KUBIG
DATA SCIENCE & AI

WikiRAG:

유사 문서 기반 단어 설명 생성기

Team | TMI절단기
20기 이예지 21기 이영서, 이예일

KUBIG
DATA SCIENCE & AI

CONTENTS

KUBIG
DATA SCIENCE & AI

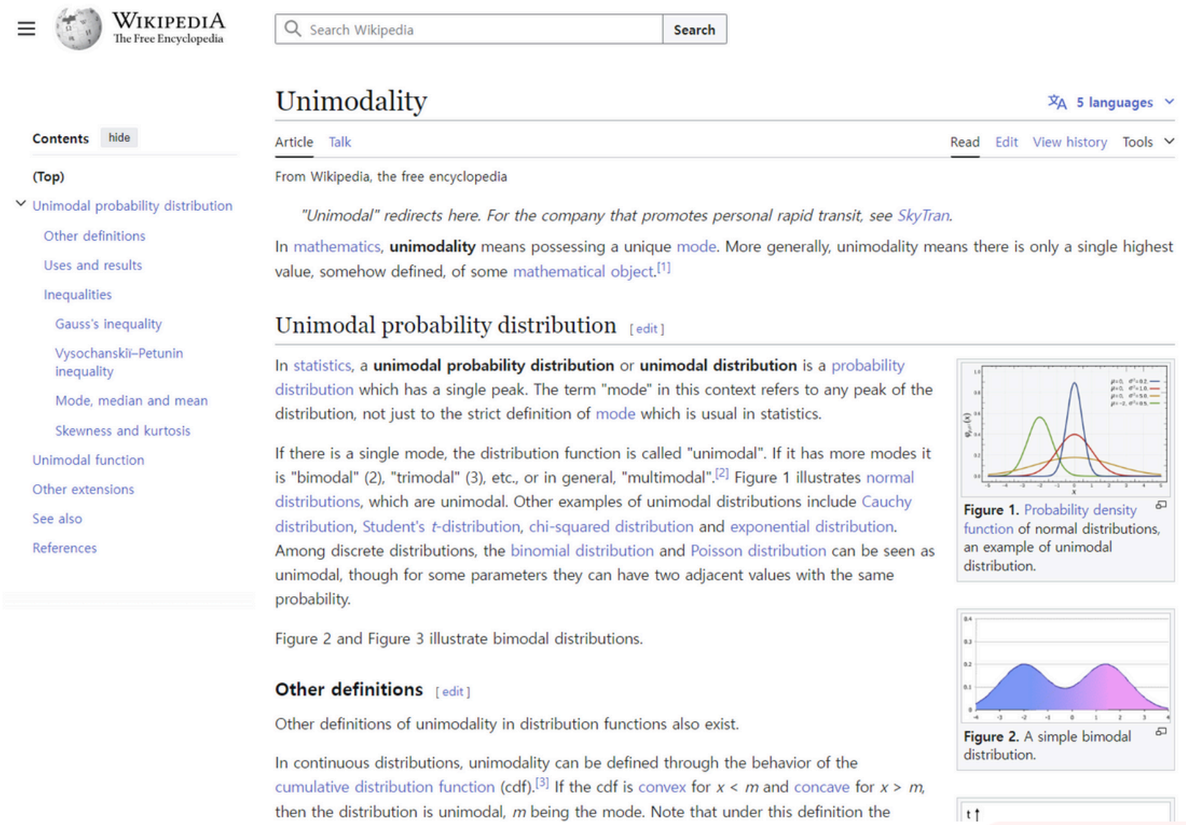
01. 개요

01. 프로젝트 소개

2. Related Work

2.1. Uni-modal for Cancer Survival Prognosis

Genome-based Methods. Genomic data quantifies the molecular profiles of patients from the microcosmic perspective in an explicit manner. Xu et al. [40] utilized the support vector machine-based recursive feature elimination (SVM-RFE) approach to discover the critical gene signatures and conduct prognosis prediction for breast cancer. Chaudhary et al. [2] specifically investigated the deep learning-based model to differentiate survival subpopulations of liver cancer patients in six cohorts. This is the first deep-learning study to identify multi-omics features linked to the differential survival of patients with hepatocellular carcinoma. Ching et al. [7] developed the Cox-nnet to predict patient prognosis from transcriptomics data and found functional biological insights. Wang et al. [35] and Xing et al. [39] employed the graph neural network to model the genomic data for cancer survival prediction based on sample similarity and co-expression gene matrix, respectively.



WIKIPEDIA
The Free Encyclopedia

Search Wikipedia | Search

Unimodality

5 languages

Article | Talk | Read | Edit | View history | Tools

Contents hide

(Top)
Unimodal probability distribution
Other definitions
Uses and results
Inequalities
Gauss's inequality
Vysochanskii–Petunin inequality
Mode, median and mean
Skewness and kurtosis
Unimodal function
Other extensions
See also
References

From Wikipedia, the free encyclopedia

"Unimodal" redirects here. For the company that promotes personal rapid transit, see SkyTran.

