

김기남, 송경준, 유태권, 최보현











Contents

01

주제

03 서비스 구현 02 Flow Chart

04 한계점 및 보완점



Detailed Topic

부문	주요 내용
서비스	국회 개방데이터(국회 Open API 등) 등 국회 정보를 기반으로 생성형 Al를 활용한 모바일 서비스 등 개발







Detailed Topic

RAG & Langchain을 활용한 사회 이슈 기반 국회 정보 서비스



Detailed Topic

RAG & Langchain을 활용한 사회 이슈 기반 국회 정보 서비스

뉴스

뉴스를 통해서 국회에 관심을 갖게 하자!

Motivation

- ☑ 높은 진입장벽
 - 정치 분야, 국회 활동과 관련된 분야는 진입 장벽이 높음
 - 관련 뉴스를 보기 위해서는 배경 지식이 필요함
- ☑ 떨어지는 접근성, 인지도
 - 국회정보포털은 다양한 정보를 제공하지만, 접근성과 인지도가 떨어짐
- ☑ 쉽게 전달
 - 쉽게 접근할 수 있는 뉴스 라는 소재를 활용하여 국회 정보를 쉽게 전달함







02 Flow Chart



Flow Chart





Clustering & Keyword Extraction





NER & Summary











User Interface



03 서비스 구현

키워드 추출 NER (개체명인식) 벡터 유사도 검색 RAG

CRAWLING

- 1. "정치" vs "법"
- 2. 국회 = "입법부" -> **"법"에 집중**
- 3. "법" 키워드로 기사 수집(하루에 약 500개 기사) -> 과도한 크롤링 시 유의미하지 않은 기사 다수 추출

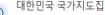
(歸) 한국경제 PiCK ⋅ 6시간 전 ⋅ 네이버뉴스

한동훈 "이재명, 법카로 샴푸 사고 초밥 사먹어도 공천할 건가" 한 위원장은 이 대표 회견에서 '법인카드 유용 의혹'에 대한 질문은 없었 냐고 기자들에게 물은 뒤 "제가 한번 물어보고 싶다. 첫째, 법카 본인이 쓴 거 맞나. 둘째, 만약 민주당 어떤 예비후보가 기업이든 국가든 법카를 자...



한동훈 "이재명, 테러로 정치장사...법카로 샴푸 산 후... 문화일보 PICK · 5시간 전 · 네이버뉴스 李 직격한 한동훈 "법카로 초밥 사먹은 후보 공천... 디지털타임스 PICK · 7시간 전 · 네이버뉴스

국회는 주권자인 국민에 의해 선출된 국회의원으로 구성되며, 국민을 대표해 국가 운영의 근간인 법률을 제정하고, 예산을 심의 · 확정하며 중요한정책 결정을 하는 대한민국의 입법 기관이다.



http://nationalatlas.ngii.go.kr > pages > page_1881

국회 - 대한민국 국가지도집

PREPROCESSING

- Konlpy의 **Okt**(Open Korean Text)를 활용하여 **명사**만 추출

당시 A씨는 빌려준 사룟값 등을 갚으라고 지인과 말다툼을 하던 중 화가 나 범행한 것으로 알려졌다 -> ['당시', '빌려', '준', '사룟값', '지인', '말다툼', '화가', '나', '범행']

- **불용어** 처리
 - 한 달 동안의 기사에서 가장 많이 등장한 단어 100개 중 유의미하다고 판단한 단어 제외

```
| Ii_stopwords = ['것', '등', '수', '고', '말', '그', '씨', '위', '대한', '대해', '위해', '관련', '며', '중', '경우', '를', '및', '때', '지난', '통해', '개', '때문', '더', '또', '건', '우리', '생각', '지금', '최근', '라며', '거', '비', '사실', '이번', '날', '포토', '속보', '종합', '法']
```

