DATA AI 분석 경진대회

- 매뉴얼 -

참가팀 : 사이냅소프트

참가 문제 : 한국어 과학기술 논문의 초록 문장분류

목차

결과물 구성

경로: /home/work/.data

파일명	설명
demo.ipynb	시연용 주피터 노트북
train.ipynb	학습용 주피터 노트북
demo.sh	시연용 셸 스크립트
abstract_classifier	결과 프로젝트
synap_manual	사이냅소프트 매뉴얼이 있는 디렉토리
synap_presentation	사이냅소프트 발표자료가 있는 디렉토리

결과 프로젝트 구조

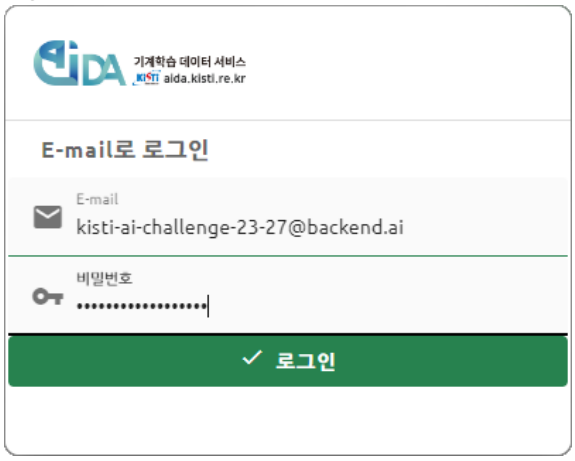
경로: /home/work/.data/abstract_classifier

디렉토리/파일명	설명
data	학습에 사용한 데이터들이 있는 디렉토리
train_config	학습 시 사용되는 config 파일들이 있는 디렉토리
outputs	학습 결과 모델이 저장되는 디렉토리
venv	Python 가상환경
train.py	학습 시 사용되는 Python 스크립트
train_korscideberta.py	korscideberta 학습 시 사용되는 스크립트
valid.py	학습 결과 모델 평가 시 사용되는 Python 스크립트
run_board.sh	Tensorboard 구동 시 사용하는 셸 스크립트
run_train.sh	학습 시 사용하는 셸 스크립트
README.md	프로젝트 환경 세팅에 대한 내용의 README 파일

