

프로젝트 방법론 - 1조

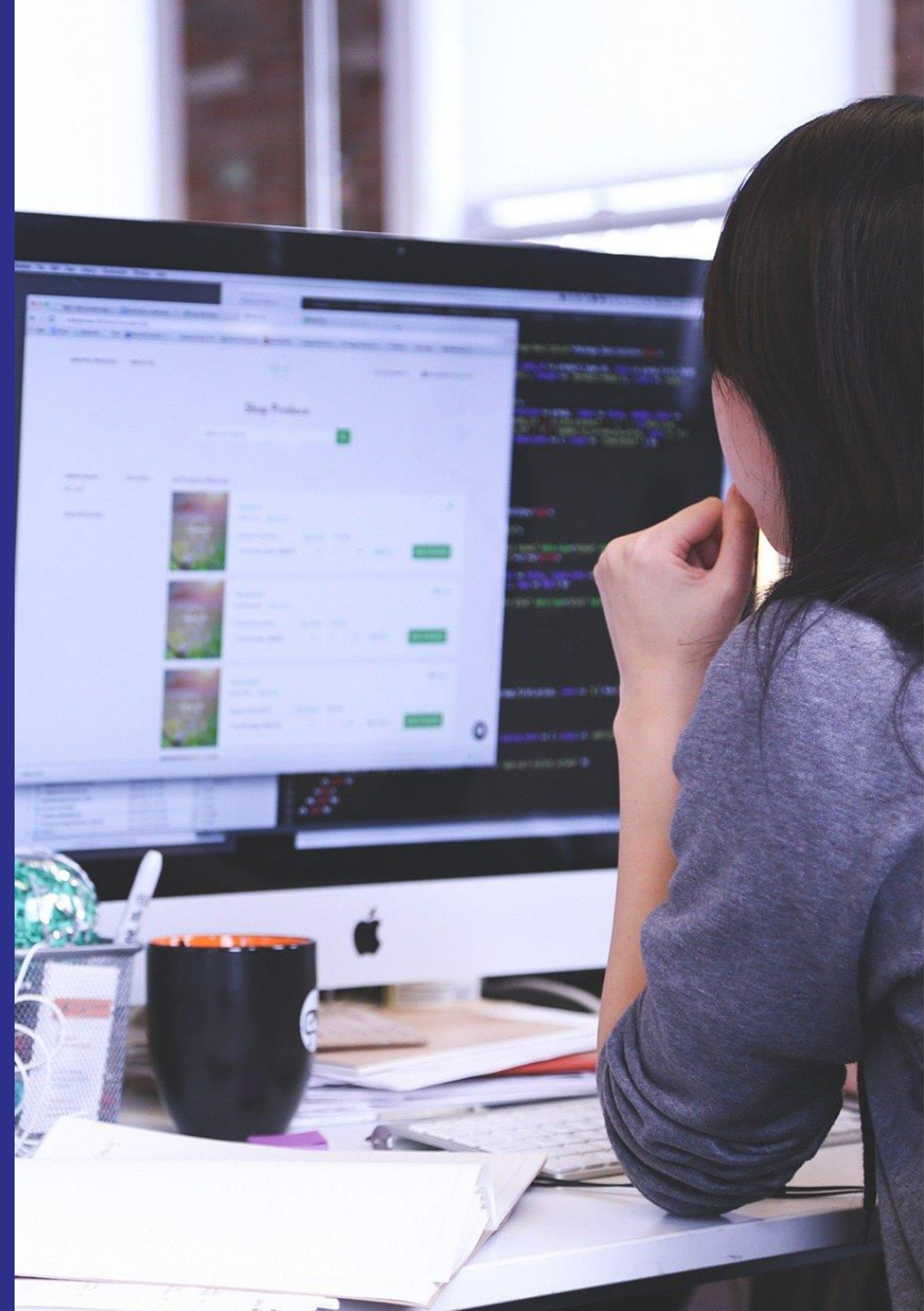
뉴스
분류 및 요약
최종 발표 자료



Contents

- 1-1. 프로젝트 추진 개요
- 1-2. 프로젝트 구축 범위
- 1-3. 프로젝트 조직 및 역할
- 1-4. 프로젝트 구성도
- 1-5. 시퀀스 다이어그램
- 1-6. WBS
- 1-7. 프로젝트 일정
- 1-8. 예상 이슈

- 2-1. 요구사항 정의서
- 2-2. 테이블 목록
- 2-3. 테이블 정의서
- 2-4. 회의록
- 2-5. 협업 도구
- 2-6. 시스템 코드



동아일보

PICK ①

바이든 “아마겟돈 올수 있다”... 푸틴의 핵위협에 경고

입력 2022.10.08 오전 3:01 기사원문

문병기 기자 >

66
 158

“쿠바 미사일 위기이후 최대 위험”
러 우크라 핵공격맨 직접개입 시사

IBM 찾아 양자컴퓨터 실험보는 바이든 조 바이든 미국 대통령(왼쪽)이 6일(현지 시간) 뉴욕주 퍼킨스 IBM 연구센터를 방문해 양자컴퓨터를 살펴보고 있다. 2천법을 쓰는 기존 디지털 컴퓨터와 달리 양자 정보 기본 단위 큐비트를 사용하는 양자컴퓨터는 슈퍼컴퓨터로 100만 년 이상 걸리는 연산을 10시간 만에 풀 수 있을 정도로 정보 처리 능력이 월등해 ‘게임을제너’ 기술을 주목 받고 있다. IBM은 이날 반도체 제조와 연구개발을 위해 10년간 200억 달러(약 28조 원)를 투자하겠다고 발표했다. 퍼킨스=AP 뉴스

볼로디미르 젤렌스키 우크라이나 대통령은 이날 “나토는 러시아가 핵무기를 사용할 가능성을 완전히 제거해야 한다”며 “선제 타격이 필요하다”고 주장했다.

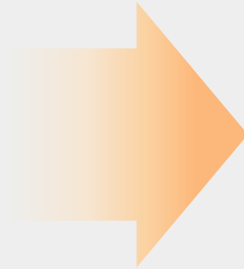
크라이나 점령지 4곳 합병 조약을 체결하
| 사용 가능성을 내비친 이후 바이든 대통

#사회, #IT, #우주비행사, #국제 우주정거장, #러시아 우주비행사

1-1. 프로젝트 추진 개요

배경

- 콘텐츠 소비 문화가 점점 짧고 빠른 소비 형태로 변화하고 있다.
- 뉴스 기사는 이러한 트렌드 변화를 따라가지 못하고 지속적으로 이용자를 잃고 있다.
- 새로운 트렌드에 맞춘 간결한 형태의 뉴스 기사에 대한 수요가 생겼다.
- 재한 외국인의 꾸준한 증가로 200만명 이상의 외국인이 국내에 체류하고 있다.
- 외국인들은 뉴스를 정확하고 빠르게 전달 받을 수 있는 방법이 없다.