PREPROCESSING

- TfidfVectorizer을 통한 벡터화
 - **TF-IDF**: 해당 문장에서는 많이 등장하지만, 전체 문서에서는 적게 사용될수록 분별력 있는 특징이라는 점을 반영
 - TF : 특정 단어가 하나의 데이터 안에서 등장하는 횟수
 - DF: 특정 단어가 여러 데이터에 자주 등장하는지를 알려주는 지표
 - IDF: DF의 역수

$$TF(t, d) = \frac{E H d M H 단 H t 가 등장한 횟수}{E H d M 등장한 모든 단어의 수}$$

$$TF-IDF(t,d,D) = TF(t,d) * IDF(t,D)$$

CLUSTERING - DBSCAN

- DBSCAN(밀도 기반 클러스터링)
 - 특정 point 기준으로 반경 eps 내에 점이 min_samples개 이상 있으면 하나의 군집으로 인식
 - eps를 낮게(0.1), min_samples를 높게(1) 설정 -> 중복 기사 클러스터링
 - 군집 별 대표 기사 추출



구글 · 애플은 플랫폼법 논의 거부…"이러다 해외 플랫폼만 반사이익"

공정위, '가맹 갑질' 맘스터치에 과징금 3억 원 공정위, 점주 협의회 만들자 '가맹 갑질' 맘스터치 과징금 3억원

"점주 혐의회 인정 못 해"…'가맹 갑질' 맘스터치 과칭금 3억 원 공정위, '가맹 갑질' 맘스터치 제재…과징금 3억

"점주 협의회 인정 못 해"…'가맹 갑질' 맘스터치 과징금 3억원 점주협의회 구성했다고 계약해지… 맘스터치 3억 과징금 잘 나가던 맘스터치. '가맹 갑질'로 과장금 3억원 부과

cluster num : 109

개포·목동부터 반월산단까지…'노후계획도시' 108개로 확대

cluster num : 70

[포토] 중대재해처벌법 유예 불발 규탄 대회

[포토] 중대재해처벌법 유예 촉구 위해 국회 찾은 중소기업인들

cluster num : 71

의료진 위장 이스라엘 군인들, 병원서 하마스 3명 사살…국제법 위반 논란

cluster num : 72

미성년자 끌여들여 '집단 성매매'…40대 남성 검찰 송치 미성년자 등 집단 성매매 알선한 40대 검찰 송치 미성년까지 끌여들여 집단 성행위 알선하고 촬영한 男

미성년자까지 끌여들여 집단 성매매 알선한 40대 검찰 송치



CLUSTERING - KMEANS

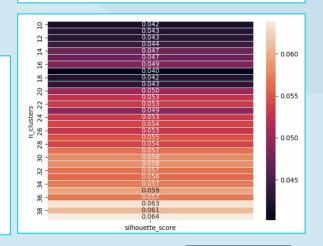
K-MEANS

- 각 클러스터와 거리 사이의 분산을 최소화하는 방식으로 동작
- Silhouette Score을 활용하여 최적의 k 찾기
- 군집 내 원소의 거리가 일정 수준 이상인 군집 제거

	cluster_label	distance	length					
12	12	1.296274	89					
13	13	1.268325	132					
14	14	0.931016	14					
15	15	1.023628	25					
16	16	1.241879	57					
17	17	1.168335	55					
18	18	1.338548	68					
19	19	1.372345	150					

```
cluster 19.0 title
        2월1일부터 전세 사기 피해지 지원 '원스톱' 이뤄진다
      출연연 공공기관 해제에 연구계 환영…"연구자율성 확대 계기"
     오늘부터 전세대출도 갈아타기 시행…'금리 인하 경쟁' 본격화?
    경기도, 올해 60억 6000만원 투입해 '농촌 일손부족 대응
     중기 옴부즈만, 지난해 기업 현장 규제애로 3천600여건 발굴
2322
        올해 전력시장 변화는…산업부, 분산에너지법
2353
       항소심도 택배노조 손들자… 산업계
2428
      원청이 하청직원 사장이라는 法…단체협약 체결 의무는 뺐지만
2447
    자본시장법 개정안 통과…금융당국 "불법리당방 엄격 규제"(종합)
         불법 리딩방에 철퇴…자본시장법 개정안, 본회의 통과
Name: title, Length: 183, dtype: object
```

$$s(i) = \frac{b(i) \cdot a(i)}{\max\{a(i), b(i)\}}, -1 \le s(i) \le 1$$



OPENALAPI

Text sample: '''{story}'''

PROMPTING

prompt = f""" Determine main one keyword that are being discussed in the following text. Please present keyword by short word phrases. Please present the answer only in Korean.

