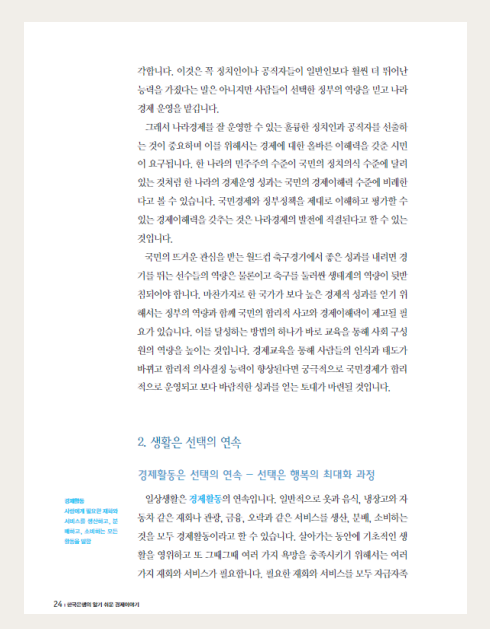
데이터 수집

SKN Family AI Camp 2기 : 최종 프로젝트 2조

프로젝트 주제	경제 지식이 없는 아동 및 국민들을 대상으로 경제 지식을 쉽게 설명하는 AI 인플루언서 웹 페이지 관리 Tool				
데이터 출처	한국은행	**데이터 형태**	pdf 파일	**데이터 정보 및 분량 (사이즈)**	경제금융용어 700선(375p), 알기 쉬운 경제이야기 : 일반인용 (335p)
데이터 수집 계기	아동 및 청소년을 대상으로 하는 정보전달 목적의 AI 인플루언서 구축이 목표이기 때문에 정확한 정보를 학습시키는 것이 중요하다고 생각했습니다. 한국은행은 자본이 개입되지 않은 독립적인 특수 법인이라 편향성이 거의 없는 데이터를 제공해 줄 수 있을 것으로 기대했습니다. 한국은행은 일반 대중들을 위해 전자책 형태의 pdf 문서를 제공합니다. 저희 팀은 그 중에서 '경제금융용어 700선'과 '알기 쉬운 경제 이야기(일반인용)'을 선택했습니다. 알기 쉬운 경제 이야기 시리즈 중에서 초/중/고등학생 용도로 쓰여진 것을 선택하지 않은 이유는 정보 압축률 때문입니다. 초/중/고등학생 입장에서 서술한 자료는 쉽게 설명하기 위해, 단순 개념을 타겟으로 길게 서술합니다. 저희는 AI에게 압축적으로 지식과 맥락을 학습시키기 위해 일반인 입장에서 쓰여진 책을 선택했습니다. 또한 일반인 입장에서 서술된 책은 '경제금융용어 700선'과 지식수준을 연계하기 좋을 것이라 판단했습니다. AI에 최대한 압축적으로 경제 지식을 습득시킨 후에 파인튜닝과 프롬프트 엔지니어링을 활용하여 읽기 쉬운 서술체로 변환시킴으로써, 고난도의 지식을 알고 있지만 쉽게 설명할 수 있는 '경제 정보 제공 목적의 AI 인플루언서'를 제작할 수 있을 것으로 기대합니다. 이와 같은 판단 하에 '경제금융용어 700선'과 '알기 쉬운 경제이야기 : 일반인용' 전자책 pdf 문서를 사용 데이터로 선정했습니다.				
데이터 출처	**[경제금융용어 700선]** https://www.bok.or.kr/portal/bbs/B0000216/view.do?nttId=10075809&type=&searchOptn8=22&menuNo=200134&listType=G&pageIndex=1 [알기 쉬운 경제이야기 : 일반인용] https://www.bok.or.kr/portal/bbs/B0000249/view.do?nttId=10061816&searchCnd=1&searchKwd=&depth2=201157&depth3=201269&depth4=200765&depth=200765&pageUnit=10&pageIndex=1&programType=newsData&menuNo=200765&oldMenuNo=201269				
원본 데이터	 **< 경제금융용어 700선 >**		 **< 알기 쉬운 경제이야기 : 일반인용 >**		
데이터 전처리 시 유의해야 할 점	'알기 쉬운 경제이야기' 문서의 경우, 날개 부분에 본문 내용을 보충하는 용어 설명 파트가 있습니다. 그대로 pdf 문서를 읽어 올 경우에 본문 내용과 날개 속 용어 설명이 섞일 가능성이 존재합니다. 따라서 각 부분을 분리해 주는 작업을 수행해야 합니다. '경제금융용어 700선' 문서는 단순 사전 형태로 구성되어 있어서 전처리할 부분이 비교적 적은 편에 속하지만, '알기 쉬운 경제이야기'는 여백에 용어 설명을 서술하고 있으므로 전처리 시 고려해야 할 사항이 확고하게 존재합니다.				