목적

분류

- 여러 언론사의 기사를 통일된 기준으로 분류하여 사용자의 뉴스 분류 파악 및 선택에 편리함 제공

요약

- 뉴스 본문의 핵심 키워드로 간결하고 핵심적인 뉴스 정보 제공

번역

- 다국적 언어로 뉴스를 번역하여 외국인 이용자에게 핵심 뉴스 제공

1-2. 프로젝트 구축 범위

소스 데이터

- YNAT 데이터셋
- 네이버 뉴스 크롤링

구축 범위

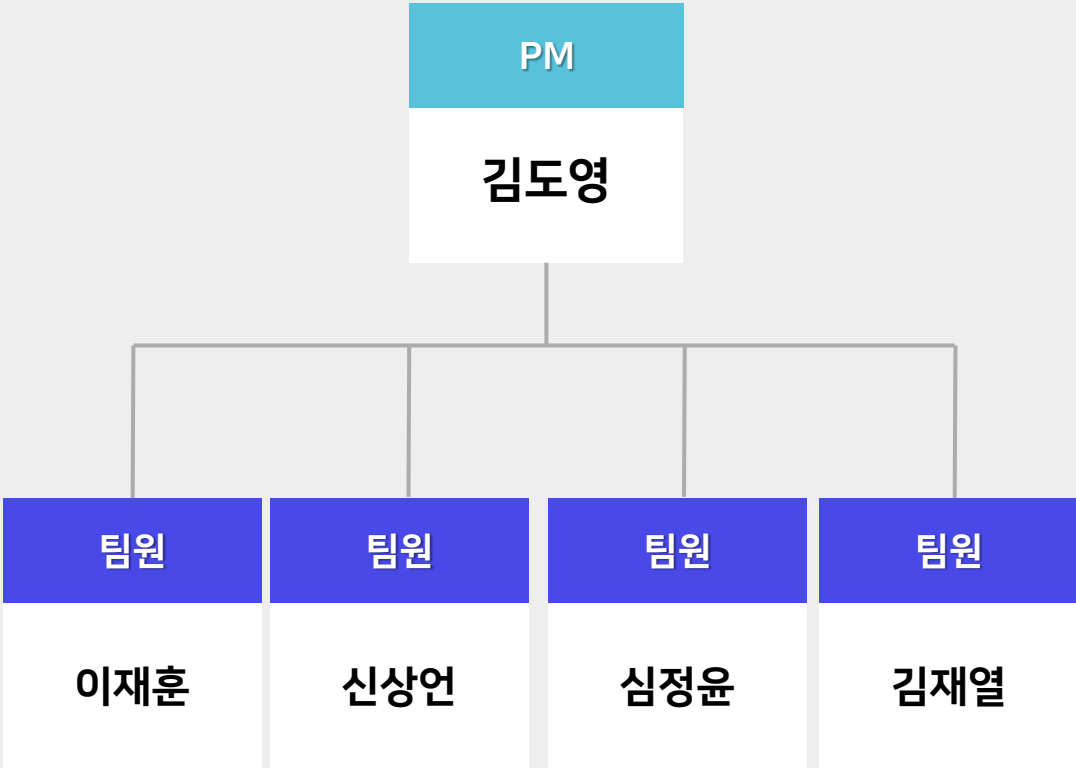
- 뉴스 기사 재분류 및 요약
 - 국내 신뢰도 1~20위의 언론사 뉴스 활용
- 뉴스 기사 번역
 - 영어, 중국어 등 5개 국어

기대 효과

- 커스터마이징 된 뉴스기사
 - 중요 뉴스 내용 및 최신 정보 습득 시간 감소
 - 뉴스 기사 열독자 증가
- 뉴스 기사 번역 서비스
 - 재한 외국인의 정보 불균형 해소