클러스터별 대표 키워드 추출 유사한 키워드 병합

	keyword	cluster	score
0	중대재해법	9.0	91
1	이태원 특별법	35.0	74
2	이태원 특별법	16.0	72
3	플랫폼법	4.0	69
4	중대재해처벌법	2.0	69
5	플랫폼법	10.0	62
6	이태원 특별법	17.0	61
7	중대재해법	36.0	58
8	노후정비	8.0	55
9	'고발사주'	23.0	52

text = ', '.join(list(keyword_df.keyword)) prompt = f"""text에 나열되어 있는 키워드들을 비슷한 것끼리 묶어줘. text는 다음과 같아 : {text}"""

response = get_completion(prompt)

"- 중대재해법, 중대재해처벌법, 중처법 협상₩n- 이태원 특별법, 이태원법, 이태원특별 법, 이태원특별법 거부권₩n- 플랫폼법₩n- 노후정비₩n- '고발사주', 민주당 돈봉투 의혹 ₩n- 선거법 위반₩n- 비대면 진료₩n- 달빛철도 특별법₩n- 검단아파트 주차장 붕괴₩n- 고 준위 특별법\n- 구속영장 청구\n- 한미 고위급 경제협의회\n- 독도\n- 신발 폭행\n- 보 험사기방지법"

KEYWORD EXTRACTION

- 이번 주 "법" 관련 주요 키워드 10가지와, 키워드에 해당하는 기사 데이터 구축

	main_keyword	total_score
0	이태원 특별법	325
1	중대재해법	269
2	플랫폼 <mark>법</mark>	131
3	선거법 위반	73
4	노후정비	55
5	비대면 진료	35
6	달빛철도 특별법	28
7	검단아파트 주차장 붕괴	25
8	고준위 특별법	24
9	구속영장 청구	22

keyword	content	date	title	uri	
노후정비	서울 양천구 목동 일대 아파트 모습.[연합뉴스] 가양(서울)·수지(용인)·행신(고	2024- 01-31	노후계획도시 특별법, 108곳으로 확대 가양·수지·정자 등 포함	https://n.news.naver.com/mnews/article/025/000	0
중대재해법	중소기업업계 3500명이 중대재해처벌법 유예 법안 즉시 통과를 촉구했다. 오는 2	2024- 01-31	전국 중기인 3500명 "중처법 유예안 통과" 호소	https://n.news.naver.com/mnews/article/014/000	1
플랫폼법	한국지역정보화학회가 31일 서울 중구 한국프레 스센터에서 개최한 '온라인 플랫폼 규	2024- 01-31	행정규제 전문가들 "플랫폼법, 옥상옥 규제 될 수 있어"…부작용 우려	https://n.news.naver.com/mnews/article/030/000	2
검단아파트 주치 장 봉고	(인천=뉴스1) 김민지 기자 = 원회룡 국토교통부 장관이 28일 오후 인천시 서구	2024- 01-31	서울시, '검단아파트 사고' GS건설 1개월 영업정지 처분	https://n.news.naver.com/mnews/article/008/000	3
노후정비	노후도시 특별법 적용 대상 108곳·215만가구로 확대 서울 남산에서 바라본 아파	2024- 01-31	서울가양·용인수지도 노후도시 특별법 대 상…"재건축 등 정비사업 속도낼 것"	https://n.news.naver.com/mnews/article/421/000	4
달빛철도 특별법	광주와 대구 잇는 총연장 198.9km 철도, 남부 경 제권 이루는 기회 될듯 ▲	2024- 01-25	영호남 화합 염원 담은 '달빛철도 특별법' 드디어 국회 통과	https://n.news.naver.com/mnews/article/047/000	1060
달빛철도 특별법	25일 국회 본회의에서 달빛철도 건설을 위한 특 별법안 가결이 선포되고 있다. (사	2024- 01-25	광주~대구 '달빚철도' 특별법 국회 본회의 통과예타 없이 사업 추진	https://n.news.naver.com/mnews/article/437/000	1061
중대재해법	유예 연장 불발에 오는 27일 50인 미만 사업장 확대 적용중대재해 발생시 사업주	2024- 01-25	[Q&A] 중대재해법 확대되면 동네빵집 사 장님 뭐가 달라지나	https://n.news.naver.com/mnews/article/001/001	1062
달빛철도 특별법	광주와 대구를 잇는 '달빛고속철도' 건설을 위한 특별법이 오늘(25일) 국회 본회	2024- 01-25	광주·대구 잇는 '달빚고속철도 특별법', 국 회 본회의서 가결	https://n.news.naver.com/mnews/article/056/001	1063
중대재해법	"중소업체 대표들, 법 적용 대상에 해당하는지조 차 알지 못해" 이정식 고용노동부	2024- 01-25	이정식 고용장관 "중대재해처벌법 개정 좌 절돼 안타깝다"	https://n.news.naver.com/mnews/article/366/000	1064

