텍스트기반 기상-AI 검색기 개발 인수인계서

1. 인수인계 담당자 연락처: 장요엘 wkddydpf@kaist.ac.kr
2. Github repo:
    a. development code repo: https://github.com/joeljang/KoBART-summarization (private repo 여서 인수인계 담당자에게 invitation 필요)
    b. deployment code repo: https://github.com/soyoung97/weather
3. TODO:
    - (COMIS) 개선사항:
        - 위험기상을 입력하는 경우 위험기상 기본 검색 결과가 아닌 분석장 결과를 전시함. pseudo url 의 결과가 실제 검색 조건과 매칭이 안되는 경우가 많이 있어서 URL Mapping 하는데 exception 이 자주 발생됨. -> 정확도 및 generalization ability 개선
        - 날짜 포함 검색어는 인식하지 못함, '어제' 키워드만 인식
        - 과학원 키워드 '지상일기도에 적외선 위성 영상 중첩'의 경우 다수의 검색 결과가 존재하지만 하나의 검색 결과만 전시
    - (NL2SQL) Tibero 사용한 sample DB 구축
    - (NL2SQL) Template 추가
    - (NL2SQL) Template 에서 각 table name 앞에 COMIS.[table_name] 으로 변경 필요. 이유는 제주도에 있는 기상과학원 내부망에서 sql query 실행시 필요하다고 피드백 받음 (12.17)
    - (NL2SQL) Template matching 성능 개선
        - Korean Sentence Transformer 로 Multilingual Sentence Transformer 대체
        - 템플릿에 대한 Contrastive learning
    - (NL2SQL) Dataset 구축은 전에 전달받은 "검색기 기본검색.sql"을 기반으로 만들어져 있습니다. 예를 들어 첫 번째 예시문의 "전기간 전지점 일단위 최고온도 3 개"는 전달받은 SQL 문의 1 번 (극값 검색 온도:: 전기간 전지점 일단위 최고온도 20 개)를 기반으로 만들어졌고, 두 번째 예시문의

"당일(2021 년 1 월 24 일) 전지점 일단위 최고온도 30 개"는 전달받은 SQL 예시의 11 번 (당일 검색 일단위 최고온도 1 개) 을 기준으로 만들어져 있습니다. 따라서 전달받은 SQL 문에 오류가 있을 시 output 되는 SQL 문에도 오류가 있을 수 있으며, 이를 고치기 위해서는 기존 template 를 수정하면 될 것으로 보입니다.

- (NL2SQL) 현재 "당일"에 해당하는 input 에 대해서는 뒤에 당일(YYYY 년 MM 월 DD 일)로 input 을 받는다고 가정하고 학습을 진행하였기 때문에, 이를 날짜가 없는 당일로 변환하고 날짜는 date 인풋에서 가져오도록 하려면 코드 수정이 필요할 것으로 보입니다.
- (Docker) 현재 NVIDIA driver version 460.73.01, CUDA version 11.2 인 azure T4 서버에서 msyoon8/weather:latest docker image 가 cpu/gpu 상황에서 모두 작동하도록 개발되었는데, 제주도에 있는 인터넷이 단절된 환경에서 해당 이미지 실행시 cuda 가 제대로 잡히지 않는다는 제보를 받았습니다. 실제로 그 환경에 가 debugging 을 할 수 없기 때문에, 현재 정확한 원인은 파악하지 못했습니다.실제 deploy 하는 상황에서 cpu 버전은 정상적으로 작동하나, 추후에 gpu 버전이 정상적으로 작동하도록 image 를 변경하는 것이 필요합니다.

4. 세부 사항:
    a. Output space vocab size
    Total Entry # : 12,943
    Max Depth : 7
    Possible outputs in each depth:
    1 * 1 * 25 * 354 * 31 * 5 * 3 = 4,115,250 possible output combinations
    b. Increase robustness to noise
    Give little turbulence to the training dataset to make it more robust to typos, etc. 1) insert/delete space 2) randomly delete punctuation
    two evaluation set (clean, noise)
    c. COMIS 모델 정확도
    EM ~ 88%
    d. Output control
    생성 후 Rule-based 필터링 진행
    e. 응답 메세지 형태
    Request format:

```
source: "오늘의 UM전구 저기압이동경로"
date: "2021-10-18 00:00:00"
sourceType: "text"
responseChannel: "aiw-response"
```

Response format:

```
{
pseudoList:[
      {site:"COMIS",
      pseudo:"일기도_예보장?동적선택1=저기압이동경로&동적버튼1=KIM전구&입력1=202109241134
      "},
      {site:"선진예보",
      pseudo:"일기도_분석장?동적선택1=KIM&동적선택2=3시간:지상&동적선택3=기본&입력1=20210
      9241134"}
  ],
extremeValue:[
      "select * from TB_COMIS_TABLE WHERE DATE=....",
      "select * from TB_COMIS_TABLE2 WHERE DATE=...."
  ]
}

요청 메시지에 날짜를 지정하는 검색어가 없는 경우 입력1 생략
```

f. SQL Template Augmentation

   Variable: 숫자, 날짜, 위치,

   Template: 최저 / 최고/ 최저최고, 온도/풍속/강수량/적설, 일단위/월단위

g. DB Dump 와 schema (LK Lab Google Dirve 에 저장됨, LK Lab 소속 아니면 인수인계시 invitation 필요)

   2021 기상과학원 > Data > COMIS_DB

h. NL2SQL 모델 정확도

   현재 우리는 14 개의 템플릿을 보유하고 있고, 템플릿 매칭의 정확도는 56 개의 평가 데이터에 대해서 21.43%(12/56)이다. 모델이 템플릿을 정확하게 골랐을 경우 룰 기반의 슬롯 채우기의 정확도는 100%이다. 따라서 구현된 모델은 56 개의 평가 데이터에 대해서 21.43%(12/56)의 정확도를 가진다.

i. GPU 버전과 CPU 버전을 선택적으로(옵션) 구동