

챗봇 만들기.. Proj

목차

- 데이터 선정
- 데이터 확인
- 데이터 정제
- finetuning
- inference

데이터 선정

- 서울공공데이터 포털에 있는 서울특별시 공영주차장 정보가 담겨 있는 데이터를 사용
- <https://data.seoul.go.kr/>



서울 열린데이터 광장

데이터 확인

```
"DESCRIPTION" : { "PKLT_KND": "주차장 종류", "LHLDY_YN": "공휴일 유, 무료 구분", "CHGD_FREE_NM": "유무료구분명",  
  "OPER_SE": "운영구분", "PRK_HM": "기본 주차 시간(분 단위)", "TPKCT": "총 주차면", "ADD_CRG": "추가 단위 요금",  
  "LHLDY_NM": "공휴일 유, 무료 구분명", "SAT_CHGD_FREE_NM": "토요일 유, 무료 구분명", "BUS_PRK_HM": "버스 기본 주차 시간(분 단위)",  
  "PRK_NOW_INFO_PVSN_YN_NM": "주차현황 정보 제공여부명", "PKLT_KND_NM": "주차장 종류명",  
  "MNTL_CMUT_CRG": "월 정기권 금액", "OPER_SE_NM": "운영구분명", "DLY_MAX_CRG": "일 최대 요금",  
  "ADDR": "주소", "BUS_PRK_ADD_CRG": "버스 추가 단위 요금", "ADD_UNIT_TM_MNT": "추가 단위 시간(분 단위)",  
  "LHLDY_BGNG": "공휴일 운영 시작시각(HHMM)", "LOT": "경도", "PKLT_CD": "주차장코드", "WE_OPER_BGNG_TM": "주말 운영 시작시각(HHMM)",  
  "LHLDY": "공휴일 운영 종료시각(HHMM)", "BUS_PRK_ADD_HM": "버스 추가 단위 시간(분 단위)", "NGHT_FREE_OPN_YN": "야간무료개방여부",  
  "WD_OPER_BGNG_TM": "평일 운영 시작시각(HHMM)", "SAT_CHGD_FREE_SE": "토요일 유, 무료 구분", "WE_OPER_END_TM": "주말 운영 종료시각(HHMM)",  
  "WD_OPER_END_TM": "평일 운영 종료시각(HHMM)", "LAT": "위도", "PRK_CRG": "기본 주차 요금", "CRB_PKLT_MNG_GROUP_NO": "노상 주차장 관리그룹번호",  
  "TELNO": "전화번호", "LAST_DATA_SYNC_TM": "최종데이터 동기화 시간", "CHGD_FREE_SE": "유무료구분", "PRK_NOW_INFO_PVSN_YN": "주차현황 정보 제공여부",  
  "PKLT_NM": "주차장명", "BUS_PRK_CRG": "버스 기본 주차 요금",  
  "NGHT_FREE_OPN_YN_NAME": "야간무료개방여부명"},
```

```
"DATA" : [ {"bus_prk_add_crg": 0, "prk_now_info_pvs_n_yn": "0", "bus_prk_hm": 0, "tpkct": 71, "oper_se": "1", "pklt_knd_nm": "노외 주차장",  
  "crb_pklt_mng_group_no": "", "telno": "", "last_data_sync_tm": "2022-01-12 15:16:35", "addr": "도봉구 창동 24-0", "prk_crg": 0,  
  "prk_hm": null, "add_crg": 300, "lhldy_yn": "N", "oper_se_nm": "시간제 주차장", "mntl_cmut_crg": "0", "add_unit_tm_mnt": 10,  
  "lhldy_nm": "무료", "chgd_free_nm": "유료", "lat": null, "sat_chgd_free_se": "N", "wd_oper_end_tm": "1900", "lhldy_bgng": "0900",  
  "dly_max_crg": 0, "pklt_knd": "NW", "sat_chgd_free_nm": "무료", "we_oper_bgng_tm": "0900", "prk_now_info_pvs_n_yn_nm": "미연계중",  
  "pklt_cd": "1010089", "bus_prk_add_hm": 0, "nght_free_opn_yn_name": "야간 미개방", "chgd_free_se": "Y", "bus_prk_crg": 0,  
  "wd_oper_bgng_tm": "0900", "pklt_nm": "초안산근린공원주차장(구)", "lot": null, "nght_free_opn_yn": "N", "we_oper_end_tm": "1900", "lhldy": "1900"  
  },  
  ...  
]
```