1-3. 프로젝트 조직 및 역할

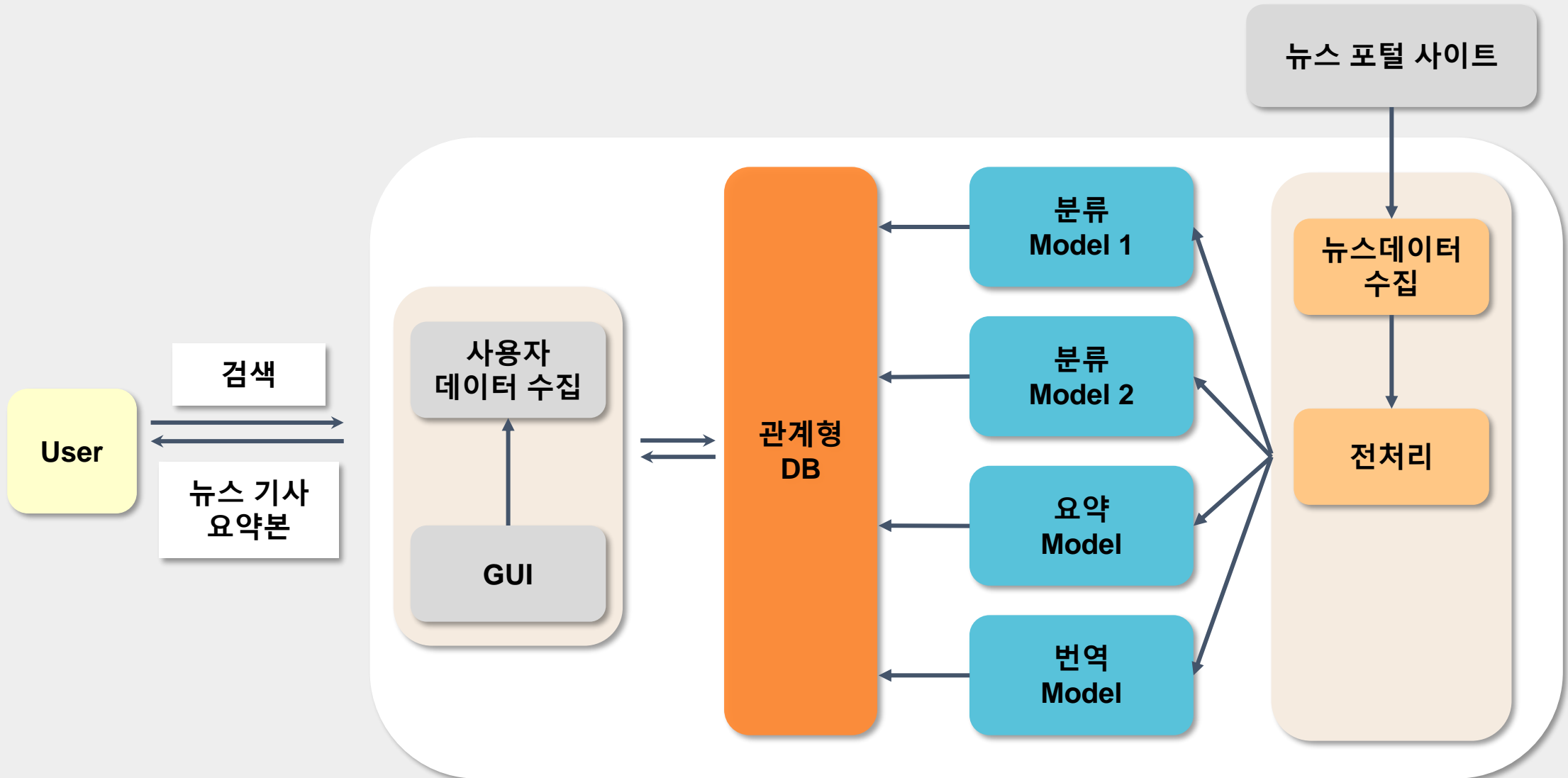
프로젝트 조직도



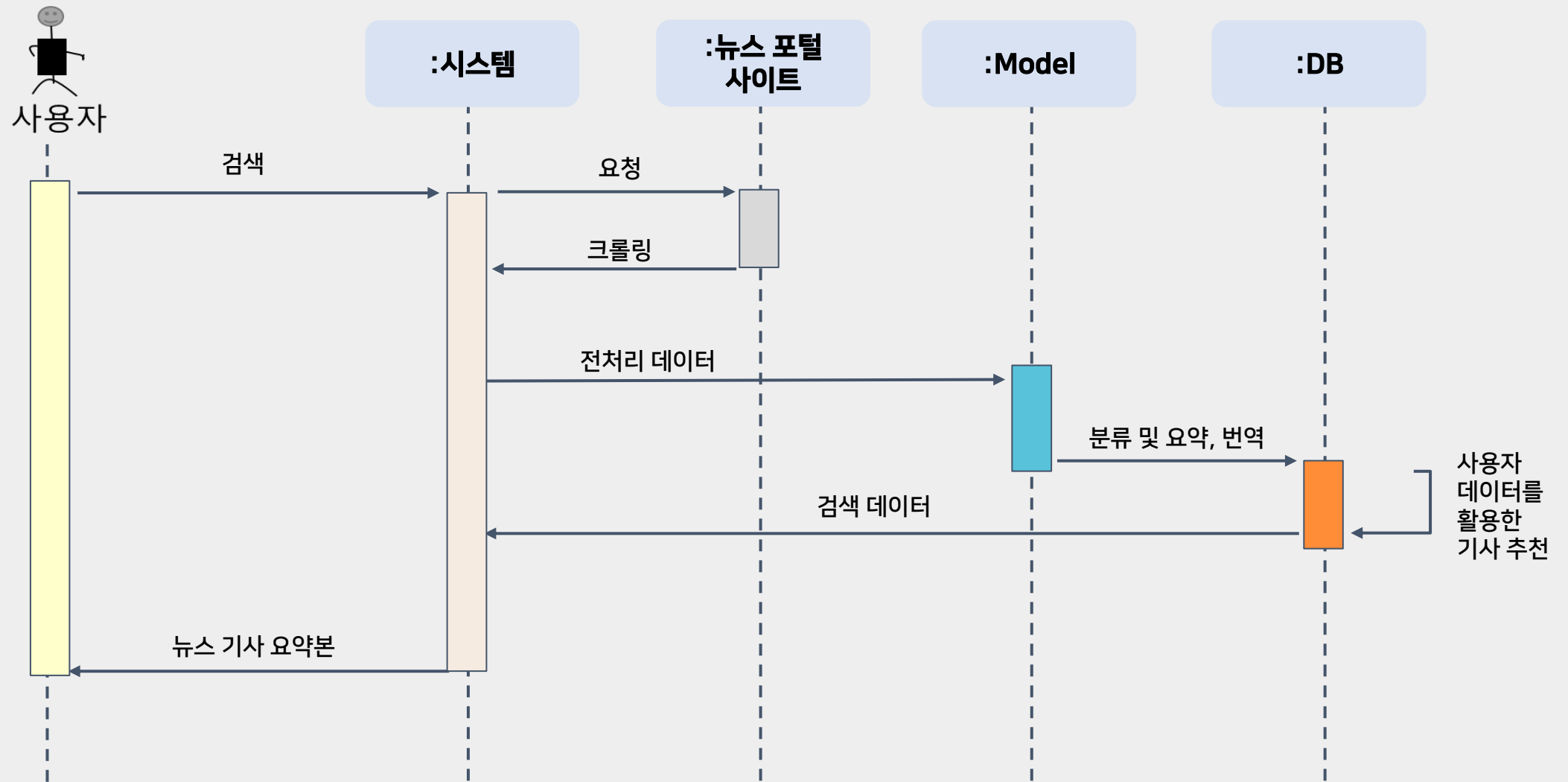
역할 분담표

구분	Role & Responsibilities	
	담당자	역할
프로젝트 관리	김도영	프로젝트 총괄
데이터 수집	이재훈	신규 데이터 크롤링
분류 모델 개발 (LSTM)	신상언	뉴스 제목을 통한 뉴스 분류 LSTM 모델 개발
분류 모델 개발 (Bert)	심정윤	뉴스 제목을 통한 뉴스 분류 Bert 모델 개발
요약 모델 개발 (SBERT)	김재열	뉴스 본문 내용 요약 모델 개발
테스트	전원	요구사항 충족 여부 확인

1-4. 프로젝트 구성도



1-5. 시퀀스 다이어그램



1-6. WBS (작업 분할 구조도)

