



## 2차 모의경진대회 NLP 과제 특강 : 쇼핑 중 대화 문장 의도 태깅

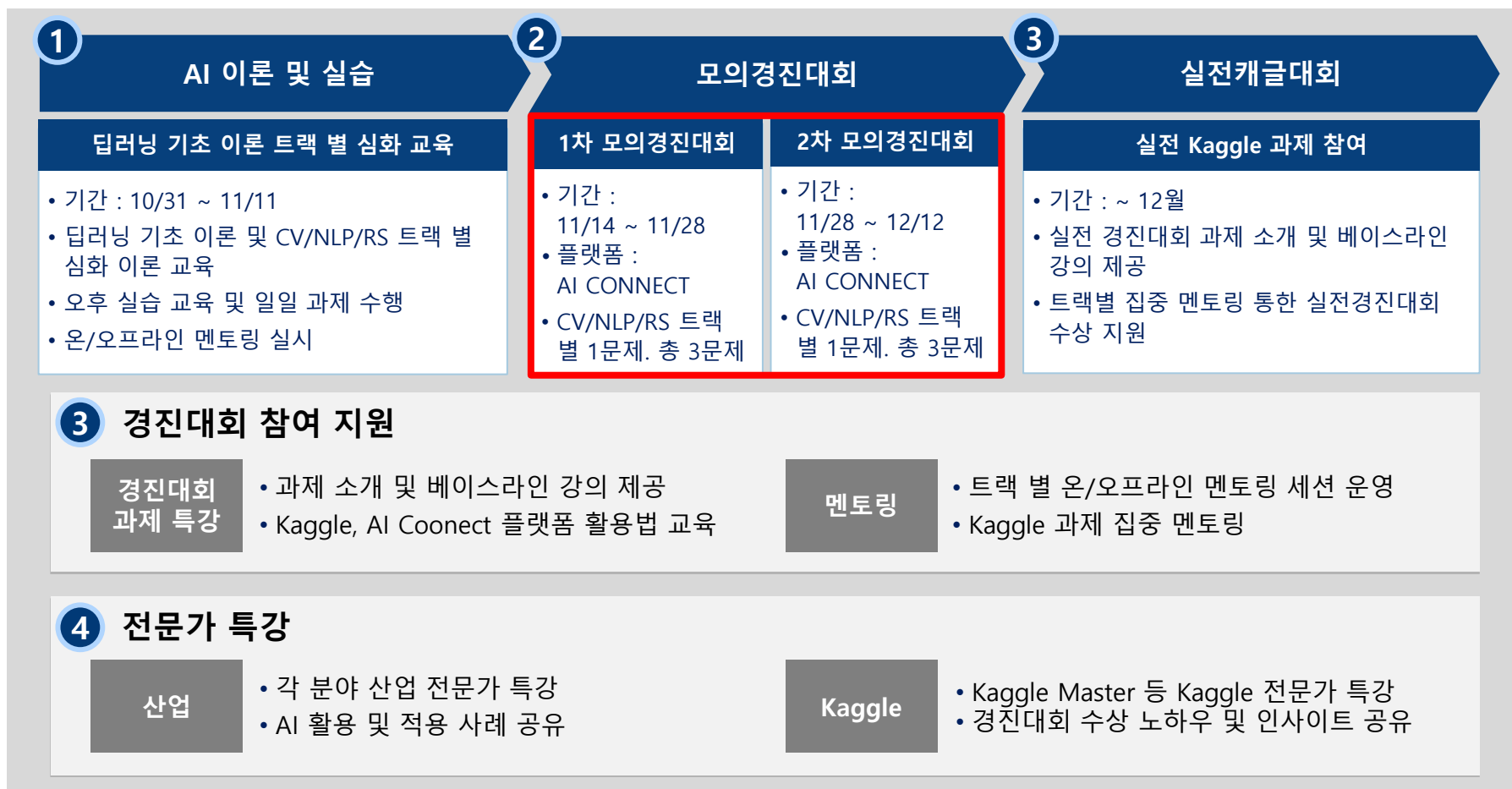
(주)마인즈앤컴퍼니 | 김태훈 매니저

**2022.11.28**

# Index

1. 실전캐글과정 커리큘럼 및 일정
2. AI CONNECT 플랫폼 사용법 안내
3. 2차 모의경진대회 – 쇼핑 중 대화 문장 의도 태깅 (Text Classification)
  - 과제 소개 (과제개요 / 데이터셋 / 베이스라인 모델 / 제한사항)
  - 베이스라인 코드

