인공지능 데이터 전처리 결과서

SKN 2기 3조 추운자리

구선아 김진유 이재원 전상욱

I. 회의록 요약 모델(sLLM) 파인튜닝 활용 데이터

A. 데이터셋

데이터	용량	출처
국회 회의록 말뭉치 2021	307 MB	국립국어원 모두의 말뭉치
국회 회의록 요약 말뭉치 2022	71 MB	국립국어원 모두의 말뭉치
국회 회의록 요약 말뭉치 2023	58 MB	국립국어원 모두의 말뭉치

B. 설명

국립국어원에서 국회 회의록 말뭉치 2021에 존재하는 회의록 중에서 선별하여 2022년, 2023년에 요약문을 작성함.

2022년에 500 건, 2023년에 300 건으로 총 800 건의 데이터셋 활용.

C. 선정 이유

회의에 대한 녹취록과 요약으로 이루어진 데이터셋으로 개발하고자 하는 sLLM의 Input - Output 쌍과 유사한 데이터

II. 데이터 전처리 결과

A. 전처리 과정

1. Pandas 패키지를 사용하여 데이터셋 프레임을 생성
2. 국회 회의록 요약 말뭉치 2022, 국회 회의록 요약 말뭉치 2023에서 쓰인 회의록 목록을 추출하여 원본과 요약 컬럼으로 정리
3. sLLM에 학습시킬 텍스트의 형식(Prompt)으로 정리하여 새 컬럼으로 생성

B. 전처리 결과

	id	minutes	summary	prompt
214	SBRW2100000215	성원이 되었으므로 제2차 법안심사소위원회를 개의하겠습니다.\r\n오늘은 어제 처리하...	본 회의는 제215회 2차 법안소위원회로 먼저 사립학교법중개정법률안을 심사하였다.\...	당신은 회의록을 요약해주는 유익하고 유능한 AI입니다.\n 회의록을 요약한 요...
219	SBRW2100000220	성원이 되었으므로 제1차 예산안조정소위원회를 개의하겠습니다.\r\n소위원회이므로 앉...	본 회의는 제215회 제1차 예산안조정소위원회로, 2000년도 제1회추가경정예산안을 ...	당신은 회의록을 요약해주는 유익하고 유능한 AI입니다.\n 회의록을 요약한 요...
230	SBRW2100000231	지금부터 예산안심사소위원회를 개의하겠습니다.\r\n오늘 전체회의에 이어 바로 소위원...	본 회의는 제215회 국회 제1차 예산심사소위원회로, 의사일정 제1항 2001년도노...	당신은 회의록을 요약해주는 유익하고 유능한 AI입니다.\n 회의록을 요약한 요...
241	SBRW2100000242	정돈해 주시기 바랍니다.\r\n제3차 예산안조정소위원회를 개의하겠습니다.\r\n의사...	본 회의는 제216차 국회 제3차 예산결산예산안 조정 소위원회회의로 금일 2001년...	당신은 회의록을 요약해주는 유익하고 유능한 AI입니다.\n 회의록을 요약한 요...
244	SBRW2100000245	성원이 되었으므로 제6차 예산안조정소위원회를 개의합니다.\r\n그러면 의사일정 제1...	본 회의는 제216회 국회 제06차 예산결산 소위원회 예산안조정 회의로 의사일정 제...	당신은 회의록을 요약해주는 유익하고 유능한 AI입니다.\n 회의록을 요약한 요...