2022년 10월

화 4

수 5

목 6

금 7

토 8

일 9

1-1. 사용자 개인정보 수집 김도영 2

1-2. 사용자 관심정보 수집 김도영 2

1-3. 사용자 요청(검색)정보 수집 김도영 2

2-1. IT 뉴스 기사 수집 이재훈 1

2-2. 경제 기사 수집 이재훈 1

2-3. 사회 기사 수집 이재훈 1

2-4. 생활문화 기사 수집 이재훈 1

2-5. 국제 기사 수집 이재훈 1

2-6. 스포츠 기사 수집 이재훈 1

2-7. 정치 기사 수집 이재훈 1

3-1. 분류 전처리 심정윤 신상언 1

3-2. 요약 전처리 김재열 1

4-2. 사용자 데이터 적재 김도영 2

10-1. DB 관리 김도영 2

5-1. 분류 모델 전처리 데이터 입력 심정윤 신상언 1

6-1. 요약 모델 전처리 데이터 입력 김재열 1

5-2. 분류값 출력 심정윤 신상언 1

6-2. 요약값 출력 김재열 1

7-1. 번역 모델 전처리 데이터 입력 심정윤 3



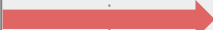
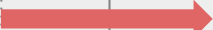
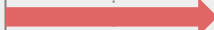
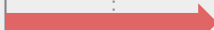


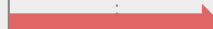

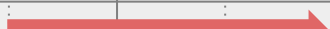
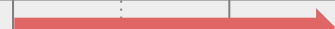
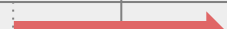

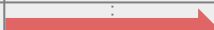
7-2. 번역값 출력 김재열 3

4-1. 분류 - 요약 데이터 적재 김도영 2

8-1. 검색어에 대한 뉴스의 분류-요약 정보 제공 심정윤 신상언 김재열 1

9-1. 사용자 요청 언어로 번역 심정윤 3

1-7. 프로젝트 일정

단계	TASK	1일차	2일차	3일차	4일차	5일차	산출물
분석기획(Planning)							
	비즈니스 이해 및 범위설정						요구사항정의서
	프로젝트 정의 및 계획설정						프로젝트수행계획서,WBS
	프로젝트 위험계획 수립						위험목록/위험관리계획서
데이터준비(Data Preparation)							
	필요데이터 정의						데이터정의서, 획득계획서
	데이터스토어설계(정형,비정형)						스토어설계서, 매핑정의서
	데이터수집 및 정합성 검증						데이터 정합성검증보고서
데이터분석(Data Analyzing)							
	분석용 데이터준비						분석용 데이터셀
	텍스트 분석						텍스트분석보고서
	탐색적분석						데이터탐색/시각화보고서
	모델링						모델링결과보고서
	모델평가 및 검증						모델평가보고서
시스템 구현(System developing)							
	설계 및 구현						구현시스템
	시스템테스트 및 운영						매뉴얼(사용자,운영자)
평가및 전개(Deploying)							
	모델발전계획 수립						발전계획서
	프로젝트 평가 및 보고						완료보고서

1-8. 예상 이슈

No	예상 이슈	대응 방안
1	언론사간 중복 뉴스 발생	수집 기간 늘려 다양한 데이터 확보
2	뉴스 기사의 특정 패턴, 노이즈 학습으로 과적합 발생	데이터양 증가 및 모델 리모델링
3	데이터 수집과 모델링 담당자간 소통 문제	실시간 협업 도구(Slack) 활용

2-1. 요구사항 정의서

요구사항 정의서					
구분	서비스(메뉴)	기능명	기능설명	우선순위	담당자
데이터 수집	1. 사용자 데이터	1-1. 사용자 개인정보 수집	사용자 개인정보 수집	2	김도영
		1-2. 사용자 관심정보 수집	사용자 쿠키 수집	2	
		1-3. 사용자 요청(검색)정보 수집	사용자 검색어 수집	2	
	2. 모델학습 데이터	2-1. IT 뉴스 기사 수집	네이버, 구글, 다음 뉴스 크롤링	1	이재훈
		2-2. 경제 기사 수집	네이버, 구글, 다음 뉴스 크롤링	1	
		2-3. 사회 기사 수집	네이버, 구글, 다음 뉴스 크롤링	1	
		2-4. 생활문화 기사 수집	네이버, 구글, 다음 뉴스 크롤링	1	
		2-5. 국제 기사 수집	네이버, 구글, 다음 뉴스 크롤링	1	
		2-6. 스포츠 기사 수집	네이버, 구글, 다음 뉴스 크롤링	1	
		2-7. 정치 기사 수집	네이버, 구글, 다음 뉴스 크롤링	1	
데이터 관리	3.전처리	3-1. 분류 전처리	데이터 정제, 불용어 제거, 토큰화, 임베딩	1	심정윤, 신상언
		3-2. 요약 전처리	데이터 정제, 불용어 제거, 토큰화, 임베딩	1	김재열
	4.관계형 DB	4-1. 분류 - 요약 데이터 적재	식별조치한 분류-요약 데이터 DB 저장	2	김도영
		4-2. 사용자 데이터 적재	사용자 관심정보 및 요청정보 DB 저장	2	
데이터 예측 및 요약	5.분류 Model	5-1. 전처리 데이터 입력	LSTM 모델 및 Bert 모델을 활용한 뉴스 제목 분류	1	심정윤, 신상언
		5-2. 분류값 출력	분류된 정보 출력	1	
	6.요약 Model	6-1. 전처리 데이터 입력	KeyBert 모델을 활용한 뉴스 본문 요약	1	김재열
		6-2. 요약값 출력	요약된 정보 출력	1	
	7.번역 Model	7-1. 전처리 데이터 입력	seq2seq 모델을 활용한 번역	3	심정윤
		7-2. 번역값 출력	번역된 정보 출력	3	
서비스	8.뉴스 분류 - 요약 시스템	8-1. 검색어에 대한 뉴스의 분류-요약 정보 제공	분류 모델과 요약 모델을 활용한 간략화된 뉴스 기사 정보 제공	1	심정윤, 신상언, 김재열
	9.번역	9-1. 사용자 요청 언어로 번역	국내 거주 외국인들을 위한 번역 기능 제공	3	심정윤
관리자	10.사후관리	10-1. DB 관리	관계형 데이터 베이스 구축으로 데이터의 축척 및 관리 편의성 제공	2	김도영

2-2. 테이블 목록

테이블 목록

테이블 목록							
시스템명		뉴스 분류 및 요약 시스템		작성일	2022.10.07	작성자	김도영
No.	주제영역명	테이블ID	테이블명	길이	초기건수	최대건수	증가건수
1	시스템 관리	TB_SYS_USER	사용자 개인정보	50	10000	100000	1000
2	시스템 관리	TB_SYS_INTEREST	사용자 관심정보	20	1000	2000000	10000
3	시스템 관리	TB_SYS_SEARCH	사용자 검색정보	20	1000	2000000	10000
4	데이터 관리	TB_DT_DCTITLE	뉴스 기본정보	200	2000	500000	1000
5	데이터 관리	TB_DT_NVTITLE	뉴스 분류정보	50	2000	500000	1000
6	데이터 관리	TB_DT_CONTENT	뉴스 요약정보	30	2000	300000	500

2-3. 테이블 정의서

테이블 정의서								
주제영역명		시스템 관리	작성일	2022.10.07		작성자		김도영
테이블ID		TB_SYS_USER	테이블명		사용자 개인정보			
테이블설명		사용자 개인정보를 관리한다.						
No.	컬럼ID	컬럼명	타입	길이	NULL	KEY	DEFAULT	비고
1	user_id	사용자ID	Int		N	PK		
2	user_name	사용자 이름	Varchar	10	N			
3	user_age	사용자 나이	Int		N			
4	user_job	사용자 직업	Varchar	15	N			
5	user_adress	사용자 주소	Varchar	50	N			
업무규칙	수정규칙 : '개인정보' 수정 시 '사용자ID' 수정 금지!							