NER(개체명인식)



✓ 개체명 (Named Entity)

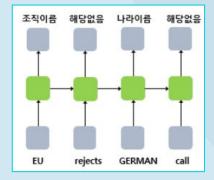
인명, 기관명, 지명 등과 같이 문서에서 특정한 의미를 가지고 있는 단어 or 어구

BIO 태깅 방식

B: Begin O: Outside

I: Inside

개체명인식을 수행할 때 코퍼스로부터 개체명을 인식하기 위한 방법



부 B-movie

산 I-movie

I-movie

보 0

0

C B-theater

G I-theater

V I-theater

가 **O**

자 0

뉴스 기사에서 국회의원 추출

'PER' tag를 인식하는 NER 모델 개발

1. 데이터셋

NAVER NLP Challenge 2018 NER 데이터셋

- -> 14 종류의 Annotated Entities (인물(PER) 포함) 한국해양대학교 NER 데이터셋
- -> 10 종류의 Annotated Entities (인물(PER) 포함)

국회의원 목록 내의 **이름**과 **데이터셋 내 PER**과 대체



뉴스에서 이름을 더 잘 인식하도록 학습

2. 국회법률정보시스템의 21대 국회의원 데이터 수집



뉴스 기사에서 국회의원 추출

'PER' tag를 인식하는 NER 모델 개발

모델 성능

• beomi/KcELECTRA-base

	precision	recall	f1-score	support
B-PER 0 I-PER	0.98 1.00 0.98	1.00 1.00 1.00	0.99 1.00 0.99	11529 2762775 22880
accuracy macro avg weighted avg	0.99 1.00	1.00	1.00 0.99 1.00	2797184 2797184 2797184

PER 외의 나머지 태그는 모두 'O'로 바꿔준뒤 학습하였음

결과



'LAW' tag를 인식하는 NER 모델 개발

1. NAVER NLP Challenge 2018 NER 데이터셋에서 법 데이터를 추출

filtered_law_words

['지방자치법,',
'미디어법',
'미도스특검법',
'유통산업발전법',
'의무신고제법',
'방송법의',
'국가공무원법은',
'국가계약법',

2. 국회법률정보시스템의 현행 법률 데이터 수집

번호	법령명 ^	분야	법령종류
825	산림조합법	제2장 산 림	법률
824	수소경제 육성 및 수소 안전관리에 관한 법률	제2장 에너지이용	법률
823	병역법	제2장 병 무	변률
322	대체역의 편입 및 복무 등에 관한 법률	제1절 군인사	법률
321	군인사법	제1관 통 칙	법률
320	선거관리위원회법	제1장 선거관리위원회	법률
319	공공단체등 위탁선거에 관한 법률	제1장 선거관리위원회	변률
318	도시 및 주거환경정비법	제3절 도시개발	튭
317	한국환경공단법	제1장 행정조직·통칙	법률
816	한강수계 상수원수질개선 및 주민지원 등에 관한 법률	제2장 환경보전	법률

