**세미프로젝트 기획안**

**과목명: 실전 빅데이터 분석 프로젝트**

**기획안 작성일자 : 2022-06-07**

|  |  |
| --- | --- |
| 팀 명 | 3조 : Best-M4.t3 |
| 프로젝트 조직  (구성원 및 역할) | 팀장  신재웅 : 총괄, 데이터 크롤링 및 모델구성, 모델검증  ​  팀원  박찬규 : 최종 발표, 데이터 크롤링 및 모델구성, 알고리즘 아키텍처 담당  ​  권기일 : 깃허브 관리, 데이터 크롤링 및 모델구성, PPT 담당  ​  강신웅 : 기획안 발표, 데이터 크롤링 및 모델구성, 데이터 전처리 |
| 프로젝트 주제 및 개요  해결하고자 하는 문제  최종 산출물의 청사진 | 법률 상담 AI |
| 프로젝트 수행 방향  수행 방법/도구 | 네이버 지식인 법률 전문가 답변 **Q**uestion **A**nswering 모델 구성  Dataset: 네이버 지식인 민사소송 및 집행 카테고리 전문가 답변  총 약 22,000건 (훈련용 18,000개 검증용 4,000 개)  Lib: Pytorch 1.11  networkx, hanspell, Mecab  Prep: 불필요한 인사말 제거 및 네이버 맞춤법 검사기 API 사용  TextRank 알고리즘을 통해 Text Summarization  질문 및 답변에 대한 핵심 문장을 추출해 학습  Model: 유사도 기반 모델과 Seq2Seq, Transformer 계열 모델 사용  질문의 유사도를 기준으로, 유사도가 높은 질문일 경우 동일한 답변 생성  유사도가 낮은 질문이 입력됐을 때, 딥러닝 모델을 통해 자연어를 생성해 법률 상담에 대한 적절한 답변 생성  토큰의 길이가 짧을 경우, Seq2Seq 모델 통과  길 경우, Transformer 계열 모델 통과  RNN 모델은 입력 문장이 길어졌을 시, Gradient Vanshing 문제가 발생하기 때문에, 해당하는 경우에 Attention Machnism을 사용한 Transformer 계열 모델을 사용  Postp: 핵심 문장을 이용해 학습을 진행하므로, 생성된 문장 또한 적절한 후처리를 통해 답변의 형태로 가공  플라스크, 장고 외 하이브리드 앱 구현을 위한 프레임워크를 사용해 웹 혹은 하이브리드 앱 형태로 최종 출력물 제공 |
| 프로젝트 추진 일정 | 2022/5/30 ~ 2022/6/6: 주제 선정 및 데이터 수집  ~ 2022/6/8: 전처리  ~ 2022/6/10: 모델링  ~ 2022/6/14: 검증 및 모델 보완, 시각화  ~ 2022/6/15: 발표자료 제작 |