# 실전캐글과정 개요



# AI 이론 및 실습 / 모의경진대회 일정

2022년 11월				
월요일	화요일	수요일	목요일	금요일
28 (오전)2차 모의경진대회 OT - 박성호, 김태훈, 한두희 - 2차 모의대회 과제 설명 및 baseline 실습 (오후)트랙별 멘토링	29	30 (오전)산업 전문가 특강 - 딜리셔스 임정욱 - 기획자와의 커뮤니케이션 (오후)트랙별 멘토링		
2022년 12월				
월요일	화요일	수요일	목요일	금요일
			1	2 (오전)Kaggle 전문가 특강 - Kaggle 수상 경험 공유 (오후)캐글 멘토링
5 (오전)Python 특강 - 박성호M - Python을 통한 딥러닝 라이브러리 실제 구현 (오후)트랙별 멘토링	6	7 (오전)전문가 특강 - LG AI Research 전기정 팀장 - AI 기술 활용 사례 (오후)트랙별 멘토링	8	9 (오전)이어드림 졸업생 특강 - 김준철M - Kaggle 수상 경험 공유 (오후)캐글 멘토링
12 (오전)전문가 특강 - 서강대 구명환 교수 - AI Production 상용화 시스템 개발 (오후)트랙별 멘토링 2차 모의경진대회 종료	13 채용 행사	14 채용 행사	15 채용 행사	16 (오전)모의대회 우승자 특강 - 과제별 우승자 - 방법론 및 인사이트 공유 (오후)캐글 멘토링
19 (오전)전문가 특강 - 테디노트 이경록 - ML/DL 공부법	20	21	22	23

