컨퍼런스 최종평가 : 1월 4일 (토)

컨퍼런스 : 1월 11일 (토)

**Base : Reference 찾기 : 중간고사 기간 전까지.(10월 9일 ~ 10월 27일)**

1. 어떤 데이터셋을 사용?
2. 참고할만한 자료 및 논문
3. 어떤 모델을 사용할지?
4. 본 프로젝트는 **챗봇 구현**이 중요하므로 챗봇만드는 Reference도 많이 필요할 것.

( Rule based가 많으나 그냥 그런 것도 다 합해서 모으기)

+) 일상적인 대화도 가능한 모델 (대화가 가능하려면, rule-based 요소도 들어가야함)

**Phase1 : Data Collection & Preprocessing : 중간고사 기간 전까지.(10월 9일 ~ 10월 27일)**

* Data Collection : 장애인 법령과 관련된 Data를 Crawling하여 Data Frame으로 만드는 과정. 현재 찾은 곳은 국민신문고게시판, naver 지식IN 이 밖에도 더 있을 것으로 추정. 일단 이 2개에 대하여 크롤링 진행하여야 함.
* Preprocessing : Tokenizing & Embedding -> index table, Lookup table구성
  + index table은 word의 index를 부여하여 table(matrix)로 만드는 것.
  + Lookup table은 word의 Embedding을 진행하여 이를 table(matrix)로 만드는 것
  + Tokenizer 뭐 쓸지는 같이 고민해봐야 할 것.
  + Word Embedding도 마찬가지
  + \*중요) Document Embedding도 필요할 것으로 예상됨. 이 Embedding으로 만으로도 사실상 챗봇 구현이 가능함.

\* 국민 신문고 → 한 사람 당 267개씩 맡아서 읽어봐야 함. (국민 신문고 데이터 셀 중 버릴 것은 버리는 과정 필요; 27일까지 끝내기)

\* 글 키워드(한 줄 이내 요약, 단어 중심) label을 따로 칼럼으로 생성해서 달아놓는기 확정!!!! (27일까지)

\* 지식인 크롤링 (유나, 윤종: 25일 이후 시작)

\* Approach 1(Rule-Based) → 전체 글 모두에 대해서 유사도를 구하는 것은 연산량이 너무 많아짐. 따라서, 토픽 모델링을 진행해서 토픽별로 유사도 계산

\* 몇개 sampling해서 phase2에 대한 실험 진행 후 어떤 모델로 진행할지 감 잡는 과정 필요(진행: 준걸, 코드 리뷰: 유진, 효은, 혜민)

**Phase2 : Modeling & Algorithm & Advanced Modeling : 늦어도 기말고사 끝나기 전까지.(11월 1일 ~ 12월 23일)**

* Approach 1 : 유사도 알고리즘을 사용한 방법 (Rule-Based) 접근
* Approach 2 : NER을 통해 질문 법령 인식 (글에서 키워드를 찾아내는 것)
* Approach 3 : QA Modeling (답변 자체를 딥러닝으로 만들자, 어려울 것으로 예상)
* Apporach 4 : 앙상블 방d법 (다 적용해 섞어보는 것(Rule-based & DL / 사실상 4번으로 진행할 듯.)
* 각각 논문을 및 Reference 찾아 정리하기.

\*NER: 글의 토큰 중 어떤 토큰이 키워드인지 찾아내는 작업

\*리뷰할 논문 추린 후, 나중에 각자에게 과제 형식으로 나갈 예정.

**Phase3 : Chatbot Development & Finish (나머지 기간)**

* 챗봇으로 구현하기 위한 작업.
* 딥러닝 Test의 연산속도가 3초이내(확실하지 않음)여야 한다고 알고 있음.

10기, 11기 → phase2, 3 중점

12기 → support