# 문서요약및

## 쉬운 말 변환기

발표 일시 : 2025.05.20(화)

발표자 : 김효진

## Why?

1. 목적

LLM서비스의 기본형을 구현해보자

2. 방향

일반적으로 문서 요약의 목적이 '학습' 또는 '정보 습득'인 경우가 많으므로, 뉴스를 참조하여 요약 결과를 한번 더 정리해보자.

단, 가능한 쉽게 요약하여 새로운 정보에 흥미를 잃지 않도록 한다.

- 3. 핵심 기술
  - LangChain: 문서 로딩, 벡터 검색, 체인 구성
  - GNews API: 뉴스 데이터 수집
  - FAISS: 뉴스 임베딩 벡터 저장소
  - Streamlit: 사용자 인터페이스

### **PSEUDO LAB 10TH ACADEMY**

## 구현 개요

	1
1. 뉴스 수집	GNews API로 최신 뉴스 기사 수집
2. 뉴스 임베딩	기사 내용을 Langchain으로 벡터화, FAISS에 저장
3. PDF 업로드	요약하고 싶은 PDF 업로드
4. PDF 텍스트 추출	텍스트 추출 후 분할
5. 연관 뉴스 참조	PDF 내용과 유사한 뉴스를 검색
6. LLM 요약	참조된 뉴스와 PDF를 바탕으로 "쉬운말 요약" 생성
7. Streamlit 출력	요약 결과를 웹에서 출력

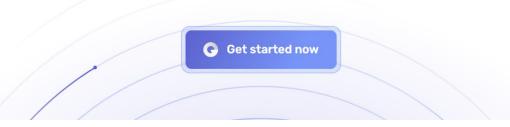
#### **GNes API**

- 기사 제목, 설명, 날짜, URL 등 수천 개의 실시간 뉴스 제공
- 무료 플랜도 꽤 쓸만함:
  - 하루 100개 이상 뉴스 호출 가능
  - 한국어 포함 20개 이상 언어 지원
  - 언론사/국가/날짜/키워드로 필터 가능
- 링크 : <a href="https://gnews.io">https://gnews.io</a>