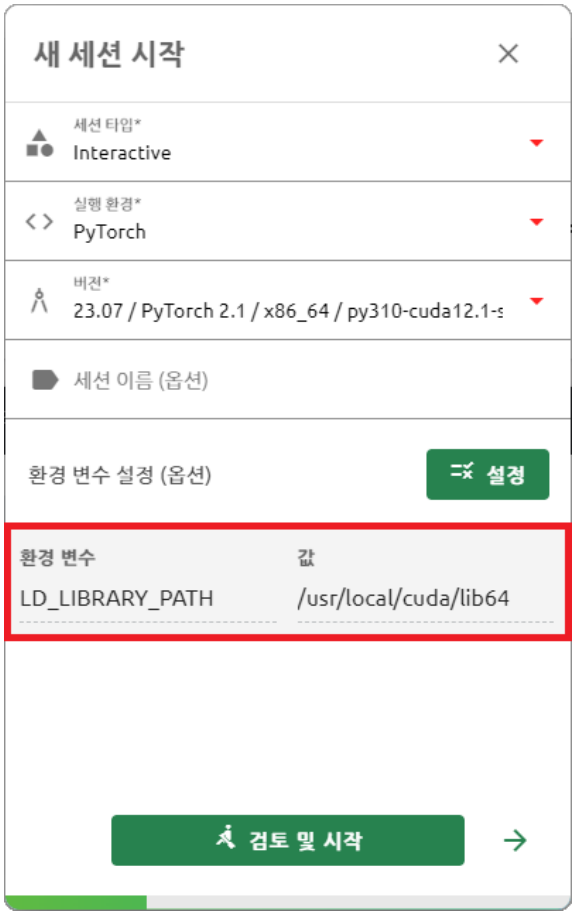
시연 방법

공통

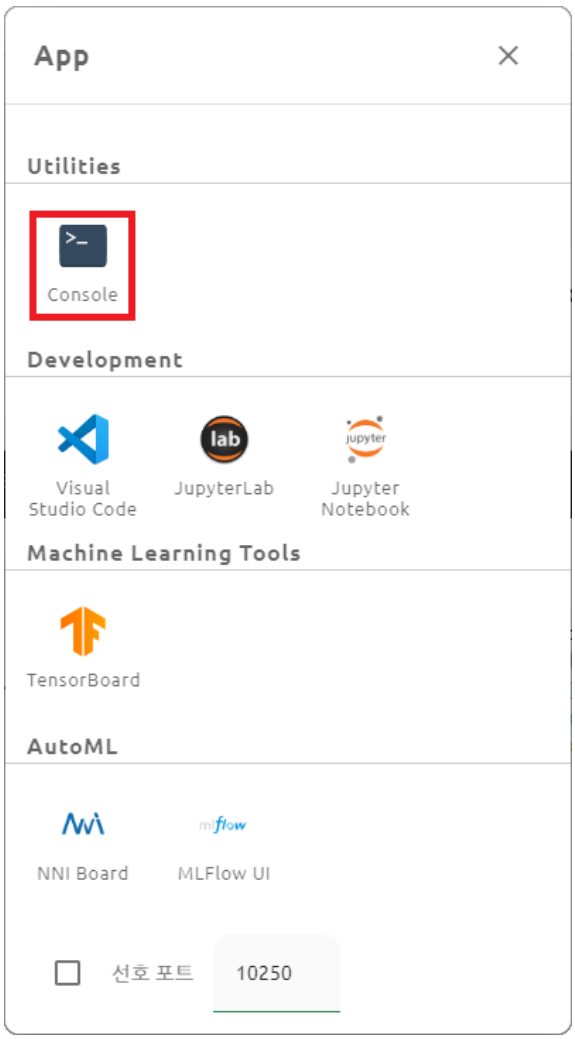
1. http://203.253.11.1152:8080 에 접속하여 로그인합니다.
(접속 정보 : kisti-ai-challenge-23-27@backend.ai / Kisti-ai-ch*!-9013)



2. 세션 페이지에서 '시작' 버튼을 클릭하여 팝업되는 창에서 환경 변수를 아래와 같이 설정한 후 세션을 시작합니다.
- 환경 변수: LD_LIBRARY_PATH
- 값: /usr/local/cuda/lib64



3. 세션이 정상적으로 실행되면 App 선택 창에서 'Console'을 선택합니다.

