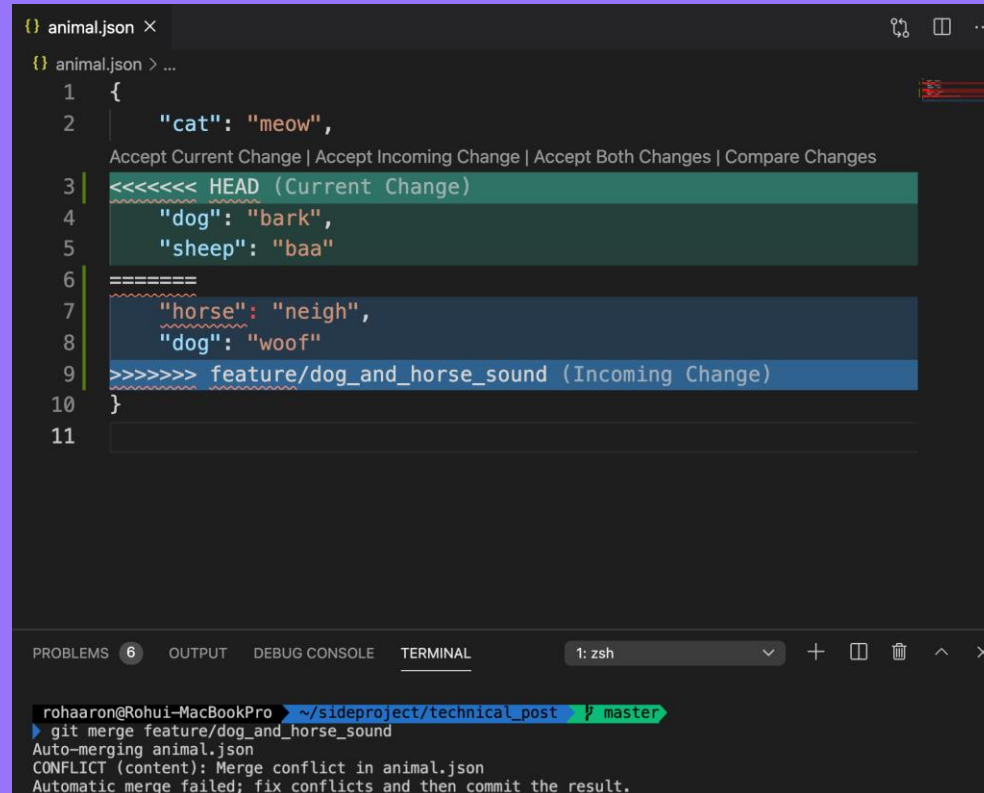
이전 베이스라인 아키텍처



새로 개선한 아키텍처

확장성, 재사용성 UP

지코실 스타일 Git Flow



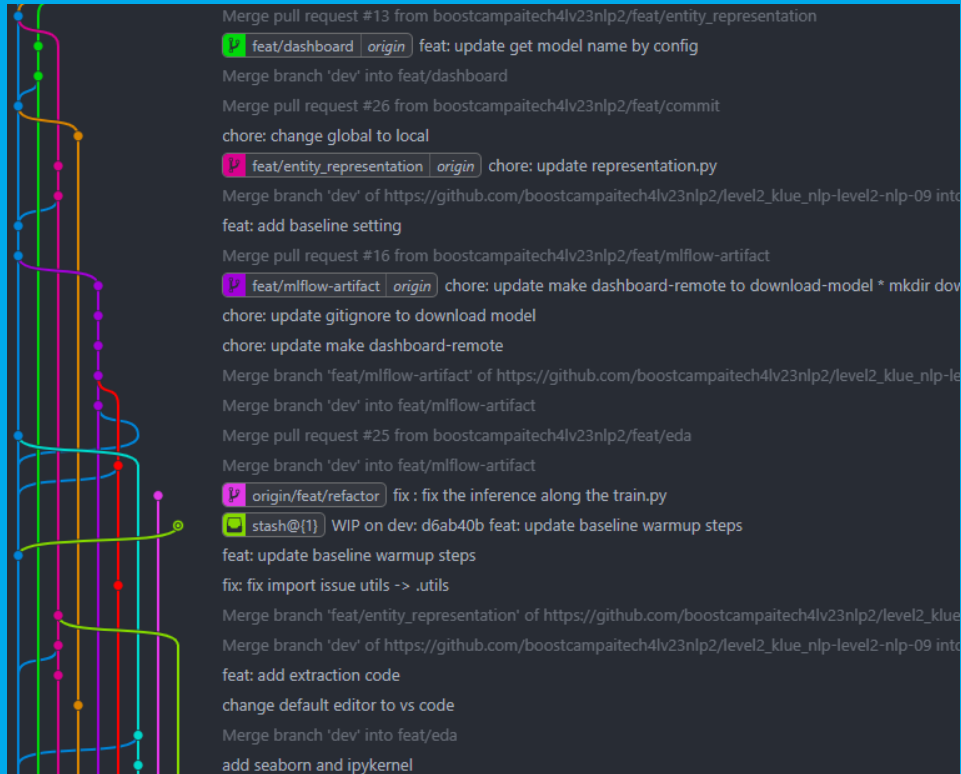
```
{} animal.json ×
{} animal.json > ...
1 {
2   "cat": "meow",
3   <<<<<< HEAD (Current Change)
4     "dog": "bark",
5     "sheep": "baa"
6   =====
7     "horse": "neigh",
8     "dog": "woof"
9   >>>>>> feature/dog_and_horse_sound (Incoming Change)
10 }
11

PROBLEMS 6 OUTPUT DEBUG CONSOLE TERMINAL 1: zsh
rohaaron@Rohui-MacBookPro ~/sideproject/technical_post master
▶ git merge feature/dog_and_horse_sound
Auto-merging animal.json
CONFLICT (content): Merge conflict in animal.json
Automatic merge failed; fix conflicts and then commit the result.
```

안돼 Merge 받아줄 생각 없어 빨리 돌아가

지코실 스타일 Git Flow

170여개의 커밋, 27개의 PR 수행
크리티컬 버전 이슈 0건



fix: wrong ^ resulted in [UNK] token	ghlrobin committed 5 days ago ✓	ebafe2d	<>
Merge branch 'dev' into feat/japanese	kyc3492 committed 5 days ago ✓	869ef8	<>
Merge pull request #31 from boostcampaitch4lv23nlp2/feat/mlflow-tran...	kyc3492 committed 5 days ago	Verified 253d64c	<>
feat: make japanese to korean, not caring about real meanings	kyc3492 committed 6 days ago ✓	d649401	<>
Commits on Nov 23, 2022			