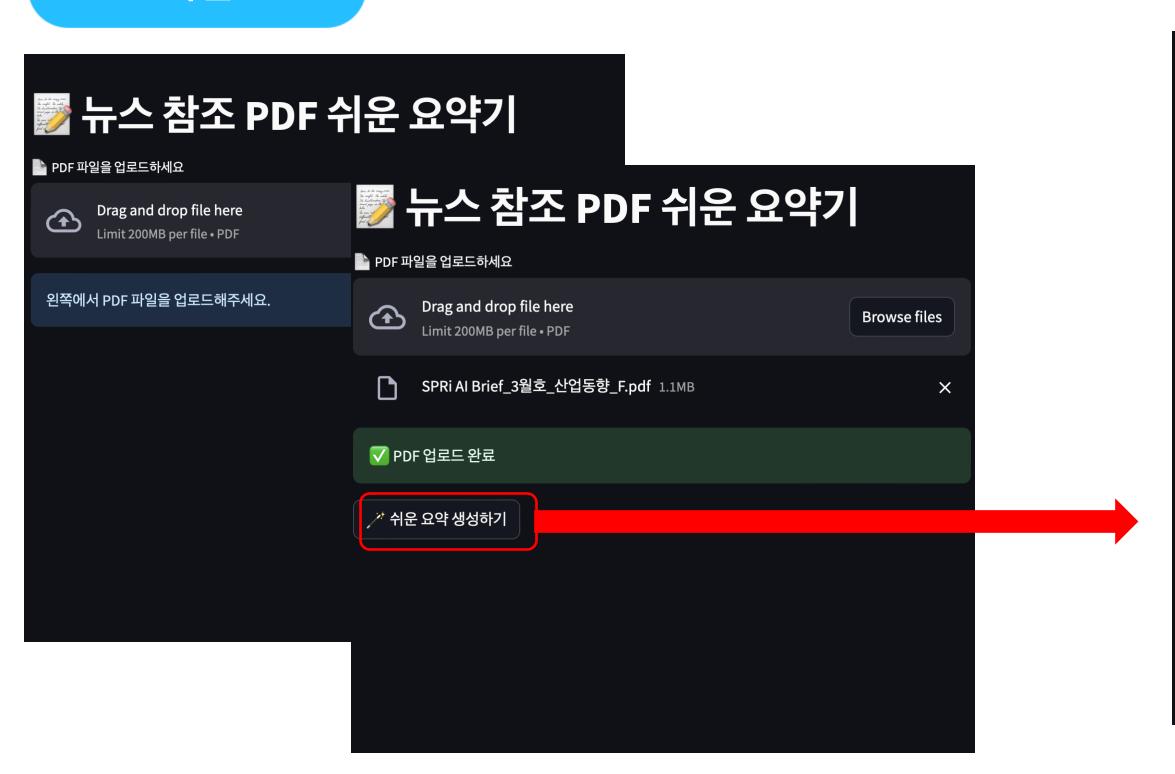


#### Your News API to Search for the Latest and Historical Worldwide News

A powerful REST News API to search for current and historical articles and get trending news in over 22 languages across 30 countries from 60,000+ sources.



### 시연



#### ■ 쉬운 요약 결과

■ 문서 개요: 이 문서는 다양한 데이터와 정보 검색 기술에 대해 설명하고 있어요. 각각의 기술이 무엇인지, 어떻게 작동하는지, 그리고 어떤 상황에서 사용되는지를 간단히 설명하고 있어요. 문서는 여러 가지 주제로 나뉘어져 있는데, 키워드 검색, 페이지 랭크, 데이터 마이닝, TF-IDF, 딥러닝, 스키마, 그리고 데이터프레임에 대해 다루고 있어요.

#### 🧠 핵심 내용:

- 1. **키워드 검색**: 우리가 인터넷에서 무언가를 찾을 때 사용하는 방법이에요. 예를 들어, "커피숍 서울"을 검색하면 서울에 있는 커피숍 목록이 나와요.
- 2. **페이지 랭크**: 웹 페이지가 얼마나 중요한지를 평가하는 방법이에요. 구글 같은 검색 엔진이 어떤 페이지를 먼저 보여줄지 결정할 때 사용해요.
- 3. **데이터 마이닝**: 많은 데이터 속에서 유용한 정보를 찾아내는 기술이에요. 예를 들어, 가게에서 어떤 물건이 잘 팔리는지 알아내는 데 사용돼요.
- 4. **TF-IDF**: 문서에서 어떤 단어가 중요한지를 평가하는 방법이에요. 자주 나오지 않는 단어일수록 더 중요하게 여겨요.
- 5. 딥러닝: 컴퓨터가 사람처럼 생각하고 배우도록 돕는 기술이에요. 사진이나 목소리를 인식하는 데 많이 사용돼요.
- 6. **스키마**: 데이터가 어떻게 저장되고 정리되는지를 설명하는 청사진 같은 거예요. 데이터베이스에서 많이 사용돼요.
- 7. 데이터프레임: 데이터를 표처럼 정리해서 쉽게 볼 수 있게 해주는 도구예요.
- ♥ 흥미로운 점: 딥러닝은 컴퓨터가 사람처럼 생각하도록 도와주는 기술인데, 이 기술 덕분에 컴퓨터가 사진 속 사람이나 사물을 알아볼 수 있게 되었어요. 예를 들어, 스마트폰 카메라가 사람 얼굴을 인식해서 자동으로 초점을 맞추는 것도 딥러닝 덕분이에요!

#### **PSEUDO LAB 10TH ACADEMY**

### 아쉬운 점

- 1. 생각보다 프롬프트 작성이 더 어렵다 어떻게 프롬프트를 바꿔도 완전하게 원하는 내용이 나오지 않는다
- 조금만 구조가 복잡해져도 문제 해결이 너무 어려웠다.
  원래는 문서 로더를 여러 개를 사용하고 싶었으나, 결국 실패...