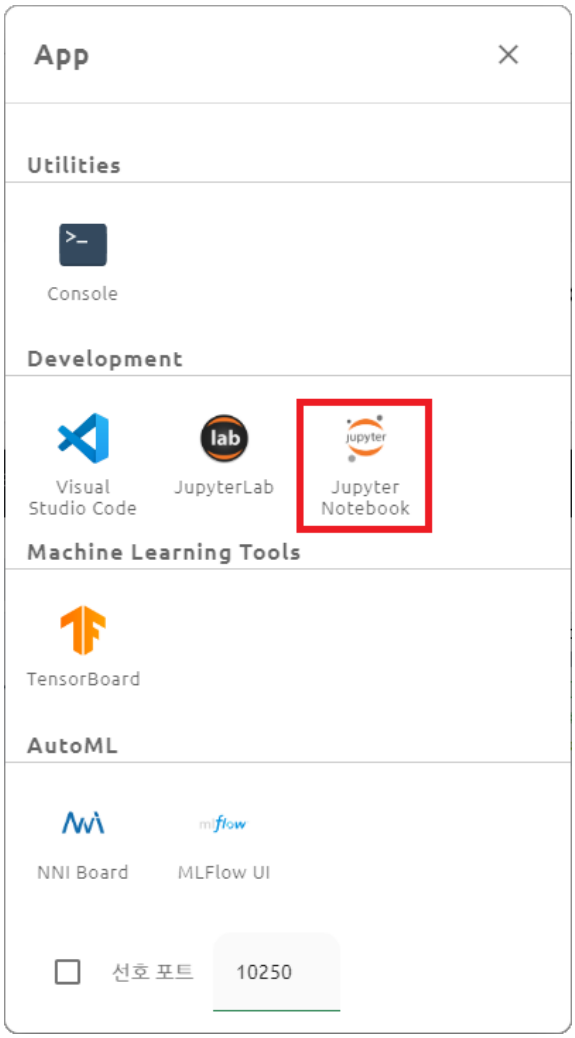


4. Console 에서 아래 명령어를 실행합니다.

```
$ cd /home/work/.data
$ source demo.sh
```

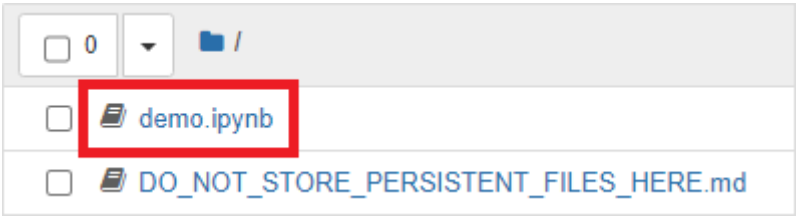
```
work@main1[T4L2Hejq-session]:~$ cd /home/work/.data
work@main1[T4L2Hejq-session]:~/.data$ source demo.sh
Installed kernelspec venv in /usr/local/share/jupyter/kernels/venv
work@main1[T4L2Hejq-session]:~/.data$
```

5. 세션 페이지로 가서 현재 실행중인 세션에 주피터 노트북을 실행합니다.

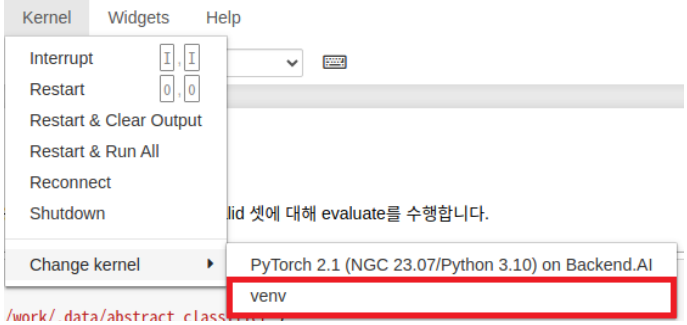


추론

1. 파일 목록 중 'demo.ipynb'를 클릭합니다.



2. 상단의 Kernel - Change kernel - venv 를 선택합니다.



3. 아래 코드가 작성된 셀을 실행하여 valid 데이터셋에 대한 추론 결과를 확인합니다.

```
import os
import sys
sys.path.append(‘/home/work/.data/abstract_classifier’)

from valid import evaluate_model

work_dir = ‘/home/work/.data/abstract_classifier’
results = {'0-10': None, '1-9': None, '2-8': None, '3-7': None}
```

```
model_path = os.path.join(work_dir, 'outputs', '0-10', 'checkpoint-15000')
evaluation_data_path = os.path.join(work_dir, 'data', 'abstract_all.json')
f1, acc = evaluate_model(model_path, evaluation_data_path)
results['0-10'] = (f1, acc)
```