Merge branch 'dev' into feat/mlflow-transformer	kyc3492 committed 6 days ago ✓	0e74eb5	<>
Merge pull request #32 from boostcampaitch4lv23nlp2/feat/preprocess	FacerAin committed 6 days ago	Verified 2ac4275	<>
fix: entity representation none sep blank	FacerAin committed 6 days ago ✓	e579f8b	<>
chore: remove check-code workflow	FacerAin committed 6 days ago ✓	c83f919	<>
chore: remove check-code workflow	FacerAin committed 6 days ago	459a0f3	<>

코드 리뷰

모든 코드에 대해 리뷰 수행
함께 성장하고, 더 좋은 코드를 쓰는 지름길

Merged

Refactoring baseline code #12
FacerAin merged 19 commits into dev from feat/refactor 11 days ago

src/train.py

Outdated

Hide resolved

101 - num_label.append(dict_label_to_num[v])
102 -
103 - return num_label
22 + from data_loader.load_data import REDataset, load_data

FacerAin 11 days ago

위 코드도 좋지만, init.py를 활용하여 아래와 같이 import를 시도해 볼 수 있을 것 같아요!
from data_loader import REDataset, load_data
참고
https://github.com/victoresque/pytorch-template/blob/master/logger/__init__.py

jinyeongAN 11 days ago

Author

이 점이 굉장히 좋다 생각해 수정을 했는데
F403 'from .util import *' used; unable to detect undefined names 에러가 발생했습니다.
따라서
setup.cfg 에 F403을 추가하고, make setup 을 했는데 여전히 lint를 통과 못합니다.