2-3. 테이블 정의서

테이블 정의서											
주제영역명		시스템 관리		작성일		2022.10.07		작성자		김도영	
테이블ID		TB_SYS_USER		테이블명			사용자 개인정보				
테이블설명		사용자 개인정보를 관리한다.									
No.	컬럼ID	컬럼명	타입	길이	NULL	KEY	DEFAULT	비고			
1	user_id	사용자ID	Int		N	PK					
2	user_name	사용자 이름	Varchar	10	N						
3	user_age	사용자 나이	Int		N						
4	user_job	사용자 직업	Varchar	15	N						
5	user_adress	사용자 주소	Varchar	50	N						
업무규칙	수정규칙 : '개인정보' 수정 시 '사용자ID' 수정 금지!										

테이블 정의서								
주제영역명		시스템 관리	작성일	2022.10.07		작성자		김도영
테이블ID		TB_SYS_INTEREST	테이블명		사용자 관심정보			
테이블설명		사용자 관심정보를 관리한다.						
No.	컬럼ID	컬럼명	타입	길이	NUL I	KEY	DEFAULT	비고
1	user_id	사용자ID	Int		N	FK		
2	user_bookmark	사용자 즐겨찾기	Varchar	20	N			
3	user_like	사용자 좋아요	Varchar	10	N			
4	user_bad	사용자 싫어요	Varchar	10	N			
업무규칙	수정규칙 : '관심정보' 수정 시 '사용자ID' 수정 금지!							

테이블 정의서										
주제영역명		시스템 관리		작성일		2022.10.07		작성자		김도영
테이블ID		TB_SYS_SEARCH		테이블명			사용자 검색정보			
테이블설명		사용자 검색정보를 관리한다.								
No.	컬럼ID	컬럼명		타입	길이	NUL I	KEY	DEFAULT	비고	
1	user_id	사용자ID		Int		N	FK			
2	user_search	사용자 검색목록		Varchar	20	N				
업무규칙	수정규칙 : '검색정보' 수정 시 '사용자ID' 수정 금지!									

테이블 정의서								
주제영역명		데이터 관리	작성일	2022.10.07	작성자		김도영	
테이블ID		TB_DT_DCTITLE	테이블명		뉴스 기본정보			
테이블설명		뉴스 기본 정보 및 ID를 관리한다						
No.	컬럼ID	컬럼명	타입	길이	NULL	KEY	DEFAULT	비고
1	news_id	뉴스ID	Int		N	PK		
2	title	뉴스 제목	Varchar	50	N			
3	content	뉴스 본문	Varchar	200	N			
업무규칙	수정규칙 : '뉴스 기본정보' 수정 시 '뉴스ID' 수정 금지!							

테이블 정의서								
주제영역명		데이터 관리	작성일	2022.10.07	작성자		김도영	
테이블ID		TB_DT_NVTITLE	테이블명		뉴스 분류정보			
테이블설명		분류 모델 1, 모델 2 의 결과 저장 및 관리						
No.	컬럼ID	컬럼명	타입	길이	NUL L	KEY	DEFAULT	비고
1	news_id	뉴스ID	Int		N	FK		
2	topic_1	분류 1	Char	50	N			IT과학:0
3	topic_2	분류_2	Char	50	N			IT과학:0
업무규칙	수정규칙 : '뉴스 분류정보' 수정 시 '뉴스ID' 수정 금지!							

테이블 정의서								
주제영역명		데이터 관리	작성일	2022.10.07	작성자		김도영	
테이블ID		TB_DT_CONTENT	테이블명		뉴스 요약정보			
테이블설명		요약 모델의 결과 저장 및 관리한다.						
No.	컬럼ID	컬럼명	타입	길이	NUL I	KEY	DEFAULT	비고
1	news_id	뉴스ID	Int		N	FK		
2	keyword_1	요약_1	Varchar	30	N			
3	keyword_2	요약_2	Varchar	30	N			
4	keyword_3	요약_3	Varchar	30	N			
업무규칙	수정규칙 : '뉴스 요약정보' 수정 시 '뉴스ID' 수정 금지!							

2-4. 회의록

회 의 록			
회의주제	주제 선정 및 역할 분담		
회의 일자/시간	2022.10.04 / 14:00~18:00	작성자	김도영
회의장소	강의실		
소 속	참석자 성명		
팀원	이재훈, 신상언, 심정윤, 김재열		
회의내용			
<div>1. 주제 및 데이터 선정</div> <div>2. 프로젝트 도구 활용</div> <div>3. 역할 분담</div> <div>4. 프로젝트 계획서 작성</div>			
회의 결과 및 향후 일정			
<div>1. 주제 : 뉴스 기사 요약 및 분류 데이터 : 뉴스 포털 크롤링</div> <div>2. GitHub : 프로젝트 형상 관리 도구, 링크 : https://github.com/knudascentists/Team1 Slack : 프로젝트 커뮤니케이션 도구, Notion : 프로젝트 문서 관리 도구 링크 : https://www.notion.so/1-2445136233d84b6ba927e9e5d9bbafab - Notion 활용하되 Googledrive를 최우선으로 활용</div> <div>3. 김도영: 프로젝트 관리 이재훈: 데이터 수집 및 전처리 신상언: LSTM 모델 개발 심정윤: Bert 모델 개발 김재열: 뉴스 요약 모델 개발</div> <div>4. 목적, 기대효과, 구성도, 예상 이슈 작성</div>			

