

그림차례	iii
------------	-----

표차례	v
-----------	---

제1장	서론	3
------------	-----------------	----------

1.1 연구 배경 및 필요성	3
1.2 연구 범위 및 내용	5
1.3 검색 UI 개발 배경 및 필요성	6
1.4 검색 UI 서비스 내용	7
1.5 연구 기대 효과	8

제2장	재난안전 지식베이스 데이터베이스 시범구축	11
------------	-------------------------------------	-----------

2.1 뉴스정보 데이터베이스 저장 방안 소개	11
2.2 국내외 뉴스 수집 결과	36
2.3 DCAT 3.0 활용 가능성	38

제3장	해외뉴스 수집 · 분석 전략 마련	43
------------	---------------------------------	-----------

3.1 해외뉴스 수집 방법	43
3.2 해외뉴스 수집 체계 및 결과 제시	63
3.3 수집한 영문 뉴스 대상 자동 번역 방안 및 결과 제시	65
3.4 향후 데이터 수집 규모에 따른 필요 예산안 산출	67

제4장 PDF 형태의 자료 내 재난안전정보 추출 · 분석 73

- 4.1 처리 대상 PDF 자료 개요 73
- 4.2 PDF 자료 내 텍스트 정보 추출 방안 81
- 4.3 추출 텍스트 저장 결과 85
- 4.4 재난안전 관련 키워드 제시 94

제5장 재난안전 지식베이스 검색 UI 프로토타입 개발 105

- 5.1 검색 웹페이지에 제공될 검색 기능 105
- 5.2 수집된 뉴스 데이터 기반 효과적인 모니터링 도구 개발 결과 127
- 5.3 영문 뉴스 대상 잠재적 재난 뉴스 탐색을 위한 LLM 활용 방법 130

제6장 결론 141

참고문헌 145

그림 1.1 재난안전분야 지식베이스 검색UI 구성의 개념도	5
그림 2.1 재난안전 지식베이스를 위한 데이터베이스 구축	13
그림 2.2 사건의 주요 구성요인 추출을 통한 사건정보 DB 구축	15
그림 2.3 사건정보 DB 테이블 구성 : doc_detail_event_info	15
그림 2.4 사건정보 DB 테이블 구성 : doc_detail_event_link_info	17
그림 2.5 사건정보 DB 테이블 구성 : doc_detail_event_modifier_info ...	19
그림 2.6 사건정보 DB 테이블 구성 : doc_detail_event_negation_info ...	20
그림 2.7 사건정보 DB 테이블 구성 : doc_detail_event_object_info	22
그림 2.8 사건정보 DB 테이블 구성 : doc_detail_event_predicate_info ..	24
그림 2.9 사건정보 DB 테이블 구성 : doc_detail_event_subject_info	25
그림 2.10 사건정보 DB 테이블 구성 : doc_detail_event_timex_info	27
그림 2.11 사건정보 DB 테이블 구성 : doc_detail_event_spatial_info	29
그림 2.12 재난안전 분야 사건의 주요 구성요인 추출	32
그림 2.13 사건정보 DB 정보 구성 방안	34
그림 2.14 재난안전 지식베이스 지도 반영 개요	36
그림 2.15 국내뉴스 적제 데이터 테이블	38
그림 2.16 사건정보 관련 영역 DCAT 3.0 추가	39
그림 3.1 해외뉴스 수집 체계 및 데이터 수집 기능 설계	45
그림 3.2 샘플 API 요청 결과	57
그림 3.3 News API 가격정책(월간)	62
그림 4.1 PDF 형태의 자료 내 재난안전정보 추출 분석	75
그림 4.2 “감사연보” PDF 텍스트 추출 대상 데이터 예시	79
그림 4.3 “국정감사보고서” PDF 텍스트 추출 대상 데이터 예시	80