FacerAin 11 days ago

이부분은 제 환경에서 확인을 해보겠습니다.

Unresolve conversation

FacerAin marked this conversation as resolved.

Merged

Feat/preprocess add translation and replace #32
FacerAin merged 9 commits into dev from feat/preprocess 6 days ago

wbin0718 approved these changes 6 days ago

View changes

wbin0718 left a comment

LGTM!!!

src/utils/representation.py

33 + entity_word = entity_word.replace(" ", "").strip()
34 + entity_type = entity_type.replace(" ", "").strip()
35 +
36 + entity_dict = {

wbin0718 6 days ago

딕셔너리 형태로 묶어서 반환하는 것 좋은 의견인 것 같습니다!

wbin0718 6 days ago • edited

representation.py entity_representation 함수 78번 패 줄 첫 번째 [SEP]가 띄어쓰기가 되어있지 않아서 이를 앞뒤로 띄어쓰기로 수정 해 주시면 감사하겠습니다!
" [SEP]" -> " [SEP] "

FacerAin 6 days ago

Author

확인했습니다! 수정사항에 반영하도록 하겠습니다!

이미 끝난 일에 대한 반복적인 작업들

우빈님 이번에 새로 추가한 기능 어떻게 사용하나요?

강혁님 이번에 리더보드 모델 파일 보내주실 수 있나요?

연철님 혹시 저번에 추가한 기능에서 문제 없는게 확실할까요?

진명님 혹시 이번에 공부하신 F1 개념 다시 설명해주실 수 있나요?

개발의 문서화

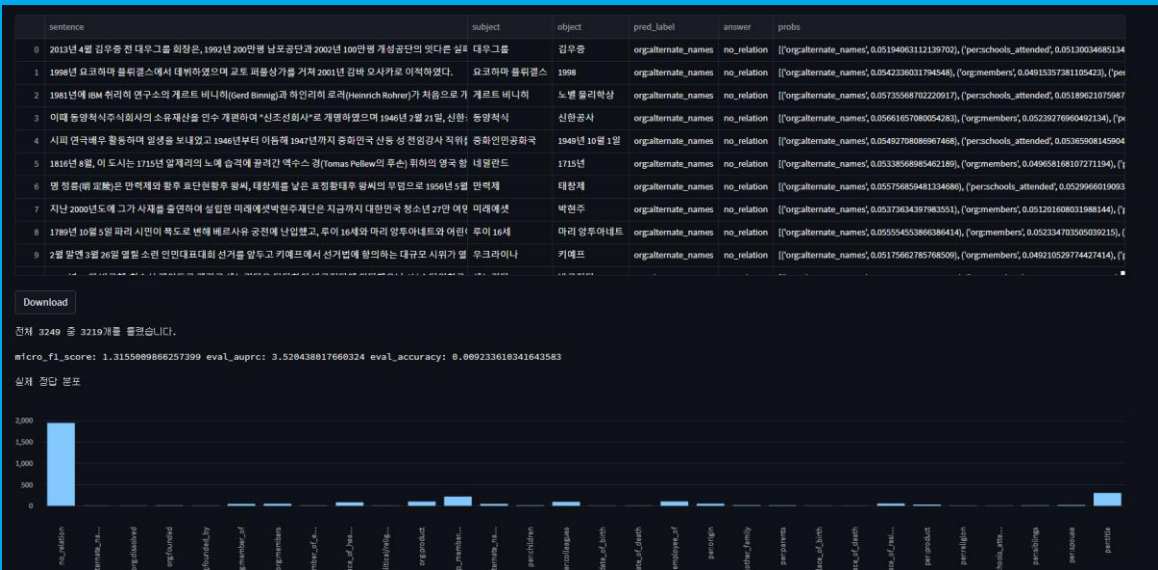
따로 시간을 내서 쓰는 글만이 문서가 아니다.
커밋 메시지, PR, 주석, 심지어 코드도 하나의 문서가 될 수 있다.