## 1 2차 모의경진대회

- 기간 : 11/28(월) 13시 ~ 12/12(월) 12시
- CV/NLP/RS track별 1문제
- 과제 소개 및 Baseline 특강

## 2 특강

- 시간 : 월/수/금 오전 9(or 10)시 ~ 12시
- 모의경진대회 및 실전 Kaggle 과제 특강
- 산업 전문가 특강

## 3 트랙별 멘토링

- 시간 : 월/수 오후 1시 ~ 4시
- 트랙 별 온/오프라인 멘토 섭외
- CV/NLP/RS 트랙 별 맞춤형 멘토링

## 4 캐글 멘토링

- 시간 : 금 오후 1시 ~ 4시
- 실전경진대회 과제 멘토링

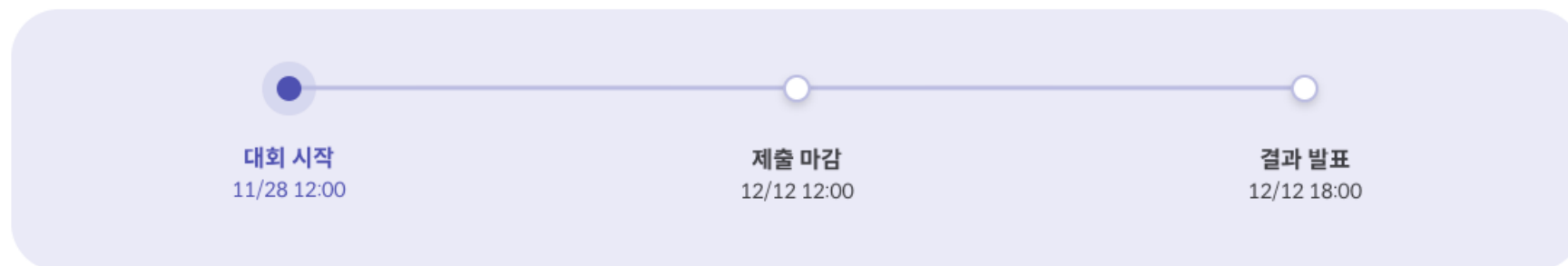
## 모의경진대회 개요

	이미지(CV)	자연어(NLP)	추천(RS)
1차 모의경진대회 (11.14 ~ 11.28)	 <p>사과 이미지를 이용한 사과 품종 분류</p>	 <p>도서자료 데이터를 이용한 기계독해</p>	 <p>고객 및 식당 데이터를 이용한 식당 만족도 예측</p>
2차 모의경진대회 (11.28 ~ 12.12)	 <p>항공 사진을 이용한 토지피복지도 객체 분할</p>	 <p>쇼핑 고객 상담 대화 문장 의도 태깅</p>	 <p>병원 리뷰 데이터를 이용한 병원 추천</p>

1, 2차 모의경진대회 과제들의 데이터셋, 베이스라인은 이어드림 입교생 공유 폴더함에 업로드되어 있습니다  
링크 : [https://drive.google.com/drive/u/0/folders/1R52BYe9icikTxTVM2ln9A92qBu\\_Ym4en](https://drive.google.com/drive/u/0/folders/1R52BYe9icikTxTVM2ln9A92qBu_Ym4en)

## 2차 모의경진대회 일정

### 일정



### 경진대회 세부 일정

- ✓ 오리엔테이션 및 과제 베이스라인 특강 : 11.28(월) 09:00 ~ 12:00
- ✓ 추론 결과 제출 : 11.28(월) 12:00 ~ 12.12(월) 12:00
- ✓ 결과 발표 : 12.12(월) 18:00
- ✓ 과제별 우승팀 코드 리뷰 세션 : 12.16 (금) 09:00 ~ 12:00


## 온/오프라인 멘토링

### 이미지(CV)

### 자연어 처리(NLP)


### 추천 시스템(RS)

#### 오프라인 멘토




**박성호 멘토**  
(631호 사무실)

- MNC Data Scientist
- Upenn Med 뇌과학 연구원
- Upenn 수학과 학사
- SCI, SSCI급 논문 제1저자
- 2022 NIPA 인공지능 온라인 경진대회




**박성일 멘토**  
(입교생 강의실)

- 서울대 데이터 사이언스 스쿨
- 대한민국 경찰(2019 ~ 2022)
- 경찰대 법학 학사
- 카카오톡 챗봇 개발 및 배포




**유승준 멘토**  
(입교생 강의실)

- 서울대 Learning and Adaptation Lab 인턴
- 서울대 DYROS 로봇틱스 부트캠프 수료
- 아주대 기계공학 학사




**김태훈 멘토**  
(631호 사무실)

- MNC Data Scientist
- 서울대 경제학부 학사
- 2022 NIPA 인공지능 온라인 경진대회
- 1기 이어드림 스쿨



**오로훈 멘토**  
(입교생 강의실)


- Megabyte School AI 데이터사이언티스트 취업완성 과정 수료
- 패스트캠퍼스 NLP 오프라인 학습 매니저



**강하예진 멘토**  
(입교생 강의실)


- Megabyte School AI 데이터사이언티스트 취업완성 과정 수료
- 패스트캠퍼스 RS 오프라인 학습 매니저

#### 온라인 멘토




**정채연 멘토**  
(ZOOM)

- 카이스트 김재철 AI 대학원 석박통합과정
- 고려대 경제학/컴퓨터학 학사
- 삼성전자 AI 교육과정 조교



**최민석 멘토**  
(ZOOM)

- 카이스트 김재철 AI 대학원 석박통합과정
- 일리노이대 컴퓨터학 학사
- 네이버 웹툰 AI Automation 팀
- 글로벌창업사관학교 조교



**곽대훈 멘토**  
(ZOOM)

- 카이스트 김재철 AI 대학원 박사과정
- 카카오엔터프라이즈 AI Lab
- 삼성전자 종합기술원 조교
- 삼성전자 DS AI Expert 조교
- 글로벌창업사관학교 조교