회 의 록			
회의주제	예상 이슈 및 방향성		
회의 일자/시간	2022.10.05 / 14:00~21:00	작성자	김도영
회의장소	강의실		
소 속	참석자 성명		
팀원	이재훈, 신상언, 심정윤, 김재열		
회의내용			
<div>1. 예상 이슈 및 대응 방안</div> <div>2. 데이터 수집 및 모델 성능</div> <div>3. 프로젝트 구성도</div> <div>4. 요구사항 정의서 작성</div>			
회의 결과 및 향후 일정			
<div>1. 팀원별 예상 이슈에 대한 대응 방안 미리 대비</div> <div>2. 포털 사이트 보안 문제로 동적 웹크롤링으로 데이터 수집 팀원별 기본 모델 구성하여 성능 체크</div> <div>3. 어제 작성한 구성도를 수정하며 전체적인 틀 수정</div> <div>4. 미래에 대한 구체적인 방향성</div>			

2-4. 회의록

회 의 록			
회의주제	프로젝트 계획서 및 계획 구체화		
회의 일자/시간	2022.10.06 / 09:00~18:00	작성자	김도영
회의장소	강의실		
소 속	참석자 성명		
팀원	이재훈, 신상언, 심정윤, 김재열		
회의내용			
<div>1. 프로젝트 계획서, 요구사항 정의서 수정</div> <div>2. WBS</div> <div>3. 테이블 목록 및 정의서</div> <div>4. 모델 완성도 증가</div>			
회의 결과 및 향후 일정			
<div>1. 프로젝트 계획서 수정, 요구사항 정의서 완성</div> <div>2. WBS 작성을 하면서 작업 일에 맞추어 계층적/점진적인 일정으로 구성</div> <div>3. 테이블 목록 및 정의서를 데이터 균형이 이루어지고 있는 것을 확인</div> <div>4. 모델 정확도를 높이기 위해 불용어 설정 및 하이퍼 파라미터 튜닝</div>			

회 의 록			
회의주제	최종 발표 자료 및 최종 결과물 완성		
회의 일자/시간	2022.10.07 / 14:00~18:00	작성자	김도영
회의장소	강의실		
소 속	참석자 성명		
팀원	이재훈, 신상언, 심정윤, 김재열		
회의내용			
<div>1. 최종 발표 자료 작성</div> <div>2. 기존 자료에 미흡한 부분 수정</div> <div>3. 모델 완성도 증가</div> <div>4. 최종 결과물 테스트 <- 요구사항 충족하는가?</div>			
회의 결과 및 향후 일정			
<div>1. 최종 발표 자료 완료</div> <div>2. 전체적인 기존 자료 수정하여 완료</div> <div>3. 최선의 모델 완성</div> <div>4. 최종 결과물(시스템 구현) 완성</div>			

2-5. 협업 도구 - Slack

The screenshot displays the Slack interface for a workspace named '14_Team_01'. The left sidebar shows the channel list with '# 일반' (General) selected. The main area shows a conversation in the '# 일반' channel. The conversation includes messages from 김도영, Jh Lee, cd ab, and 심정윤. A thread is visible for the message from 김도영 at 1:52, titled 'PowerPoint 프레젠테이션'. The thread shows a response from 심정윤 and a link to a Notion page. The right sidebar shows the thread details for the selected message, including the message content and a link to the Notion page.

Left Sidebar (Channel List):

- 14_Team_01
- 스레드
- 멘션 및 반응
- Slack Connect
- 더 보기
- 채널
- # 랜덤
- # 일반 (Selected)
- 공부
- 채널 추가
- 다이렉트 메시지
- 김도영 나
- 김재열
- 심정윤
- cd ab
- Jh Lee
- Joonion Bae (joonion)
- 팀원 추가
- 무료 평가판 사용 중

Main Channel View (# 일반):

- 김도영 오후 1:10: 예상이슈 대응방안 알려주세요 ㅎㅎ
- Jh Lee 오후 1:17: 데이터 수집 담당) 예상 이슈 : 언론사 별로 동일한 뉴스가 올라오는 경우가 많아 수집된 데이터가 중복인 경우가 많을 것 같음 대응 방안 : 뉴스의 수집 기간을 최대한 늘려 중복되지 않은 데이터를 최대한 수집
- cd ab 오후 1:22: 예상 이슈 데이터의 양이 적어 특정 패턴이나 노이즈까지 쉽게 암기됨에 따라 과적합 현상이 발생할 확률이 높을 것이다
- 재열 오후 1:24: <기사요약>
 - 예상문제 okt.pos를 통해 품사 부착해 형태소 분리를 했을때 데이터 양에 따라 처리 시간이 길어질 수 있음.
 - 해결방안 'morphs'로 형태소를 추출하면 시간은 짧아지나 불용어를 정해줘야함. 샘플데이터로 처리시간 확인해서 처리
- cd ab 오후 1:25: 예상 방안 데이터 수집 담당과 지속적인 커뮤니케이션을 통한 데이터 수집과 모델 리모델링을 통한 모델 개선
- 심정윤 오후 1:25: 예상 이슈 : 데이터가 계속 추가됨에 따라 모델 성능에 영향을 미치는 요인이 부정확해질 수 있음 해결 방안 : 새로운 데이터 추가 전 동일한 모델로 결과 확인 후 모델 파라미터 변경
- 김도영 오후 1:25: 감사합니다 ㅎㅎ
- 김도영 오후 1:52: PowerPoint 프레젠테이션

Thread View (Threaded Conversation):

- 심정윤 오늘, 오전 3:23: Bert모델 합본입니다. 전체 실행 시 오류는 발생하지 않으나 Colab에서 실행 시 마지막 print에서 기사 제목이 제대로 나오지 않습니다. 현재 로컬에서 다시 돌려보고 있고 결과는 추후 업데이트하도록 하겠습니다. 또한, Notion에 회의록 업데이트 및 추가적인 수정을 하였습니다. 확인 부탁드립니다. <https://www.notion.so/a73d006b0294447d8c74b428b8c64d3e?v=3d47df52abca4a58be9ce770955f2ad6>
- 이전
- 01 merging.ipynb 이전

Right Sidebar (Thread Details):

- 심정윤 11시간 전: Bert모델의 epoch는 1로 설정되어있으며, Colab에서 실행 시 Colab=True로 수정 후 실행하시면 됩니다. (편집됨)

Bottom Bar:

- #일반(으)로도 전송

2-5. 협업 도구 - Github

프로젝트 완료를 위해 날짜별 진행사항 기록 및 공유

The screenshot shows a GitHub repository page for 'knudatascientists / Team1'. The repository is public and has 1 branch (main) and 0 tags. The file list shows several folders and a README.md file, all updated 10 minutes ago, except for README.md which was updated 10 hours ago. The README.md content is in Korean and describes a project methodology for analyzing news articles. The right sidebar shows repository statistics: 0 stars, 1 watching, 0 forks, and 5 contributors. The language statistics show 100.0% Jupyter Notebook.