데이터 구조 변경

```
[ "서울시 공영주차장1" : {  
  "주차장 위치" : ""  
  "이름" : "",  
  "운영시간" : {"평일" : "", "주말" : "", "공휴일" : ""},  
  "요금정보" : {"기본 요금 및 기본 시간" : "", "추가 단위 시간 당 요금" : ""},  
  "야간 개방" : ""  
},  
...  
"서울시 공영주차장2" : {  
  "주차장 위치" : ""  
  "이름" : "",  
  "운영시간" : {"평일" : " ", "주말" : " ", "공휴일" : " "},  
  "요금정보" : {"기본 요금 및 기본 시간" : " ", "추가 단위 시간 당 요금" : " "},  
  "야간 개방" : ""  
}  
]
```

데이터 구조 변경

```
"서울시 공영주차장 num" : {  
  "주차장 위치" : dev_data['DATA'][num][addr]  
  "이름" : dev_data['DATA'][num][pklt_nm],  
  "운영시간" : {"평일" : dev_data['DATA'][num][wd_oper_bgng_tm]에서dev_data['DATA'][num][wd_oper_end_tm],  
                "주말" : dev_data['DATA'][num][we_oper_bgng_tm]에서dev_data['DATA'][num][we_oper_end_tm],  
                "공휴일" : dev_data['DATA'][num][lhldy_bgng]에서dev_data['DATA'][num][lhldy]  
                },  
  "요금정보" : {"기본 요금 및 시간" : dev_data['DATA'][num][prk_hm] 분에 dev_data['DATA'][num][prk_crg]원,  
                "추가 단위 시간 당 요금" : dev_data['DATA'][num][add_add_unit_tm_mnt]분 당 dev_data['DATA'][num][add_crg]원  
                },  
  "야간 개방" : dev_data['DATA'][num][nght_free_opn_yn]  
}
```

데이터 정제

1. 초기 학습 목표

Instruction: 서울시 공영주차장 0 ### Response: {'위치': '도봉구 창동 24-0', '이름': '초안산근린공원주차장(구)', '운영 시간': {'평일': '0900에서 1900', '주말': '0900에서 1900', '공휴일': '0900에서 1900'}, '요금 정보': '요금 정보 없음', '야간 개방 여부': 'N'}

Instruction: 서울시 공영주차장 1 ### Response: {'위치': '노원구 상계동 770-2', '이름': '마들스타디움(근린공원)(구)', '운영 시간': {'평일': '0000에서 2400', '주말': '0000에서 2400', '공휴일': '0000에서 2400'}, '요금 정보': '5분 당 150원', '야간 개방 여부': 'N'}

Instruction: 서울시 공영주차장 2 ### Response: {'위치': '성동구 마장동 463-2', '이름': '마장동(건물) 공영주차장(구)', '운영 시간': {'평일': '0000에서 2400', '주말': '0000에서 2400', '공휴일': '0000에서 2400'}, '요금 정보': '5분 당 50원', '야간 개방 여부': 'N'}

Instruction: 서울시 공영주차장 3 ### Response: {'위치': '영등포구 신길동 184-3', '이름': '영등포여고 공영(구)', '운영 시간': {'평일': '0000에서 2400', '주말': '0000에서 2400', '공휴일': '0000에서 2400'}, '요금 정보': '5분 당 50원', '야간 개방 여부': 'N'}

-> 데이터 학습시간이 너무 오래걸려서 코랩의 런타임을 초과해버림...

2. 수정

Instruction: 도봉구 창동 24-0에 있는 공영주차장 정보 ### Response: 이름은 초안산근린공원주차장(구)이고 요금은 요금 정보 없음이야.