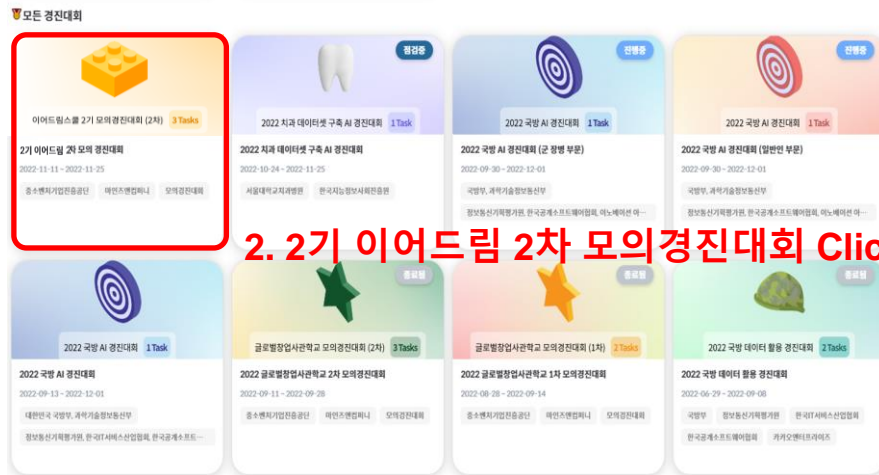
# Index

1. 실전캐글과정 커리큘럼 및 일정
2. AI CONNECT 플랫폼 사용법 안내
3. 2차 모의경진대회 – 쇼핑 중 대화 문장 의도 태깅 (Text Classification)
  - 과제 소개 (과제개요 / 데이터셋 / 베이스라인 모델 / 제한사항)
  - 베이스라인 코드