Search or jump to... Pull requests Issues Marketplace Explore

knudatascientists / Team1 Public

Edit Pins Watch 1 Fork 0 Star 0

<> Code Issues Pull requests Actions Projects Wiki Security Insights Settings

main 1 branch 0 tags

Go to file Add file Code

About

No description, website, or topics provided.

Readme 0 stars 1 watching 0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 5

Languages

Jupyter Notebook 100.0%

Aravis0309 최종정리 e2fa25f 10 minutes ago 21 commits

__최종모델__	최종정리	10 minutes ago
도영(팀장---최종산출물)	최종정리	10 minutes ago
상언(분야분류-LSTM)	최종정리	10 minutes ago
재열(키워드_도출)	최종정리	10 minutes ago
재훈(크롤링)	최종정리	10 minutes ago
정윤(분야분류-Bert)	최종정리	10 minutes ago
README.md	Update README.md	10 hours ago

README.md

프로젝트방법론 1조

Notion : <https://www.notion.so/1-2445136233d84b6ba927e9e5d9bbafab>
GDrive : <https://drive.google.com/drive/u/0/folders/1-dtKFqXJqOMxqMU4fLUuz5HD3Wh2nXUu>

프로젝트 주제 : 새로운 콘텐츠 소비 문화에 맞는 뉴스 기사 커스터마이징

- 뉴스 기사 제목을 통한 주제 분류
- 뉴스 기사 본문 중 키워드 도출하여 요약
- 뉴스 기사 번역

2-5. 협업 도구 - notion

프로젝트방법론 1조

댓글 추가

GDive : <https://drive.google.com/drive/u/0/folders/1-dtKFqXJqOMxqMU4fLUuz5HD3Wh2nXUu>

GitHub : <https://github.com/knudatascientists/Team1>

팀원

김도영 : 팀장

이재훈 : 데이터 크롤링

신상연 : 뉴스 제목 분류 모델 개발 (LSTM)

심정윤 : 뉴스 제목 분류 모델 개발 (Bert)

김재열 : 뉴스 기사 요약 모델 개발

문서

요구사항정의서

WBS

보드

회의록

담당자

마감일

시작 전 0

진행 중 0

완료 13

자료 수집

자료 전처리

LSTM 모델 개발

Bert 모델 개발

정윤 심

숨긴 그룹

보관됨 0

Status 없음 0

회의록

전체

유형별

필터

정렬

...

새로 만들기

Aa 이름	유형	최종 편집자	참석자	+	...
10/08 회의록	FINAL	정윤 심	정윤 심 Jh Lee 재 재열 김 도영 도영 김 cd ab		
10/07 회의록	Daily	정윤 심	정윤 심 Jh Lee 재 재열 김 도영 도영 김 cd ab		
10/06 회의록	Daily	정윤 심	정윤 심 Jh Lee 재 재열 김 도영 도영 김 cd ab		
10/05 회의록	Daily	정윤 심	정윤 심 Jh Lee 재 재열 김 도영 도영 김 cd ab		
10/04 회의록	Initial	정윤 심	정윤 심 Jh Lee 재 재열 김 도영 도영 김 cd ab		

+

새로 만들기

커버 추가

댓글 추가

참고

(아래 목록을 클릭하시면 상세내용 페이지로 연결됩니다.)

<< 최종 산출물 목록 >>

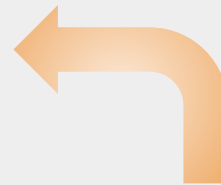
<< 요구사항 정의서 >>

<< PPT 폰트 다운로드 링크 >>

2-6. 데이터 크롤링

크롤링 모듈

```
def getTitles(date, type) :  
    # 리스트의 리스트에서 입력된 리스트를 불러오기  
    list_li = list_list[type]  
    num = list_num[type]  
    if num == 99 :  
        print("스포츠 뉴스는 다른 함수를 사용해주시길 바랍니다.")  
    else :  
        # 타이틀을 담은 리스트  
        list_title = []  
        for i in range(0, len(list_li)) :  
            print(i)  
            # 입력한 날짜의 리스트 페이지의 데이터를 저장  
            url = (f'https://news.naver.com/main/list.naver?mode=LS2D&sid2=731&mid=shm&sid1=105&date=20221003&page=1')  
            print(url)  
            html = urlopen(url) # url 주소 html로 저장  
            soup = BeautifulSoup(html.read(), 'html.parser') # html 데이터 BeautifulSoup으로 요약  
            text = soup.find_all('dt')  
            j = 1  
            for j in range(1, 40, 2) :  
                print(j)  
                try :  
                    te = text[j].text.strip()  
                    if len(te) >= 1 : # 내용이 있는 경우에만  
                        list_title.append(te)  
                        print(te)  
                except :  
                    continue  
            time.sleep(10)  
        # 데이터를 데이터 프레임으로 변환  
        df = pd.DataFrame(list_title)  
        # 데이터에 입력했던 라벨 지정  
        df['label'] = type  
        return df
```



```
# 입력한 날짜의 리스트 페이지의 데이터를 저장  
driver = webdriver.Chrome("./chromedriver")
```

```
driver.get(f"https://news.naver.com/main/list.naver?mode=LS2D&sid2={list_li[i]}&mid=shm&sid1=10{num}&date={date}&page=1")  
driver.implicitly_wait(3)
```

```
html = driver.page_source
```

2-6. 분류 모델 - LSTM

전처리 모듈

```
def makeTextlist(data):
    stopwords_01 = ['의', '가', '이', '은', '들', '는', '좀', '잘', '강', '과', '도', '를', '으로', '자', '에', '와', '한', '하다']
    okt = Okt()
    text_list = []
    for text in tqdm(data['title']):
        text = re.sub(r"[^uAC00-uD7A30-9a-zA-Z#s]", " ", text) # 특수문자 제거
        text = text.strip() # 문자 처음과 끝 공백 제거
        tokens = okt.morphs(text) # 단어 추출
        text = [word for word in text if not word in stopwords_01] # 불용어 처리
        text = " ".join(text)
        text = text.replace(' ', ' ')
        text_list.append(text)

    data["title"] = text_list
```

```
makeTextList(data)
```

100% ██████████ | 63931/63931 [02:25<00:00, 439.59it/s]

```
data['title']
```

0 인천 핀란드 항공기 결항 휴철 여행객 분통
1 실리곤밸리 넘어서겠다 구글 15조원 여 전역 거점화
2 란 외무 긴장완화 해결책 미국 경제전쟁 멈추 것
3 NYT 클린턴 측근 기업 특수관계 조명 공 사 맞물려종합
4 시진핑 트럼프 중미 무역협상 조속 타결 희망