In mathematics, **unimodality** means possessing a unique mode. More generally, unimodality means there is only a single highest value, somehow defined, of some mathematical object.[1]

Unimodal probability distribution [edit]

In statistics, a **unimodal probability distribution** or **unimodal distribution** is a probability distribution which has a single peak. The term "mode" in this context refers to any peak of the distribution, not just to the strict definition of mode which is usual in statistics.

If there is a single mode, the distribution function is called "unimodal". If it has more modes it is "bimodal" (2), "trimodal" (3), etc., or in general, "multimodal".[2] Figure 1 illustrates normal distributions, which are unimodal. Other examples of unimodal distributions include Cauchy distribution, Student's *t*-distribution, chi-squared distribution and exponential distribution. Among discrete distributions, the binomial distribution and Poisson distribution can be seen as unimodal, though for some parameters they can have two adjacent values with the same probability.

Figure 2 and Figure 3 illustrate bimodal distributions.

Figure 1. Probability density function of normal distributions, an example of unimodal distribution.

Figure 2. A simple bimodal distribution.

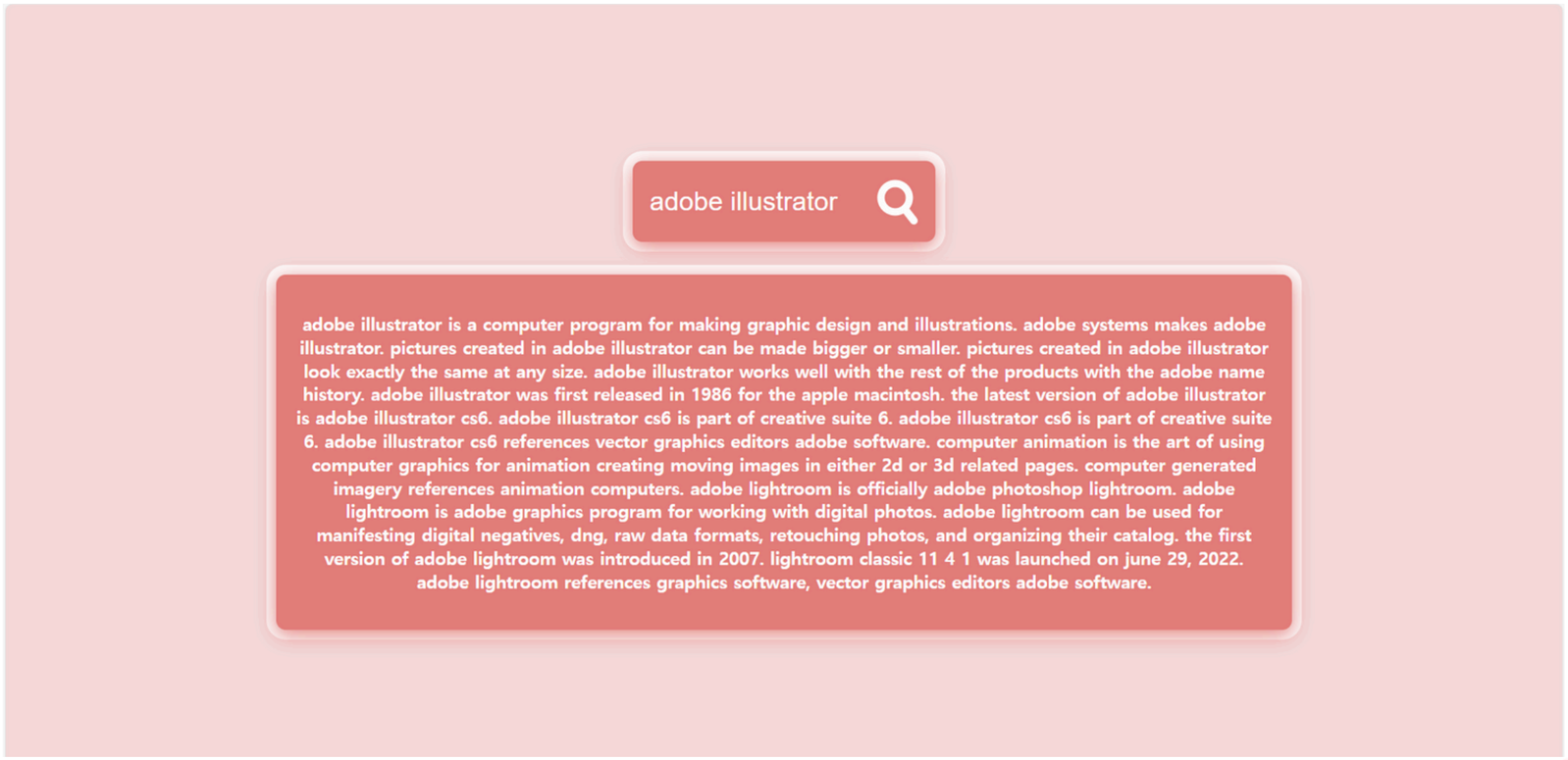
Other definitions [edit]

Other definitions of unimodality in distribution functions also exist.