'LAW' tag를 인식하는 NER 모델 개발

- 3. NAVER NLP Challenge 2018 NER 데이터셋 법 데이터를 랜덤으로 '국회법률정보시스템의 현행 법률 데이터'로 바꾼 후 'LAW' tag 생성
- 4. 나머지 태그는 전부 'O'로 통일
 - for i, tag in enumerate(unique_tags):
 print(tag) # 학습을 위한 label list

 O
 B-LAW
 I-LAW

'LAW' tag를 인식하는 NER 모델 개발

5. KcELECTRA 사전학습 모델을 fine-tuning하여 '법'을 인식하는 모델 구현

text3 = ''<mark>선거법</mark> 위반을 제외한 나머지, 즉 공무상비밀누설, 개인정보보호법 위반, 형사사법절차전자화촉진법<mark>위반 혐의를 가지고 집행유예 없는 실형을 선고한 것은 꽤 중한 처분이다.</mark>

laws3 = ner_inference(text3)

laws3

- ['선',
- '##'H'
- '##법',
- 'ንዘ',
- '##인',
- '##정',
- '##早'
- ' ### U
- '##= '
- ##豆,
- '##법',
- '렴',
- '##人上'。
- · ##64 ·
- ·##20',
- '##절',
- '#** / /
- '##전',
- '##|||'
- ##화,
- ##季',
- '##진',
- '##법']

'LAW' tag를 인식하는 NER 모델 개발

6. 저장된 모델을 불러와서 새로운 문장 속 법률을 추출하는 함수 생성

text3 = ''<mark>선거법</mark> 위반을 제외한 나머지, 즉 공무상비밀누설, <mark>개인정보보호법</mark> 위반, 형사사법절차전자화촉진법<mark>위반 혐의를 가지고</mark> 집행유예 없는 질형을 선고한 것은 꽤 중한 처분이다. 게다가 손 검사장은 전과도 없고 현직 검사장 신분이다.

print(extract_law(text3))

['선거법', '개인정보보호법', '형사사법절차전자화촉진법']

추출된 법률과 유사한 현행 법률 제시

TF-IDF 기반의 벡터 유사도 검색

- 1. TfidfVectorizer를 이용하여 현행 법률 텍스트 데이터를 벡터 형태로 변환
- 2. 뉴스에서 추출된 법 텍스트를 벡터 형태로 변환
- 3. 두 벡터 간 코사인 유사도 계산
- 4. 현행 법률 중 유사도가 0.5 이상인 법률 출력

idices = extract tfidf('형사사법절차전자화촉진법', tfidf matrix)

추출된 법률: 형사사법절차전자화촉진법

유사도 0.5 이상의 기존 법률:

형사사법절차 전자화 촉진법 (유사도: 1.0000)

idices = extract_tfidf('줌대재해법', tfidf_matrix)

추출된 법률: 중대재해법

유사도 0.5 이상의 기존 법률:

중대재해 처벌 등에 관한 법률 (유사도: 0.7728) 중대재해 처벌 등에 관한 법률 시행령 (유사도: 0.7523)

문제 발생

idices = extract tfidf('개인정보보호법', tfidf_matrix)

추출된 법률: 개인정보보호법

유사도 0.5 이상의 기존 법률:



SBERT 시도

추출된 법률과 유사한 현행 법률 제시

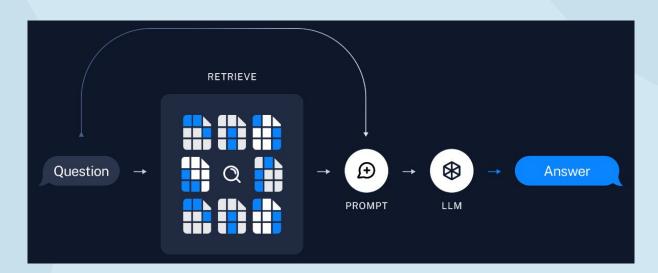
SBERT(Sentence-BERT) 기반의 벡터 유사도 검색

- 1. 사전훈련된 SBERT 모델을 통해 현행 법률 텍스트 데이터에 대한 문장 임베딩 벡터 생성
- 2. 뉴스에서 추출된 법 텍스트를 벡터 형태로 변환
- 3. 두 벡터 간 코사인 유사도 계산
- 4. 유사도가 높은 순으로 정렬하여 제시