Earth Coding Lab (ECL) / ECL Secret Notes			
Type	Name	Date	
Dev Docs	Loss와 Accuracy가 동시에 증가한다?	2022년 11월 25일	
Dev Docs	Phase2 결과 분석	2022년 11월 23일	
Dev Docs	학습 데이터를 손질해보자	2022년 11월 21일	
Dev Docs	PHASE 2 실험 보고서	2022년 11월 17일	
Dev Docs	F1 Score: Deep Dive	2022년 11월 16일	열기
Dev Docs	NER 대회 EDA	2022년 11월 15일	
Dev Docs	처음부터 시작하는 MLflow + SFTP 서버	2022년 11월 13일	
Dev Docs	baseline + LSTM Layer		
Dev Docs	baseline + GRU Layer		2
Dev Docs	PHASE 3 실험 예측		
Dev Docs	train과 validation 힘하게 분리하는 법		
Dev Docs	ECL Style 협업 후기		
Dev Docs	RE Phase1 Research Report		
Dev Docs	RE Phase1 Technical Report		
Dev Docs	Area Under Precision-Recall Curve (AUPR)		
Dev Docs	RE 기초대회 베이스라인 리팩토링		
Dev Docs	알잘딱깔센하게 논문 읽는 팁		
Dev Docs	허깅페이스 컨트리뷰션 후기		
Dev Docs	RE 논문 & 모델 LIST-UP		
Dev Docs	Github Guide		
Dev Docs	랩업 리포트 오픈 피드백		
Dev Docs	연접 질문 리스트		
Dev Docs	Huggingface Transformer Deep Dive		
Dev Docs	지구코딩실을 소개합니다.		
Dev Docs	Pythonic Code Guide		
Dev Docs	Commit Message Guide		
Dev Docs	Code Review Guide		

Code Issues 6 Pull requests 1 Actions Projects Security Insights Settings			
Filters	is:issue is:open	Labels 9	Milestones 0
New issue			
6 Open 6 Closed Author Label Projects Milestones Assignee Sort			
MLflow - Huggingface enhancement 1			
#30 opened 7 days ago by kyc3492			
ner_label 함수 작성 enhancement 1			
#28 opened 7 days ago by wbin0718			
Utils 구조 정리 enhancement			
#21 opened 11 days ago by FacerAin			
VScode port forwarding to use dashboard app good first issue			
#15 opened 11 days ago by FacerAin			
train 전처리 방안 enhancement 3 tasks			
#8 opened 12 days ago by FacerAin			
commit message through VS code editor enhancement			
#2 opened 14 days ago by ghlobin			