모델 모듈

```
# 양방향 LSTM
def create_model():

    model1=Sequential()
    model1.add(Embedding(10000,64,input_length=sent_length))
    model1.add(Bidirectional(LSTM(50)))
    model1.add(Dropout(0.3))
    model1.add(Dense(7,activation='softmax'))
    model1.compile(loss='CategoricalCrossentropy',optimizer='adam',metrics=['accuracy'])
    print(model1.summary())
    return model1
```

```
model1 = create_model()
```

Model: "sequential 2"

Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, 27, 64)	640000
bidirectional_2 (Bidirectional)	(None, 100)	46000
dropout_2 (Dropout)	(None, 100)	0
dense_2 (Dense)	(None, 7)	707

```
Total params: 686,707
Trainable params: 686,707
Non-trainable params: 0
```

None

2-6. 분류 모델 - Bert

전처리 모듈

```
1 tokenizer_bert= BertTokenizer.from_pretrained("bert-base-multilingual-cased",
2       | cache_dir='bert_ckpt', do_lower_case=False)
3
4 def bert_tokenizer(stc, MAX_LEN):
5
6     encoded_dict = tokenizer_bert.encode_plus(
7         text = stc,
8         add_special_tokens = True,      # Add '[CLS]' and '[SEP]'
9         max_length = MAX_LEN,          # Pad & truncate all sentences.
10        pad_to_max_length = True,
11        return_attention_mask = True
12    )
13
14
15    input_id = encoded_dict['input_ids']
16    attention_mask = encoded_dict['attention_mask']
17    token_type_id = encoded_dict['token_type_ids']
18
19    return input_id, attention_mask, token_type_id
```

모델 모듈

```
1 # MODEL CLASS -----
2 class TFBertClassifier(tf.keras.Model):
3     def __init__(self, model_name, dir_path, num_class):
4         super(TFBertClassifier, self).__init__()
5
6         self.bert = TFBertModel.from_pretrained(model_name, cache_dir=dir_path)
7         self.dropout = tf.keras.layers.Dropout(self.bert.config.hidden_dropout_prob)
8         self.classifier = tf.keras.layers.Dense(num_class,
9             kernel_initializer=tf.keras.initializers.TruncatedNormal(self.bert.config.initializer_range),
10            name="classifier")
11
12     def call(self, inputs, attention_mask=None, token_type_ids=None, training=False):
13
14         #outputs 값: # sequence_output, pooled_output, (hidden_states), (attentions)
15         outputs = self.bert(inputs, attention_mask=attention_mask, token_type_ids=token_type_ids)
16         pooled_output = outputs[1]
17         pooled_output = self.dropout(pooled_output, training=training)
18         logits = self.classifier(pooled_output)
19
20         return logits
21 # -----
22
23 model2 = TFBertClassifier(model_name='bert-base-multilingual-cased',
24       | dir_path='bert_ckpt',
25       | num_class=7)
26
27 optimizer = tf.keras.optimizers.Adam(3e-5)
28 loss = tf.keras.losses.SparseCategoricalCrossentropy(from_logits=True)
29 metric = tf.keras.metrics.SparseCategoricalAccuracy('accuracy')
30
31 model2.compile(optimizer=optimizer, loss=loss, metrics=[metric])
```


2-6. 요약 모델 - SBERT

모델 모듈

Max Sum Similarity

- 데이터 쌍 사이의 최대 합 거리는 데이터 쌍 간의 거리가 최대화되는 데이터 쌍으로 정의
- 후보 간의 유사성을 최소화하면서 문서와의 후보 유사성을 극대화

```
def max_sum_sim(doc_embedding, candidate_embeddings, words, top_n, nr_candidates):
    # 문서와 각 키워드들 간의 유사도
    distances = cosine_similarity(doc_embedding, candidate_embeddings)

    # 각 키워드들 간의 유사도
    distances_candidates = cosine_similarity(candidate_embeddings,
                                           candidate_embeddings)

    # 코사인 유사도에 기반하여 키워드들 중 상위 top_n개의 단어를 pick.
    words_idx = list(distances.argsort()[0][-nr_candidates:])
    words_vals = [candidates[index] for index in words_idx]

    # (참고) numpy ix_ https://bit.ly/3CBFrUa , explain: can quickly construct index arrays that
    distances_candidates = distances_candidates[np.ix_(words_idx, words_idx)]

    # 각 키워드들 중에서 가장 덜 유사한 키워드들간의 조합을 계산
    min_sim = np.inf
    candidate = None
    # combination(iterable, r) => iterable에서 원소 개수가 r개인 조합 뽑기
    for combination in itertools.combinations(range(len(words_idx)), top_n):
        sim = sum([distances_candidates[i][j] for i in combination for j in combination if i != j])
        if sim < min_sim:
            candidate = combination
            min_sim = sim

    return [words_vals[idx] for idx in candidate]
```

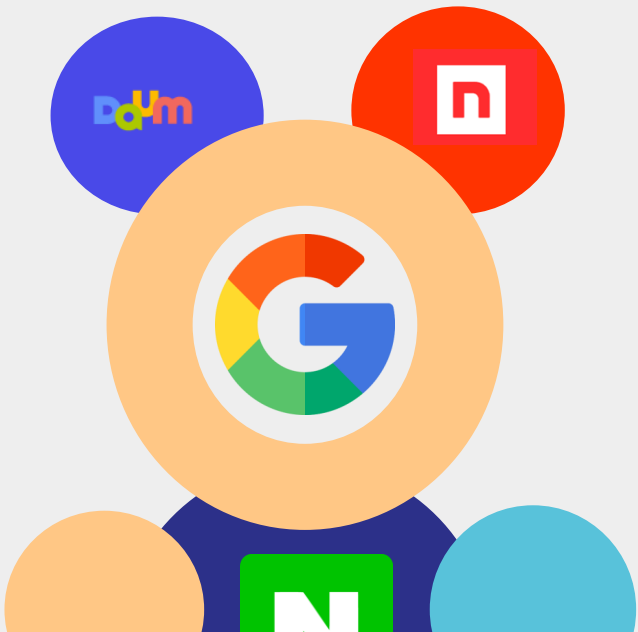
모델 결과

결론 : 상위 10개 키워드중 유사성이 낮은 n개를 선택해 다양한 키워드 도출

```
# 상위 10개의 키워드를 선택하고 이 10개 중에서 서로 가장 유사성이 낮은 3개를 선택
# 상대적으로 높은 nr_candidates는 다양한 키워드 3개를 도출
max_sum_sim(doc_embedding, candidate_embeddings, candidates, top_n=3, nr_candidates=30)
```

['재보궐선거 차례 서울', '대가 불법 정치자금', '수십 차례 형의']

QnA



Thank you