그림 4.4 PDF 자료 내 텍스트 정보 추출 방안 개요	82
그림 4.5 원본 PDF 메타정보 데이터 테이블	88
그림 4.6 PDF TXT 텍스트 추출 메타정보 데이터 테이블	90
그림 4.7 PDF XML 텍스트 추출 메타정보 데이터 테이블	91
그림 4.8 PDF 이미지 텍스트 추출 메타정보 데이터 테이블	93
그림 4.9 재난안전 유형별 키워드 자동추출 기술	94
그림 4.10 주요 키워드 간 연관도 분석 기능	97
그림 4.11 클러스터링 알고리즘	99
그림 4.12 재난안전 관련 문서 주제 분류	100
그림 5.1 지식베이스 검색 개념(안)	111
그림 5.2 지식베이스 상세 검색(안)	113
그림 5.3 검색 UI 설계안 : 데이터 검색 및 재난분류 선택 기능	117
그림 5.4 검색 UI 설계안 : 상세검색(검색기간, 범위, 검색어조합)	119
그림 5.5 검색 UI 설계안 : 국내뉴스 검색결과	121
그림 5.6 검색 UI 설계안 : 국내뉴스 상세검색결과	123
그림 5.7 국회 행정안전위원회 국정감사보고서 검색결과 화면	125
그림 5.8 국회 행정안전위원회 국정감사보고서 상세검색결과 화면	127
그림 5.9 재난안전 뉴스 모니터링 화면안	128
그림 5.10 LLM 기반 최신정보 우선순위 선정 방법 개요	131
그림 5.11 LLM 도입시 고려사항	133
그림 5.12 거대언어모델 아키텍처 적용 방법	135
그림 5.13 LLM(대규모 언어모델) 적용 : LLM기반 검색 화면 시안	137

표 3.1 해외뉴스 수집을 위한 뉴스제공 API 조사 항목	48
표 3.2 News API 주요 기능	50
표 3.3 News API 구성요소	52
표 3.4 News API 요청 Parameter	54
표 3.5 샘플 API 요청 결과 파싱(parsing) 예시	60
표 3.6 해외 뉴스 데이터 수집 및 번역 시스템 운영 예산안(1안)	67
표 3.7 해외 뉴스 데이터 수집 및 번역 시스템 운영 예산안(1안)	67
표 4.1 PDF 형태의 자료 내 재난안전정보 추출 분석	77
표 4.2 형식 교정 예시	83
표 4.3 PDF 자료 변환 대상 목록	86
표 5.1 재난안전 지식베이스 데이터베이스 설계(안)	109

"시간", 공간적 범위를 지칭하는 "공간", 감정 상태를 형으로 구분된다. 수식표현 정보는 해당 수식표현가 하며, 예를 들어 강도의 경우 "강" 또는 "약", 시간의 경우 같은 정보를 제공한다.

특성을 보다 세부적으로 분류하고 설명할 수 있도록 채풍이 전국적으로 심각한 피해를 입혔다"라는 문장에서 "같은 사건의 시간적 맥락을, "전국적으로"는 공간적 맥락을 나타내는 수식표현로 해석될 수 있다. 이처럼 수식표현을 다각도로 분석하여 재난 상황의 의미를 구체화하고, 사용자에게 유용한 정보를 제공한다.

데이터베이스에서 수식 표현은 사건의 맥락과 의미를 보다 명확히 파악하는 데 중요한 도구로 기능한다. NLP 기술은 수식 표현을 활용해 사건의 스케일, 강도, 시간적 및 공간적 범위를 분류할 수 있으며, 이는 데이터 기반 의사결정의 핵심

프로젝트명 : 재난안전 지식베이스 구축 요소기술 및 검색 UI 개발				
테이블 코드 : doc_detail_event_negation_info				
테이블명 : 문서의 사건 negation 정보				
No.	필드 ID	Key	Null	필드명
1	doc_uid	PK	N	사건 정보를 담은 문서를 구분하기 위한 고유 식별자입니다. 하나의 문서는 여러 문장으로 구성됩니다.
2	sent_uid	PK	N	문서 내 특정 문장을 식별하기 위한 고유 식별자입니다. 문서와 문장의 계층적 관계를 나타냅니다.
3	doc_date	PK	N	문서가 작성된 날짜로, 사건의 시간적 맥락을 제공합니다.
4	negation_idx	PK	N	문장에서 부정어가 등장한 위치(단어 인덱스)에: 5(문장의 다섯 번째 단어에 부정어가 등장)
5	position	PK	N	부정어가 포함된 텍스트의 시작과 끝 위치(문장 내 인덱스): 10-35 (10번째 문자부터 35번째 문자까지가 부정어 표현)
6	negation_info			부정어를 포함한 텍스트 내용 또는 부정어 유형: 예: "하지 않았다", "없다", "불가능하다".

테이블 구성 : doc_detail_event_negation_info

스 상에서 사건 정보를 처리할 때, 부정어(Negation)는 : Negation은 특정 문장에서 사실로 보이는 정보를

TABLE 1			
2	sent_uid	PK	N
3	doc_date	PK	N
4	predicate_idx	PK	N
5	position	PK	N
6	predicate_info		

케아

스어

는

이

가

타

과

on)오

이터

각

조

장

수

의

은 "

치한

다.