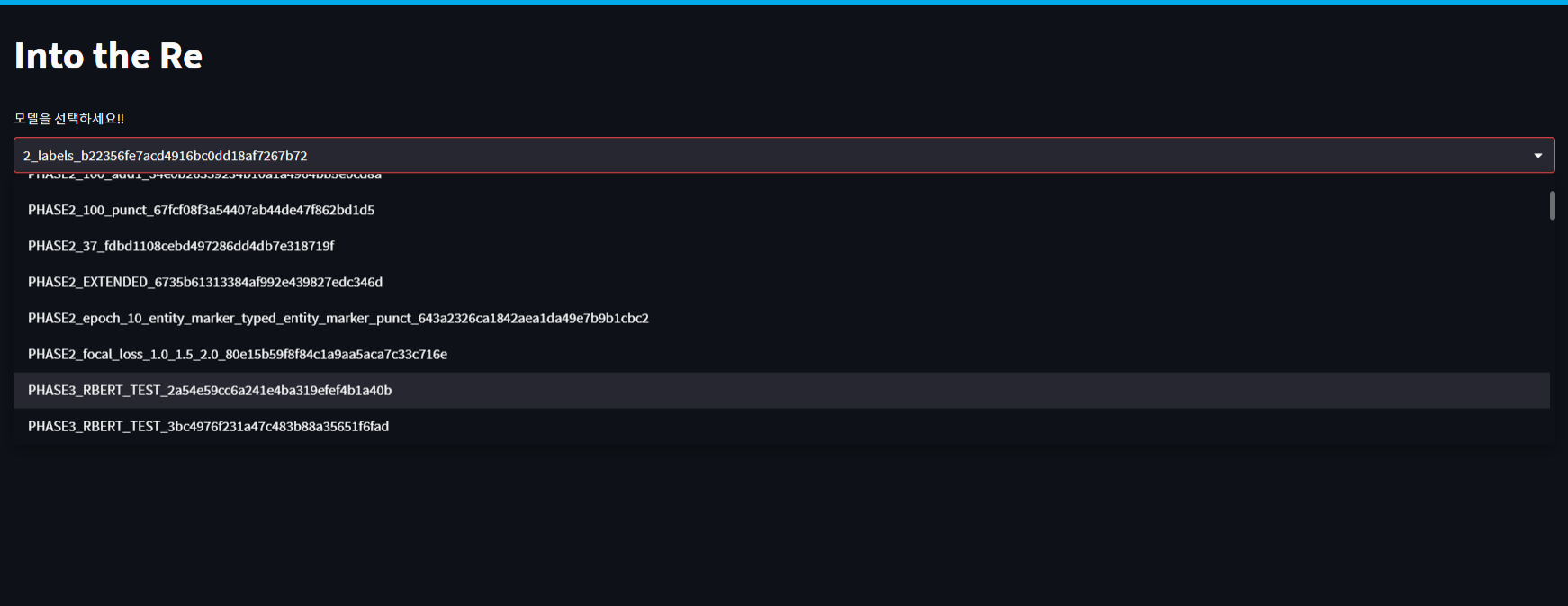
결과분석 대시보드 앱

모델의 성능과 인사이트 (예측 라벨 분포, 틀린 문장 예시)
등을 한눈에 보기 쉽게 정리해보자.



Custom MLFlow

언제 어디서나 누구나 모델을 바로 사용할 수 있도록
SFTP로 모델 파일을 관리하자



테스트 코드 작성

테스트 코드를 짤 수 없다면,
그것은 아직 코드를 100% 이해하지 못한 것이다,
나의 구현이 옳음을 증명해보자.

```
tests/test_preprocess.py::PreprocessTester::test_bracket_symbol FAILED
tests/test_preprocess.py::PreprocessTester::test_replace_symbol PASSED
tests/test_representation.py::RepresentationTester::test_chinese PASSED
tests/test_representation.py::RepresentationTester::test_custom FAILED
tests/test_representation.py::RepresentationTester::test_japanese PASSED
tests/test_representation.py::RepresentationTester::test_none PASSED
tests/test_representation.py::RepresentationTester::test_replace PASSED
```

```
class RepresentationTester(unittest.TestCase):
    def test_none(self):
        for example_object, answer in zip(test_objects, none_answers):
            sentence, subject, object = example_object
            generate_text = representation(subject, object, sentence, entity_method=None)
            self.assertEqual(generate_text, answer)

    def test_chinese(self):
        for example_object, answer in zip(test_objects, chinese_answers):
            sentence, subject, object = example_object
            generate_text = representation(
                subject, object, sentence, entity_method=None, translation_methods=["chinese"]
            )
            self.assertEqual(generate_text, answer)

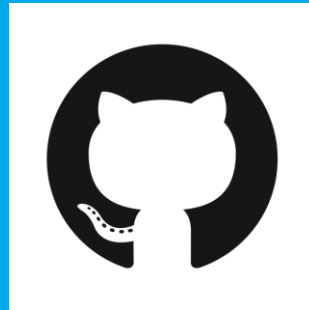
    def test_replace(self):
        for example_object, answer in zip(test_objects, replace_symbols_answers):
            sentence, subject, object = example_object
            generate_text = representation(
                subject, object, sentence, entity_method=None, translation_methods=[None], is_replace=False
            )
            self.assertEqual(generate_text, answer)
```