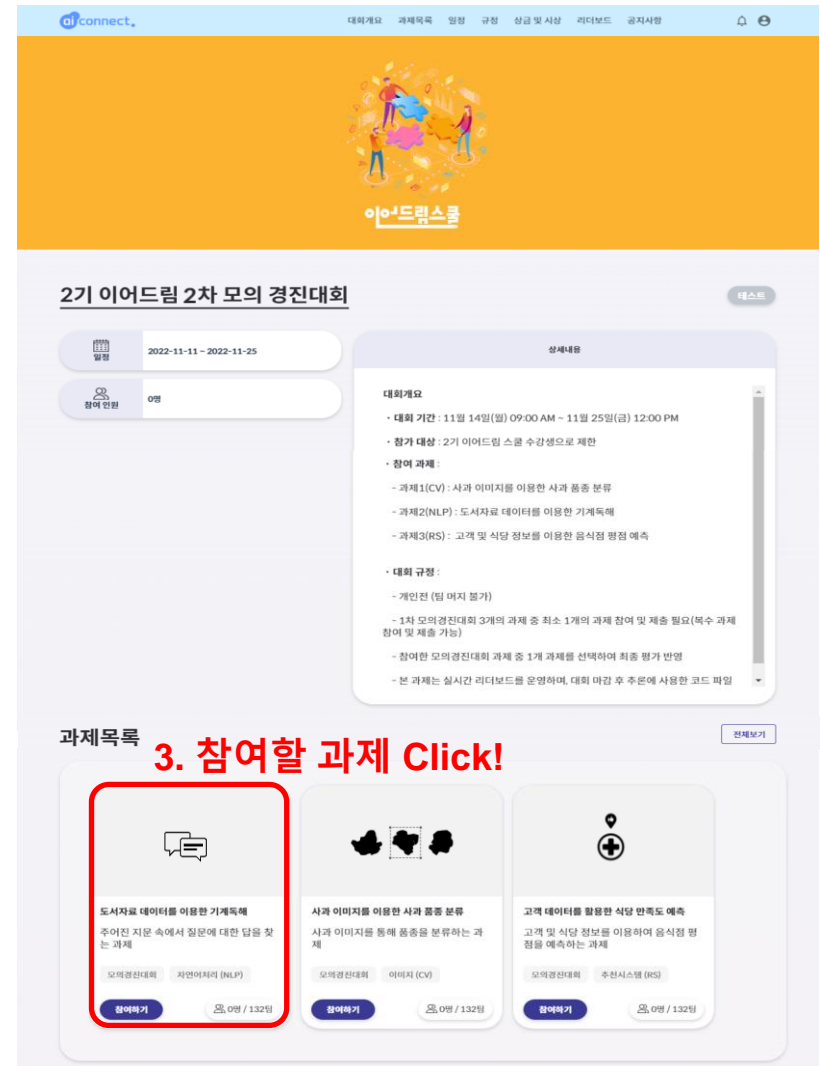


# AI CONNECT 과제 확인 방법

## 1. AI CONNECT 플랫폼 링크 Click!



## 2. 2기 이어드림 2차 모의경진대회 Click!



## 3. 참여할 과제 Click!

# AI CONNECT 과제 기능 탭 확인

2기 이어드림 1차 모의 경진대회 / 도서자료 데이터를 이용한 기계독해

과제 개요, 제한 사항, 평가지표 확인

리더보드 순위 및 score 확인

팀 머지 기능  
개인전 과제에서는 사용 XTrain/Test 데이터,  
Sample\_submission.csv 다운로드

베이스라인 코드 다운로드

과제 참여불가

HOME

과제 설명

리더보드

팀 구성

팀 빌딩

대회 공지/문의

데이터 설명/다운로드

코드공유

결과제출

과제 공지/문의

일정  
테스트참여인원  
0명

## 과제개요

도서자료 데이터 기계독해  
자연어 처리(NLP) | 개방형 문제 | Accuracy

### 문제정의

- 주어진 지문 속 질문에 대한 답을 찾는 Machine Reading Comprehension

## 제한사항

### <일반>

- 외부 데이터 사용 가능
- Pre-trained 모델 사용 가능
- 부정행위 적발 시 페널티 부여

### <제출 관련>

- 결과 제출 제한: 1일 최대 24회
- sample\_submission.csv과 동일한 형태로 예측 파일을 만들어 제출
- 최종 제출 파일 선택 ('결과제출' 탭에서 해당 파일의 '최종선택' 체크박스 선택)
- \* 최종 파일 미선택 시, public 스코어가 높은 제출 파일 자동 선택

## 평가지표

### Exact Matching(EM)

추론한 문자열이 정답 문자열과 완전히 일치할 경우 정답으로 인정

$$EM(Accuracy) = \frac{TP + TN}{TP + FP + TN + FN}$$

# AI CONNECT 결과 제출 방법

‘결과제출’ 탭을 통해  
추론 결과 파일 제출

과제 참여중

HOME

과제 설명

리더보드

팀 구성

대회 공지/문의

데이터 설명/다운로드

코드공유

결과제출

과제 공지/문의

과제참여취소

일정

테스트

참여인원

2명

결과제출

1일 최대 24회의 ‘결과제출’ 제한

결과제출 제한사항

파일크기

최대 1MB

제출횟수

하루 최대 24회까지 제출 가능

파일첨부

최대 1MB

제출설명

실험 및 코드 버전 관리를 위해 ‘제출설명’ 기능 활용

제출 파일에 대한 설명을 작성하세요.

제출하기

제출현황

#	제출파일	Public Score	성공여부	제출시간	최종선택
제출이력이 없습니다.					

Public score / CV score 등을 고려하여 제출 결과물 중 하나의 결과물 ‘최종선택’ 선택하지 않을 시 Public score가 가장 높은 결과물이 자동으로 최종선택됨

# Index

1. 실전캐글과정 커리큘럼 및 일정
2. AI CONNECT 플랫폼 사용법 안내
3. 2차 모의경진대회 – 쇼핑 중 대화 문장 의도 태깅 (Text Classification)
  - 과제 소개 (과제개요 / 데이터셋 / 베이스라인 모델 / 제한사항)
  - 베이스라인 코드