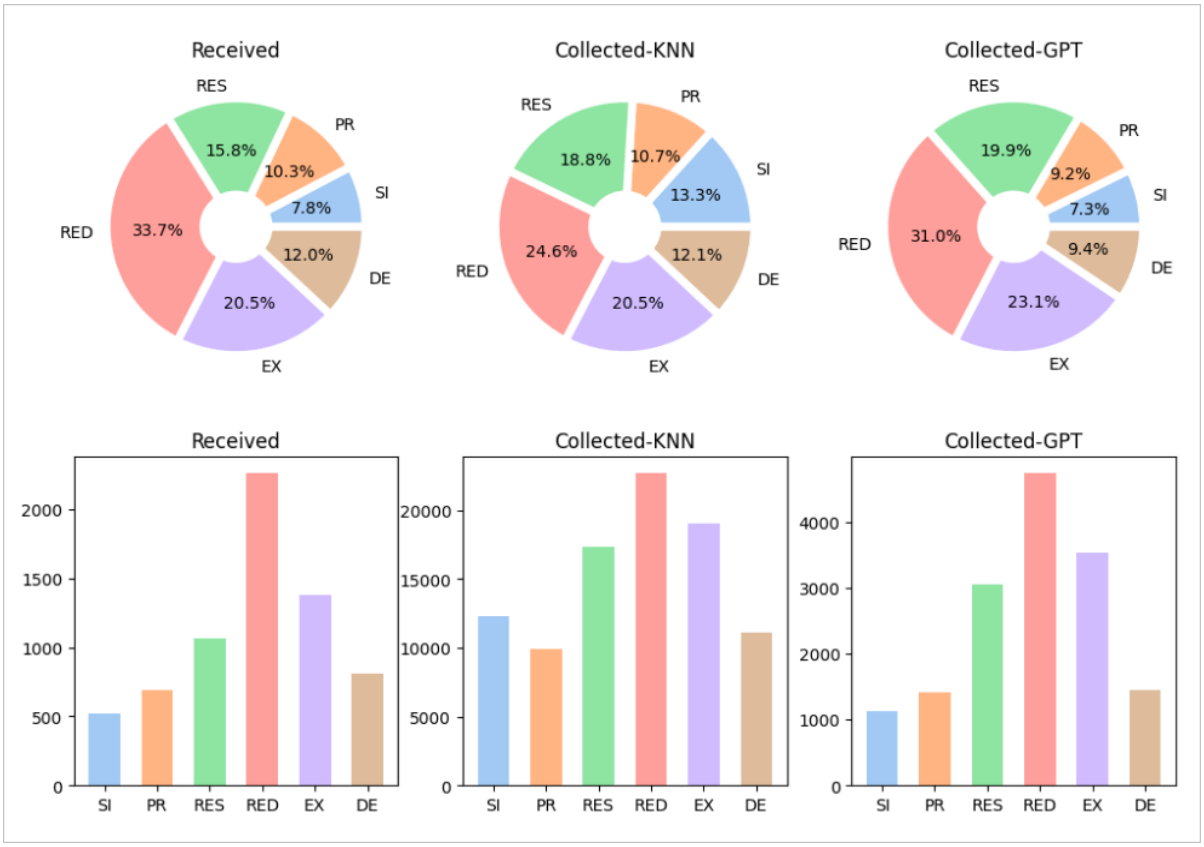
In [1]:
```python
import os
import sys
sys.path.append('/home/work/.data/abstract_classifier')

from valid import evaluate_model

work_dir = '/home/work/.data/abstract_classifier'
results = {'0-10': None, '1-9': None, '2-8': None, '3-7': None}

model_path = os.path.join(work_dir, 'outputs', '0-10', 'checkpoint-15000')
evaluation_data_path = os.path.join(work_dir, 'data', 'abstract_all.json')
f1, acc = evaluate_model(model_path, evaluation_data_path)
results['0-10'] = (f1, acc)
```

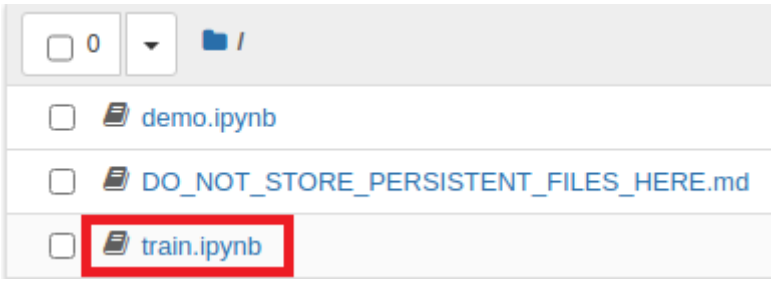
4. 코드 셀 아래에서 학습에 사용된 데이터들의 라벨 분포를 시각화한 내용을 확인할 수 있습니다.