함

ite

사건의 핵심 동작이나 상태를 명시적으로 "침수되다"와 같은 동사형 표현이 이에

NLP 관점에서 Predicate 서술어는 ‘부여(Semantic Role Labeling)를 연결하여 사건의 구조를 파악할 수 있는 문장에서 "파괴하다"는 서술어로, "산지" 관계를 추출(Relation Extraction)을 통해 관계를 도출할 수 있다. 예컨대, "태풍<피해 → 도시"와 같은 관계를 <Extraction)에서는 Predicate를 통해 "발생했다"라는 문장에서 "화재 발생" 데이터 요약(Text Summarization) 내용을 간결하게 표현할 수 있다.

따라서 재난안전 사건정보 데이터베이스를 이해하고 분석하는 데 중요한 역할을 작업을 수행할 수 있다. 이 데이터를 효율적으로 구조적 정보를 효과적으로 추출하고, 사방안을 마련하는 데 기여할 수 있을 것

프로젝트명	재난안전 지식베이스 구축 요소기술 및 검색 UI 개발
테이블 코드	doc_detail_event_subject_info

구분	필드명	필드명	필드명	비고
3	doc_date	PK	N	문서가 작성된 날짜로, 사건의 시간적 맥락을 제공합니다.
4	subject_idx	PK	N	주어의 시작 위치를 나타내며, 텍스트 내에서 주어를 정확히 추출하는 데 사용됩니다.
5	position	PK	N	주어가 문장에서 나타나는 전체 범위(시작과 끝 위치)를 나타냅니다.
6	subject_info			주어의 텍스트 값(예: '정부', '소방서', '대중')으로, 주체의 구체적인 정보를 제공합니다.

고유 아이디	문장 고유 아이디	문서 일자	subject 인덱스	위치	subject 정보
DOC001	SENT001	2024-12-01	0	0-2	정부
DOC001	SENT002	2024-12-01	15	15-17	소방서
DOC002	SENT001	2024-12-02	8	8-10	대중
DOC003	SENT003	2024-12-03	3	3-5	주민

그림 2.9 사건정보 DB 테이블

분
대

을 :
로 :

프로젝트명	재난안전 지식베이스 구축 요소기술 및 검색 UI 개발
-------	-------------------------------

1	doc_uid	PK	N	사건 정보를 담은 문서를 구분하기 위한 고유 식별자입니다. 하나의 문서는 여러 문장으로 구성됩니다.
2	sent_uid	PK	N	문서 내 특정 문장을 식별하기 위한 고유 식별자입니다. 문서와 문장의 계층적 관계를 나타냅니다.
3	doc_date	PK	N	문서가 작성된 날짜로, 사건의 시간적 맥락을 제공합니다.
4	spatial_idx	PK	N	장소 정보가 텍스트에서 등장하는 위치를 인덱싱한 값입니다. (예: 문서의 몇 번째 단어에 등장했는지 등)
5	position	PK	N	장소 정보를 문장에서 실제로 언급된 형태를 나타낸 값입니다. (예: "서울시 송파구", "제주도", "뉴욕")
6	spatial_info			장소 정보를 표준화하거나 구조화된 형태로 나타낸 값입니다. (예: "서울", "제주", "New York")

문
있
추
대

고
문
"s:
정
명
"s:
포

테
스어
타니
한 국
으로
는
식별
기
spat
이를
중에
가
서울
NLI

API 명	기능 및 제공항목	가격	데이터 형식
Google News API	Google 검색의 주제, 헤드라인, 인기 기사, URL 및 기타 뉴스 항목을 애플리케이션이나 웹페이지에 통합 가능	Free	JSON
Bloomberg API	플랫폼 데이터에 대한 24시간 프로그래밍 방식 액세스 제공 스트리밍 실시간 뉴스, 지연된 뉴스 및 과거 뉴스, 속보, 과거 데이터, 심층 분석 및 기타 금융 시장 정보 제공	Free / 비상업적	JSON, XML, HTML
News API	키워드나 문구, 언어, 출판 소스 이름, 출판 날짜 및 출판 소스의 도메인 이름을 사용하여 출판된 기사 검색	449\$/Month	GET HTTP / JSON
New York Times API	11개의 공개 API를 제공 월별 출판물 기사, 헤드라인 및 기타 세부 정보 등	Free / Day1000회 제한	Restful / JSON
ESPN API	NBA, 농구, NFL, 축구, MLB 등 다양한 스포츠에 대한 최신 뉴스 보도, 리뷰, 점수 및 기타 하이라이트 제공	Free	XML, JSON
Bing News API	Query를 이용한 최신 뉴스, 흥미로운 기사, 인기 주제 및 이미지에 대한 포괄적인 결과	종량제 가격	JSON
Guardian API	콘텐츠, 사용된 태그, 사용 가능한 섹션, 출판물 데이터베이스에 있는 에디션 등	Free / 비상업적	JSON
Yahoo News API	최신 뉴스 기사와 헤드라인, 포괄적인 영상 보도, 의견, 이미지 제공	Query 1,000개/\$1.80	Restful/ XML, JSON
Financial Times API	비즈니스 및 경제 뉴스, 정치, 논평, 전문가 분석 및 기타 금융 데이터 제공	종량제 가격	JSON