In continuous distributions, unimodality can be defined through the behavior of the cumulative distribution function (cdf).[3] If the cdf is convex for $x < m$ and concave for $x > m$, then the distribution is unimodal, m being the mode. Note that under this definition the

영문 정보 탐색 과정에서의 피로도를 줄여 콘텐츠에 집중 가능한 환경을 제공하고자 함

01. 프로젝트 소개



위키 문서 기반 단어 설명 생성 예시

01. 개발 흐름도

데이터셋 로드 및 전처리

- Hugging Face의 lsb/simplewiki2023 사용
- 소문자 변환 및 텍스트 전처리
- 너무 긴 텍스트 삭제
- 필요 컬럼 선정

VectorDB 구축

- SemanticChunker 사용
- all-MiniLM-L6-v2 사용
- ChromaDB 사용

설명 생성

- RAG 기반
- 위키기반 T5 모델 사용
- 프롬프트 튜닝

서비스 구현

- Codepen을 활용하여 프론트엔드 구현
- FastAPI를 사용하여 백엔드 구현

KUBIG
DATA SCIENCE & AI

02. 개발 과정

02. 데이터 소개

Dataset Viewer

Auto-converted to Parquet | API | Embed | Data Studio

Split (1)
train · 225k rows

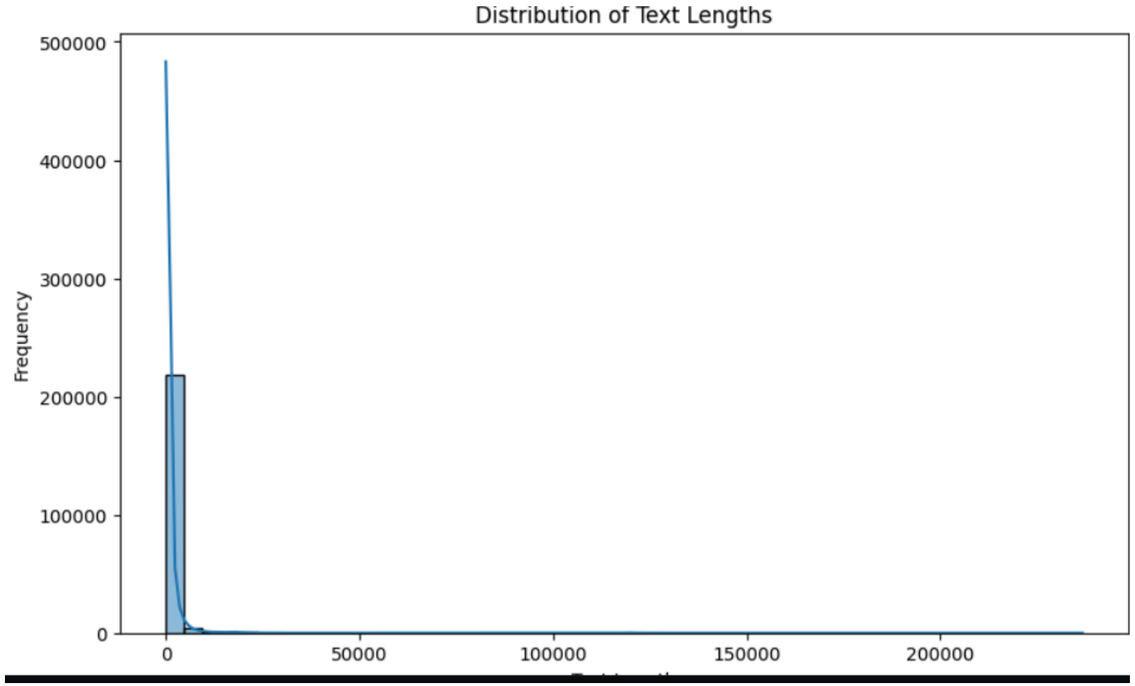
Search this dataset

id string · *lengths* 1 – 6	url string · *lengths* 35 – 214	title string · *lengths* 1 – 118	text string · *lengths* 1 – 237k
1	https://simple.wikipedia.org/wiki/April	April	April is the fourth month of the year in the Julian and Gregorian…
2	https://simple.wikipedia.org/wiki/August	August	August (Aug.) is the eighth month of the year in the Gregorian…
6	https://simple.wikipedia.org/wiki/Art	Art	Art is a creative activity that expresses imaginative or…
8	https://simple.wikipedia.org/wiki/A	A	A or a is the first letter of the English alphabet. The small…
9	https://simple.wikipedia.org/wiki/Air	Air	Air is the Earth's atmosphere. Air is a mixture of many gases…
12	https://simple.wikipedia.org/wiki/Autonomous%20communities%20of%20Spain	Autonomous communities of Spain	Spain is divided in 17 parts called autonomous communities.…