파이프라인 반자동화



문서화
Task 정리

코드 포매팅
코드 테스트
코드 리뷰
버전관리



코드 리뷰
리마인더



코드 개발

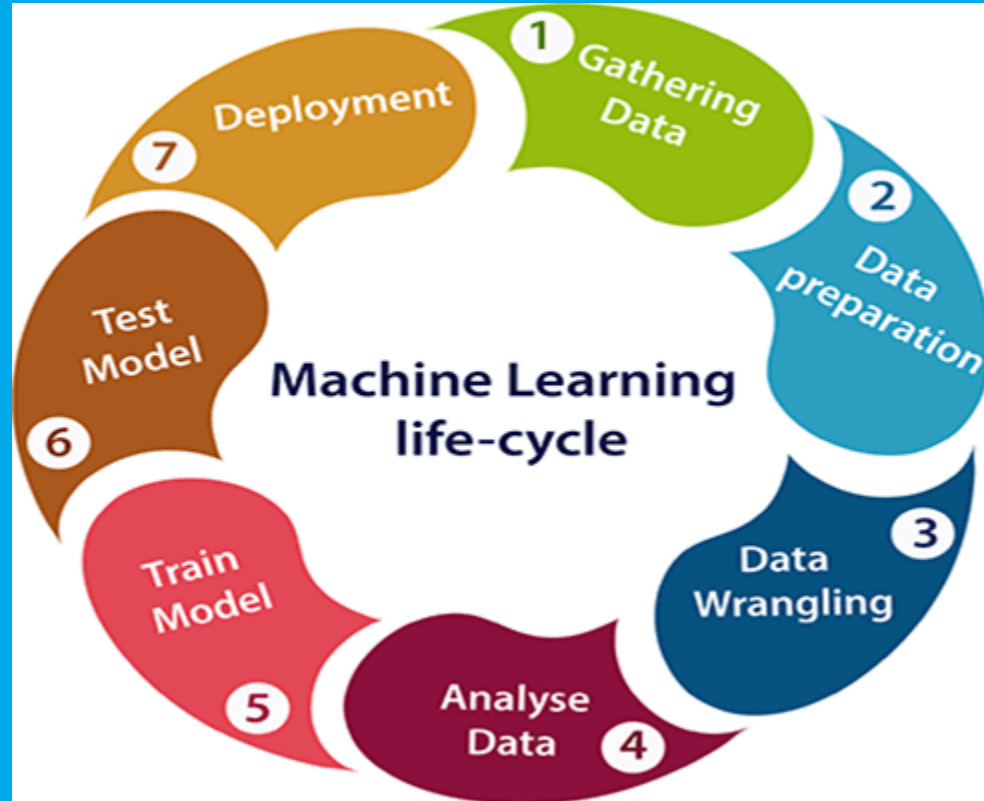
실험 기록



결과 분석
모델 파일
관리