항목	설명
뉴스 검색	•특정 주제나 키워드를 기반으로 최신 뉴스 검색
카테고리 필터링	•스포츠, 경제, 기술 등 다양한 카테고리별 뉴스 필터링
출처 필터링	•특정 뉴스 출처나 출판사를 기반으로 뉴스 필터링
날짜 범위 지정	•특정 날짜 범위 내의 뉴스 기사 검색
언어 필터링	•원하는 언어로 작성된 뉴스 기사만 검색
정렬 옵션	•최신 순, 관련성 순 등으로 검색 결과 정렬
기사 미리보기	•뉴스 기사 제목, 요약, 출처 등을 포함한 기사 미리보기 제공
지역 뉴스 검색	•특정 지역에 대한 뉴스 검색

항목	설명
API 키	• API 접근을 위한 인증 키
검색 쿼리	• 뉴스 기사를 검색하기 위한 키워드 또는 문구
필터 옵션	• 카테고리, 출처, 날짜 범위, 언어 등 다양한 필터링 옵션
정렬 옵션	• 검색 결과를 정렬하기 위한 옵션
응답 형식	• JSON 형식으로 제공되는 검색 결과
기사 메타데이터	• 기사 제목, 요약, URL, 출처, 발행 날짜 등 뉴스 기사와 관련된 메타데이터
쿼터 및 제한	• 하루 또는 분당 호출 수에 대한 쿼터 및 제한
에러 핸들링	• 잘못된 요청, 인증 실패 등 다양한 에러 상황에 대한 응답 처리

항목	설명
APIKey (필수)	<ul style="list-style-type: none"> Your API key. Alternatively you can provide this via the X-API-Key HTTP header.
Category	<ul style="list-style-type: none"> Find sources that display news of this category. Possible options : business entertainment general health science sports technology. Default: all categories
Language	<ul style="list-style-type: none"> Find sources that display news in a specific language. Possible options : ar de en es fr he it nl no pt ru sv ud zh. Default : all languages.
Country	<ul style="list-style-type: none"> Find sources that display news in a specific country. Possible options : ae ar at au be bg br ca ch cn co cu cz de eg fr gb gr hk hu id ie il in it jp kr lt lv ma mx my ng nl no nz ph pl pt ro rs ru sa se sg si sk th tr tw ua us ve za. Default : all countries.

[illegible]

T
 3
 T

각
 의

1. Developer : \$ 0	2. Business : \$ 449	3. Advanced : \$ 1,749	4. Enterprise : Contact
Totally Free	For production and published commercial projects.	Per month, Billed monthly	For enterprise projects that require premium data or solutions.
• Search articles and get live top headlines	• Search all articles and get live top headlines	• Search all articles and get live top headlines	• Access to extra articles and an extended source library
• Articles have a 24 hour delay	• New articles available in real-time	• New articles available in real-time	• Articles are enriched with 20 additional data-points
• Search articles up to a month old	• Search articles up to 5 years old	• Search articles up to 5 years old	• Article & story clustering
• CORS enabled for localhost	• CORS enabled for all origins	• CORS enabled for all origins	• Custom classifications, tagging, & information extraction
• 100 requests per day	• 250,000 requests per month included	• 2,000,000 requests per month included	• Unlimited requests
• No extra requests available	• \$0.0018 per extra request	• \$0.0009 per extra request	• Add sources on demand
• No uptime SLA	• No uptime SLA	• 99.95% uptime SLA	• Custom SLA & dedicated chat
• Basic support	• Email support	• Priority email support	• On-premise deployment