# 과제 개요

## 쇼핑 고객 상담 대화 문장 의도 태깅 | NLP

고객과 상담 내용 중 문장을 Classification하는 문제

### Input :

- 쇼핑 고객과의 상담 중 이루어진 대화문
- .txt 파일

평가  
지표

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

데이터

### Output :

- 총 14개의 대화의 의도(Speech Act) 클래스
- '인사하기', '진술하기', '질문하기', '약속하기 (개인적 수준의 서약)', '주장하기', '감사하기', '명령하기/요구하기', '사과하기', '충고하기', '긍정감정표현하기 (칭찬하기 포함)', '반박하기', '부탁하기', '거절하기', '부정감정표현하기 (비난하기 포함)'

```
A.반갑습니다 #@소속# 상담사 #@이름#입니다
B.안내해주신 방법을 시도해 봤는데 안 돼서요
A.그러시다면 두 번째 방법입니다
A.휴대폰 명의자 본인이에요
A.휴대폰하고 신분증 챙기셔서 저희 서비스센터 쪽으로 방문 하셔야 되세요 고객님의
B.이게 왜 이렇게 된거죠
B.그거 내가 설정 안 해봤는데요
A.그건 고객님의 설정하신 번호 눌러도 진입이 안 되세요 고객님의
B.그거 해봤지요 안 되네요
A.일단 고객님의 고객센터서 설정하신 걸 입력하셨는 데도 안 되신다면요
A.조금 번거로우시겠지만 센터 쪽으로 내방해 주셔야 할 것 같으세요 고객님의
B.어쩔 수 없지요 센터 가야 되겠네요
B.그러면 회선 쪽에 좀 알려주셔요
A.예 고객님의 회선지구 쪽으로 확인해 드리겠습니다
A.저희가 부산 해운대구에는 #@주소#에 있는 센터가 있고요
A.아니시면 #@주소#에 있는 작은 점점도 확인되고 있습니다
B.#@주소#이 나을 거 같아요
A.그러세요 그러시다면 #@주소#은 지하철 장산역 십 사 번 출구에 있구요
A.고객님 까르떠움 이 흥으로 방문하시면 됩니다
B.그래요 이 흥미요 알았어요
A.저희 서비스센터로 오시면 됩니다 고객님의
A.감사합니다 행복한 하루 보내십시오
```

← train.txt ▾ lines: [ ] 22 items

▾ 0:

id: 1

text: "A.반갑습니다 #@소속# 상담사 #@이름#입니다"

norm\_text: "A.반갑습니다 #@소속# 상담사 #@이름#입니다"

speaker:

speechAct: "인사하기"

morpheme: "A/SL+반갑/VA+습니다/EF+##/SY+@/SY+소속/NNG+##/SY+상담/NNG+사/NNG+##/SY+@/SY+이름/NNG+##/SY+입니다/VCP+EC"

label.json →

# 데이터셋

## 데이터 디렉터리 구조

```

DATA/
├── train/
│   ├── texts/
│   │   ├── shopping1_0001.txt
│   │   ├── ...
│   │   └── shopping7_2536.txt
│   └── labels/
│       ├── shopping1_0001.txt
│       ├── ...
│       └── shopping7_2536.txt
├── test/
│   └── texts/
│       ├── test_0001.txt
│       ├── ...
│       └── test_1456.txt
└── sample_submission.csv
  
```

## 데이터 수량

```

DATA/
├── train/
│   ├── texts/ : 15036개의 .txt 파일
│   └── labels/ : 15036개의 .json 파일
└── test/
    └── texts/ : 1456개의 .txt 파일
  
```

< shopping@\_####.txt 파일 >

A.반갑습니다 @@소속# 상담사 @@이름#입니다  
 B.안내해주신 방법을 시도해 봤는데 안 돼서요  
 A.그러시다면 두 번째 방법은요  
 A.휴대폰 명의자 본인미요