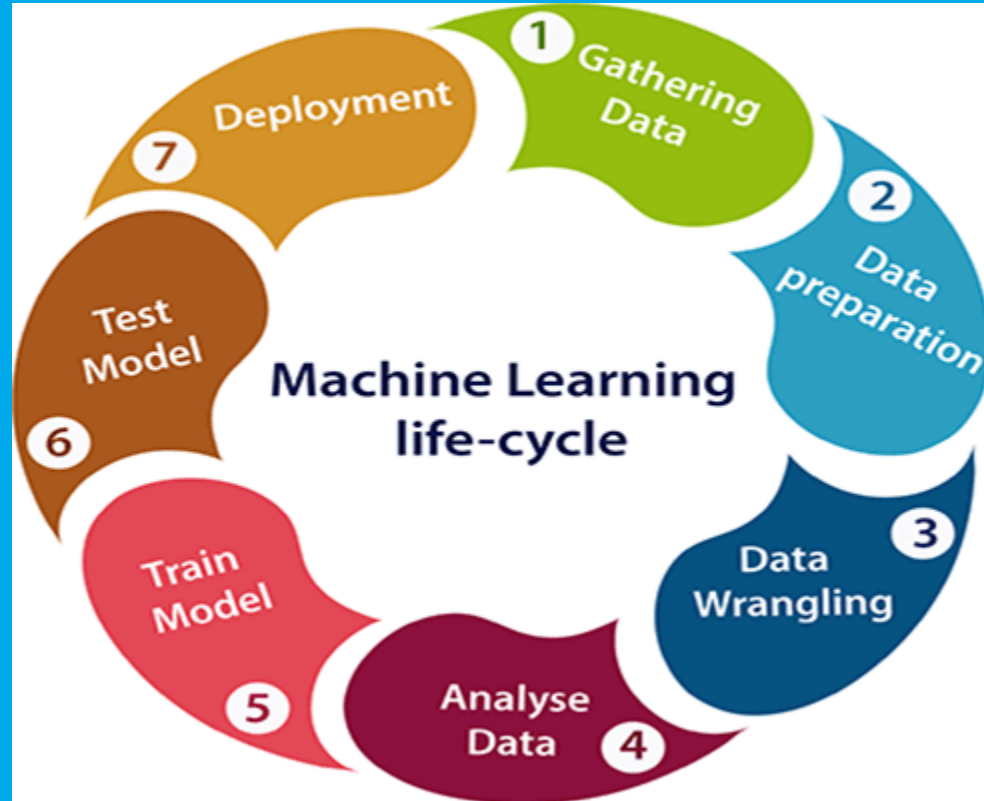


지구코딩실에게 애자일이란



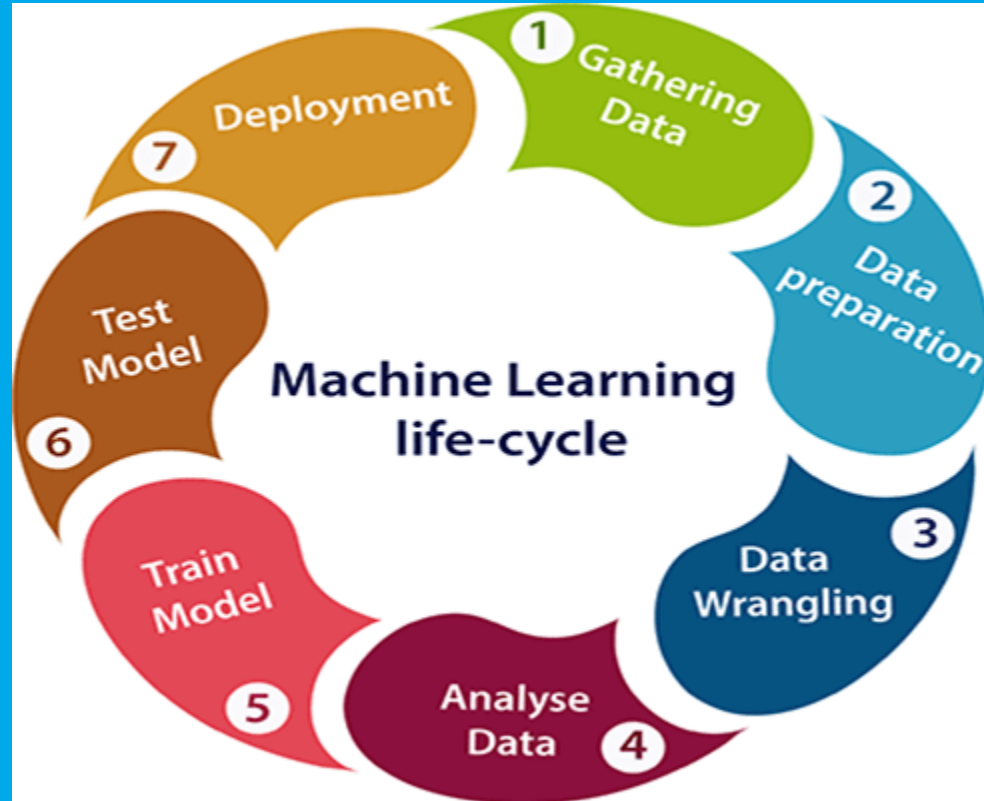
Build your first system quickly and then iterate!
- Andrew Ng