J
 2
 3
 3
 4
 3
 3
 4
 3
 3
 3

개
 세
 기
 석
 기
 은
 },
 },
 1.
 기
 을



변환대상 목록				
구분	데이터명	원본 용량	텍스트 용량	이미지 용량
감사연보	2013 ~ 2022 감사연보	829MB	289MB	추출완료
국정감사결과보고서	2008 ~ 2022년 행정안전위원회 국정감사결과보고서	15MB	107MB	추출완료

데이터명	페이지 수	데이터 형식
2013감사연보	924	pdf
2014감사연보	748	pdf
2015감사연보	818	pdf
2016감사연보	850	pdf
2017감사연보	834	pdf
2018감사연보	980	pdf
2019감사연보	971	pdf
2020감사연보	901	pdf
2021감사연보	916	pdf
2022감사연보	708	pdf
2008년행정안전위원회국정감사결과보고서	203	pdf,hwp
2009년행정안전위원회국정감사결과보고서	219	pdf,hwp
2010년행정안전위원회국정감사결과보고서	250	pdf,hwp
2011년행정안전위원회국정감사결과보고서	259	pdf,hwp
2012년행정안전위원회국정감사결과보고서	298	pdf,hwp
2013년행정안전위원회국정감사결과보고서	301	hwp
2014년행정안전위원회국정감사결과보고서	317	pdf,hwp
2015년행정안전위원회국정감사결과보고서	331	pdf,hwp
2016년행정안전위원회국정감사결과보고서	162	pdf,hwp
2017년행정안전위원회국정감사결과보고서	158	pdf,hwp
2018년행정안전위원회국정감사결과보고서	246	pdf,hwp
2019년행정안전위원회국정감사결과보고서	260	pdf,hwp
2020년행정안전위원회국정감사결과보고서	334	pdf,hwp
2021년행정안전위원회국정감사결과보고서	264	pdf,hwp
2022년행정안전위원회국정감사결과보고서	202	pdf,hwp

교정 전	교정 후
재난안전 대책 국정감사결과보고서 제5페이지 재난안전 관련 대응체계가 중요합니다. 추가적인 조치가 필요합니다.	재난안전 대책 재난안전 관련 대응체계가 중요합니다. 추가적인 조치가 필요합니다.

구분	데이터명	원본 용량
감사연보	2013 ~ 2022 감사연보	829MB
국정감사결과보고서	2008 ~ 2022년 행정안전위원회국정감사결과보고서	15MB

프로젝트명		(재난연) 재난안전 지식베이스 구축 요소기술 및 검색 UI 개발					
데이터 코드		pdf_meta_info				스키마명	DW
데이터 설명		PDF메타정보				작성일	2024.07.07
No.	필드 ID	Key	Null	필드명	Type	길이	비고
1	title			title	character varying	255	
2	producer			producer	character varying	255	
3	author			author	character varying	255	
4	pages			pages	integer	-	
5	encrypted			encrypted	character varying	20	
6	pdf_file_version			pdf_file_version	character varying	20	
7	page_layout			page_layout	character varying	255	
8	page_mode			page_mode	character varying	255	
9	pdf_id			pdf_id	character varying	255	
10	fonts_unembedded			fonts_unembedded	text	-	
11	fonts_embedded			fonts_embedded	text	-	
12	attachments			attachments	text	-	
13	images			images	character varying	255	
14	file_path			file_path	character varying	255	
15	file_name			file_name	character varying	255	
16	pdf_mod_file_name			PDF Modify File name	character varying	255	
17	file_permissions			file_permissions	character varying	9	
18	file_size			file_size	numeric	-	
19	creation_date			creation_date	timestamp without time zone	-	
20	modification_date			modification_date	timestamp without time zone	-	
21	access_date			access_date	timestamp without time zone	-	
22	load_dt			적재 일시	timestamp without time zone	-	
데이터 설명							

테이블명	테이블 설명
Disaster_Table (재난 기본 정보를 저장하는 테이블)	disaster_id: 재난 고유 식별자 disaster_type: 재난 유형 (예: 태풍, 화재, 지진) location: 피해 지역 date: 재난 발생 날짜 damage_scale: 피해 규모 (인명 피해, 재산 피해 등) description: 재난에 대한 설명
Damage_Report_Table (각 재난의 피해 이력 및 상세 정보를 저장하는 테이블)	report_id: 보고서 고유 식별자 disaster_id: 재난과 연결된 식별자 (외래 키) human_damage: 인명 피해 수 property_damage: 재산 피해 금액 economic_loss: 경제적 손실 recovery_status: 복구 상태