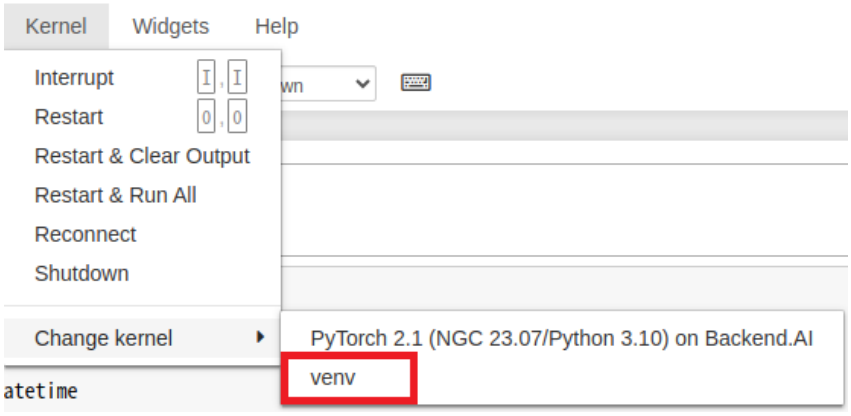
학습

jupyter notebook 에서 학습

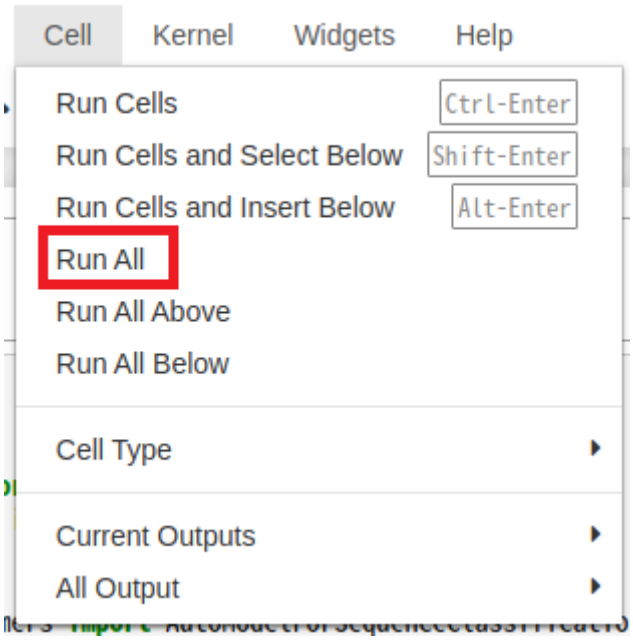
1. 파일 목록 중 'demo.ipynb'를 클릭합니다.



2. 상단의 Kernel - Change kernel - venv 를 선택합니다.



3. 상단의 Cell - Run All 를 선택하여 학습 진행 가능합니다.



5. 학습 진행 확인 가능합니다.

In [ ]:
```
with open(train_config_file, 'r', encoding='utf-8') as f:
        train_config = json.load(f)
if train_config['base_data']['do_prepare']:
    data_prepare(train_config)
train_main(train_config)
```

```
Dataset({
    features: ['text', 'label'],
    num_rows: 107702
})
Dataset({
    features: ['text', 'label'],
    num_rows: 3363
})
```

```
pytorch_model.bin: 100% 452M/452M [00:05<00:00, 80.9MB/s]
Some weights of ElectraForSequenceClassification were not initialized from the model checkpoint at monologg/koelectra-base-v3-discriminato
r and are newly initialized: ['classifier.out_proj.weight', 'classifier.dense.bias', 'classifier.dense.weight', 'classifier.out_proj.bia
s']
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Detected kernel version 3.10.0, which is below the recommended minimum of 5.5.0; this can cause the process to hang. It is recommended to
upgrade the kernel to the minimum version or higher.
```

[1001/38470 08:29 < 5:18:37, 1.96 it/s, Epoch 0.26/10]

Step	Training Loss	Validation Loss	Accuracy	F1
100	1.631000	1.479170	0.490039	0.443052
200	0.931500	1.024335	0.611062	0.594136
300	0.623600	0.970041	0.644068	0.648856
400	0.506800	0.915124	0.676479	0.672696
500	0.455400	0.969031	0.669938	0.667378
600	0.447800	0.892643	0.701160	0.700822
700	0.440100	0.875756	0.689860	0.694265
800	0.363900	0.988623	0.688968	0.683442
900	0.384400	0.913188	0.699673	0.697519
1000	0.342200	0.950695	0.690455	0.691947

터미널에서 학습

1. Console 에서 아래 명령어를 실행합니다.

```
$ cd/home/work/.data/abstract_classifier/
$ source venv/bin/activate
$ ./run_train.sh
```

```
work@main1[6jKitMxt-session]:~$ cd .data/abstract_classifier/
work@main1[6jKitMxt-session]:~/.data/abstract_classifier$ source venv/bin/activate
(venv) work@main1[6jKitMxt-session]:~/.data/abstract_classifier$ ./run_train.sh
```

2. 학습 진행 확인 가능합니다.