지구코딩실에게 애자일이란

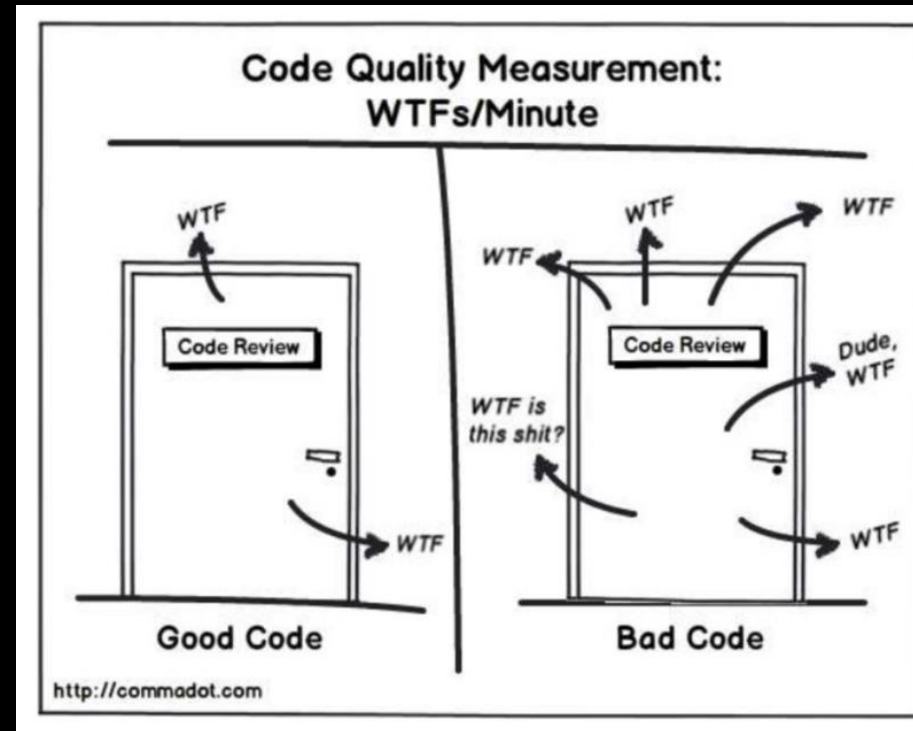


세상에 미래를 예측할 수 있는 개발자는 없다.
하지만 미래에 유연하게 대응할 수 있는 개발자는 있다.

지구코딩실에게 애자일이란



불필요한 잡음과 군더더기를 줄이자.
반복 프로세스를 최적화하자.
그럼 변화에 유연하게 대응할 수 있다.



Better Code = Better Team Work

PART 0.

앞으로의 지구코딩실 여정

앞으로의 지구코딩실의 여정

1. 실제 End User까지 배포를 목적으로 서비스 개발
2. 테스트 커버리지 50% 이상 달성
3. 파이프라인 및 내부 툴 오픈소스화
4. 캠퍼분들과 함께하는 주기적인 세미나 및 지식 공유

자세한 이야기가 더 궁금하다면

- RE 기초대회 베이스라인 리팩토링 [\[링크\]](#)
- Deep Dive: F1 Score [\[링크\]](#)
- 처음부터 시작하는 MLflow + SFTP 서버 세팅하기 [\[링크\]](#)
- train과 validation 힙하게 분리하는 법 [\[링크\]](#)
- 허깅페이스 컨트리뷰션 후기 [\[링크\]](#)
- 지코실 스타일 협업 후기 [\[링크\]](#)

Contact

- 이강혁 @ghlrobin
 - math@kakao.com
- 안진명 @jinmyeongAN
 - jinmyeong.an@gmail.com
- 강연철 @kyc3492
 - kyc3492@gmail.com
- 박우빈 @wbin0718
 - wbbin0718@gmail.com
- 송용우 @facerein
 - ywsong.dev@kakao.com

Reference

- [WTF per Minute – commandot.com](https://www.commandot.com/)
- [Life cycle of Machine Learning – Javatpoint](https://www.javatpoint.com/)
- [AI Index 2022 | Stanford HAI](https://aiindex.stanfordhai.org/)