RAG

- 1. 지식저장소의 연관성 검색을 이용해, **질문과 가장 가까운 문서**를 LLM의 Context로 활용
- 2. 정보를 DB에 저장해두고, 사용자의 질의가 들어올 때 관련된**정보를 검색**하여 해당 정보가 담긴 문서들을 프롬프트를 통해 LLM에 전달하는 방식



RAG

Documents Chunked Texts Embeddings

- 국회의원 관련 문서들 로드 및 분할
- Embedding Model: jhgan/ko-sroberta-multitask
 - KorNLU 데이터셋으로 한국어 사전학습 모델 파인튜닝한 ko-sentence-transformers
 - HuggingFace에 있는 한국어 임베딩 모델들 중 다운로드 수 및 좋아요 수 1순위

RAG



- Vector DB 및 Retriever: FAISS (Facebook AI Similarity Search)
 - Facebook Al에서 개발한 유사도 검색 모델
 - 대규모 데이터베이스에서 벡터를 효율적으로 검색하는 것을 목표로 함
 - ㅇ 대규모 데이터 세트에서 빠른 유사성 검색이 필요한 애플리케이션에 유용
- LLM: ChatOpenAl gpt-3.5-turbo
 - OpenAI의 채팅 전용 LLM
 - o temperature = 0.1 (창의성)

RAG 구현 예시









RAG

문서에 있는 내용에 대해 물으면, 정확하게 답변

Ask me anything!

강기윤 국회의원 정보 요약해줘

Response: content='강기윤 국회의원은 1960년 6월 4일에 태어났으며, 경남 창원시 성산구에서 선거구를 가지고 있습 니다. 그는 보건복지위원회에 소속되어 있으며, 현재는 대구경북신공항 여야정 협의체 위원, 국민의힘 천안함 생존 장병과 유가족 지원을 위한 TF 위원, 한중차세대리더포럼 회원, 국회 아시아정당국제회의 의원연맹 회원 등의 활동을 하고 있습 니다. 또한, 강기윤 국회의원은 미래통합당 경남도당 민생특위 위원장이며, 제19대 국회의원으로 활동하였고, 새누리당 원내부대표도 역임한 적이 있습니다.

RAG 구현 예시









RAG

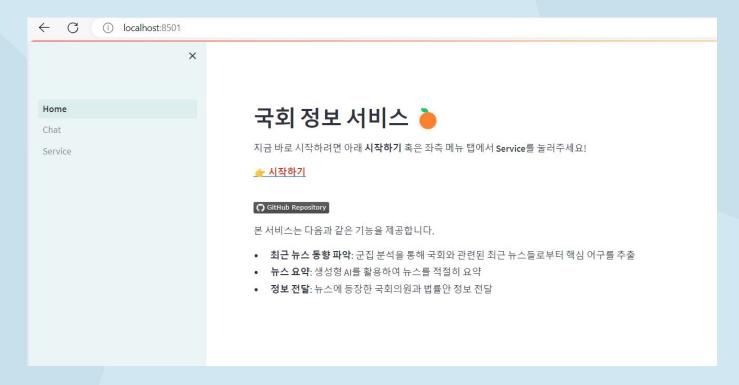
문서에 없는 내용에 대해 물으면, 해당 정보가 없다고 답변

Ask me anything!