Instruction: 노원구 상계동 770-2에 있는 공영주차장 정보 ### Response: 이름은 마들스타디움(근린공원)(구)이고 요금은 5분 당 150원이야.

Instruction: 성동구 마장동 463-2에 있는 공영주차장 정보 ### Response: 이름은 마장동(건물) 공영주차장(구)이고 요금은 5분 당 50원이야.

Instruction: 영등포구 신길동 184-3에 있는 공영주차장 정보 ### Response: 이름은 영등포여고 공영(구)이고 요금은 5분 당 50원이야.

-> 데이터 학습시간은 짧아져서 학습을 시키기는 했는데 학습시키는 정보가 너무 부족함...

finetuning

```
import os
os.environ['HF_TOKEN']="hf_YceBcrSpLarkpIVTkCI MVDRMFogYRqaEfM"
```

```
!autotrain llm --train #
  --project-name "llama3-park-finetuning-da-8B" #
  --model "meta-llama/Meta-Llama-3-8B" #
  --data-path "park_prompt_data" #
  --text-column "text" #
  --peft #
  --quantization "int4" #
  --lr 2e-4 #
  --batch-size 8 #
  --epochs 10 #
  --trainer sft #
  --model_max_length 256
```

```
- {'loss': 0.9811, 'grad_norm': 0.5538907051086426, 'learning_rate': 0.00018253968253968255, 'epoch': 1.7543859649122808}
- {'loss': 0.4589, 'grad_norm': 0.39669400453567505, 'learning_rate': 0.00014285714285714287, 'epoch': 3.5087719298245617}
- {'loss': 0.37, 'grad_norm': 0.6180463433265686, 'learning_rate': 0.00010317460317460319, 'epoch': 5.2631578947368425}
- {'loss': 0.3021, 'grad_norm': 0.6176848411560059, 'learning_rate': 6.349206349206349e-05, 'epoch': 7.017543859649122}
- {'loss': 0.2271, 'grad_norm': 0.7432821393013, 'learning_rate': 2.380952380952381e-05, 'epoch': 8.771929824561404}
```

loss가 줄어들긴 하였으나 epoch를 10번 밖에 돌리지 못하여 더 줄이지 못한 것 같음.

inference

```
[ ] gen("도봉구 창동에 있는 공영주차장에 대해서 알려줘")
```

```
⇒ Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.  
'<|begin_of_text|> 이름은 창2동 1공영주차장(구)이고 요금은 5분 당 100원야.<|end_of_text|>'
```

```
[ ] gen("도봉구에 있는 공영주차장 정보를 알려줘")
```

```
⇒ Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.  
'<|begin_of_text|> 이름은 도봉2동 2공영주차장(구)이고 요금은 5분 당 100원야.<|end_of_text|>'
```

```
[ ] gen("도봉구에 있는 공영주차장 여러개 알려줘")
```

```
#음... 학습을 잘못된거지 답변은 무조건 한개이다... 도봉구에 있기는 한데 한개만 알려줌...
```

```
⇒ Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.  
'<|begin_of_text|>1. 이름은 쌍문동 3-1 공영주차장(구)이고 요금은 5분 당 50원야.<|end_of_text|>'
```

```
[ ] gen("서울 강북구에 있는 공영주차장 이름 알려줘")
```

```
⇒ Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.  
'<|begin_of_text|> 이름은 구청앞(구)이고 요금은 5분 당 250원야.<|end_of_text|>'
```

질문을 하고 대답을 시켜 보았으나
자신이 학습한 내용(--에 있는 주차
장 정보는 요금이 --이다.)에 대해서
는 잘 대답하는 것 같으나,
“여러개 대답해줘”/“어디 주변에 있
는” 등.. 다양한 질문에 대한 답변은
잘 수행하지 못함.

마무리

추후 project를 더 진행할 기회가 있다면 데이터를 조금더 잘 정제하여서 학습시간을 줄여나가는 방법을 고민해봐야 할 것 같습니다.

라마 3를 이용하여 학습했는데, 다음에는 3.1을 이용해서 학습해보도록...