< shopping@\_####.json 파일 >

EX) 화자 : A , SpeechAct : 인사

EX) 화자 : B , SpeechAct : 문의

# label 데이터(.json) 구조

```

{ "dataset": { "identifier": 5701,
  "name": "s1_202101824_0001_0154_01",
  "src_path": "/data/file/cubeManager/PROJECT018/39/txt20210908151222005558/s1_202101824_0001_0154_01/",
  "label_path": "/data/file/cubeManager/PROJECT018/39/txt20210908151222005558/s1_202101824_0001_0154_01/",
  "category": 2,
  "type": 0},
  "licenses": { "name": "Apache License 1.0",
    "url": "http://www.apache.org/licenses/LICENSE-1.0" },
  "info": [ { "id": 8227,
    ...
    "annotations": { "subject": "AS문의",
      ...
      "text": "A.반갑습니다 #@소속# 상담사 #@이름#입니다\nB.안내해주신 방법을 시도해 봤는데 안 돼서요\n",
      "lines": [ { 'id': 1,
        "text": "A.반갑습니다 #@소속# 상담사 #@이름#입니다",
        "norm_text": "A.반갑습니다 #@소속# 상담사 #@이름#입니다",
        "speaker": { "id": "A",
          "sex": "여성",
          "age": "1그룹(10대~20대)"
        },
        "speechAct": "인사하기",
        "morpheme": "A/SL+반갑/VA+습니다/EF+#/SY+@/SY+소속/NNG+#/SY+상담/NNG-
      }, ...
    }
  ]
}
  
```

하나의 전체 텍스트  
파일에 부여된 id

json 파일 내  
전체 텍스트

텍스트를 구성하는 각 문장의  
id, text, 화자 정보, 의도 클래스  
등의 정보

대화 의도 클래스  
(총 14개)

# 베이스라인 모델 : RoBERTa

## RoBERTa: A Robustly Optimized BERT Pretraining Approach

Yinhan Liu<sup>\*§</sup> Myle Ott<sup>\*§</sup> Naman Goyal<sup>\*§</sup> Jingfei Du<sup>\*§</sup> Mandar Joshi<sup>†</sup>  
Danqi Chen<sup>§</sup> Omer Levy<sup>§</sup> Mike Lewis<sup>§</sup> Luke Zettlemoyer<sup>†§</sup> Veselin Stoyanov<sup>§</sup>

<sup>†</sup> Paul G. Allen School of Computer Science & Engineering,  
University of Washington, Seattle, WA  
{mandar90, lsz}@cs.washington.edu

<sup>§</sup> Facebook AI  
{yinhanliu, myleott, naman, jingfeidu,  
danqi, omerlevy, mikelewis, lsz, ves}@fb.com



- RoBERTa : A **R**obustly Optimized **B**ERT Pretraining **A**pproach
- BERT와 아키텍처는 동일하나, 데이터를 10배 늘려서 학습한 모델



## 제한 사항

- 외부 데이터 사용 가능
- **Pre-trained 모델 사용 가능**
  - 모델 라이브러리를 통해 이미 학습이 진행된 모델을 로드하여 사용할 수 있음
  - .pt, .pth, .h5 등의 pre-trained 가중치 파일을 업로드하여 사용할 수 있음
  - Transfer learning, Fine tuning 가능함
- 결과 제출 제한 : 1일 최대 24회
- sample\_submission.csv와 동일한 형태로 예측 파일을 만들어 제출
- 최종 제출 파일 선택  
(‘결과제출’ 탭에서 해당 파일의 ‘최종선택’ 체크박스 선택)
  - \* 최종 파일 미선택 시, public 스코어가 가장 높은 제출 파일 자동 선택

# Index

1. 실전캐글과정 커리큘럼 및 일정
2. AI CONNECT 플랫폼 사용법 안내
3. 2차 모의경진대회 – 쇼핑 중 대화 문장 의도 태깅 (Text Classification)
  - 과제 소개 (과제개요 / 데이터셋 / 베이스라인 모델 / 평가지표 / 제한사항)
  - 베이스라인 코드



**End of document**