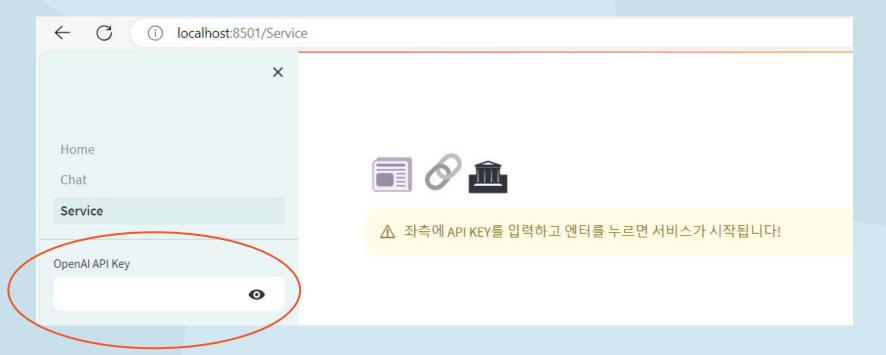
김김김 국회의원 정보 알려줘

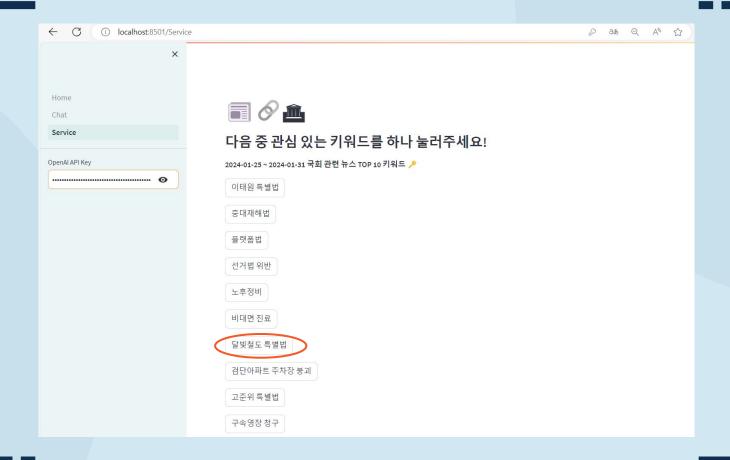
Response: content='김김김 국회의원에 대한 정보는 제공되지 않습니다.'

Streamlit 구현 예시



Streamlit 구현 예시





비대면진료

달빛철도 특별법

■ 다음은 키워드와 관련된 뉴스의 원문입니다.

광주광역시 서구의회 의원 일동은 31일 열린 제318회 임시회 제2차 본회의 직후 '달빛철도특별법'(이하 '특별법')의 국회 본회의 통과를 환영하며 달빛철도 건설의 조속한 추진을 촉구했다. 서구의회는 "달빛철도의 건설은 영호남 주민의 30년 숙원사업이다. 국가 국토 균형 발전과 동서화합을 이끌고 영호남 남부경제권이 새로운 도약을 할 수 있는 중추적 기반 마련이 될 것이다"며 "국가균형발전을 위한 중요한 사업이 수도권 중심의 단순한 경제 논리와 정치적 이해관계로 진행되서는 안된다"고 주장했다.건의안을 대표발의한 백종한 의원은 "해방이후 끊임없이 정치적으로 이용되어 온 극단적인 영호남 지역갈등의 해소를 위해서라도 달빛철도는 하

▲ 다음은 뉴스 요약 결과입니다.

31일 제318회 임시회 제2차 본회의에서 광주광역시 서구 의회 의원들은 '달빛철도특별법'의 국회 통과를 환영하고 달빛철도 건설을 촉구했습니다. 이들은 영호남 지역의 오랜 소망으로 국가 균형 발전과 남부경제권 발전을 이끌 것이라고 주장했으며, '달빛철도특별법'의 국회 통과는 영호남 지역의 상생과 화합을 위한 기폭제가 될 것이라고 강조했습니다.

W

다음은 인식된 국회의원과 법률안(유사 법률안)입니다.

교 국회의원 : **백종한**

법률안: 달빛철도특별법(철도안전법, 철도사업법)

검단아파트 주차장 붕괴



04 한계점 & 보완점

한계점 & 보완점

- 키워드 추출 자동화 한계
- ChatOpenAl 외 hugging face에서 제공하는 모델들 중 QA 성능이 좋은 다양한 LLM 시도
- Streamlit의 다양한 기능들을 활용하여 **챗봇 UI**로 업데이트 진행 중

End Of Document