‹ Previous 1 2 3 ... 2,254 Next ›

Hugging Face의 lsb/simplewiki2023

02. 데이터 전처리



count	225332.000000
mean	1125.944877
std	4086.367741
min	1.000000
25%	263.000000
50%	522.000000
75%	987.000000
max	236695.000000

위키 데이터 내 설명 text의 길이가 너무 긴 경우를 제외하고자
Q3 경계로 1000자 이내의 설명 가진 데이터만 추출

02. VectorDB 구축

청킹 및 임베딩 생성

- Semantic Chunker 사용하여 청킹
  - 먼저 문장으로 나누고 의미상 유사한 경우 주변에 있는 항목을 결합하는 방식
  - CharacterTextSplitter, RecursiveCharacterTextSpitter 비교 진행
  - 의미가 상이한 문장들만 효율적으로 청킹하기 위해 Semanic Chunker 채택
- 임베딩 생성
  - title, text 구조로 저장

SemanticTextSplitter	RecursiveCharacterTextSplitter	CharacterTextSplitter
adobe illustrator is a computer program for making graphic design and illustrations it is made by adobe systems pictures created in adobe illustrator can be made bigger or smaller and look exactly the same at any size it works well with the rest of the products with the adobe name history it was first released in 1986 for the apple macintosh the latest version is adobe illustrator cs6 part of creative suite 6 release history references vector graphics editors adobe software	adobe illustrator is a computer program for making graphic design and illustrations it is made by adobe systems pictures created in adobe illustrator can be made bigger or smaller and look exactly the same at any size it works well with the rest of the products with the adobe name history it was	adobe illustrator is a computer program for making graphic design and illustrations it is made by adobe systems pictures created in adobe illustrator can be made bigger or smaller and look exactly the same at any size it works well with the rest of the products with the adobe name history it was
	the same at any size it works well with the rest of the products with the adobe name history it was first released in 1986 for the apple macintosh the latest version is adobe illustrator cs6 part of creative suite 6 release history references vector graphics editors adobe software	the same at any size it works well with the rest of the products with the adobe name history it was first released in 1986 for the apple macintosh the latest version is adobe illustrator cs6 part of creative suite 6 release history references vector graphics editors adobe software

청킹 메소드 비교

02. 유사 문서 검색

검색

- 1차 기준: Title이 정확히 일치하는가
- 2차 기준: Title이 유사한가
- Input 데이터와 Title이 일치하거나 유사한 위키 데이터 3개 선택

```
[adobe illustrator] adobe illustrator is a computer program for making graphic design and illustrations it is made by
****
[computer animation] computer animation or cgi animation is the art of using computer graphics for animation creating
****
[adobe after effects] adobe after effects is a video editing app developed by adobe systems the current version is 22
****
```

adobe illustrator 검색 예시

02. 설명 생성

T5 -large (Text2Text Generation 모델)	propositionizer-wiki-flan-t5-large (위키기반 Text2Text Generation 모델)	pythia-70m-wikipedia-paragraphs (위키기반 Text Generation 모델)
adobe after effects is a video editing app developed by adobe systems Using the following information, provide a detailed explanation of 'adobe illustrator'. 'adobe illustrator'. 'adobe illustrator'. Using the following information, provide a detailed explanation of 'adobe illustrator'. Explain 'adobe illustrator'. a illustrator. Explain	adobe illustrator is a computer program for making graphic design and illustrations. adobe systems makes adobe illustrator. pictures created in adobe illustrator can be made bigger or smaller. pictures created in adobe illustrator look exactly the same at any size. adobe illustrator works well with the rest of the products with the adobe name history. adobe illustrator was first released in 1986 for the apple macintosh. the latest version of adobe illustrator is adobe illustrator cs6. adobe illustrator cs6 is part of creative suite 6. adobe illustrator cs6 is part of creative suite 6. adobe illustrator cs6 references vector graphics editors adobe software. computer animation is the art of using computer graphics for animation creating moving images in either 2d or 3d related pages. computer generated imagery references animation computers. adobe after effects is a video editing app developed by adobe systems. the current version of adobe after effects is 22 6. after effects is an advanced app which lets you create animations, edits, special effects, and more. adobe software references adobe illustrator.	adobe illustrator is the most popular graphic platform available in the world, with the exception of the.info files and.info files of these files. The software also provides a free desktop application and a built-in graphic editor for Windows.

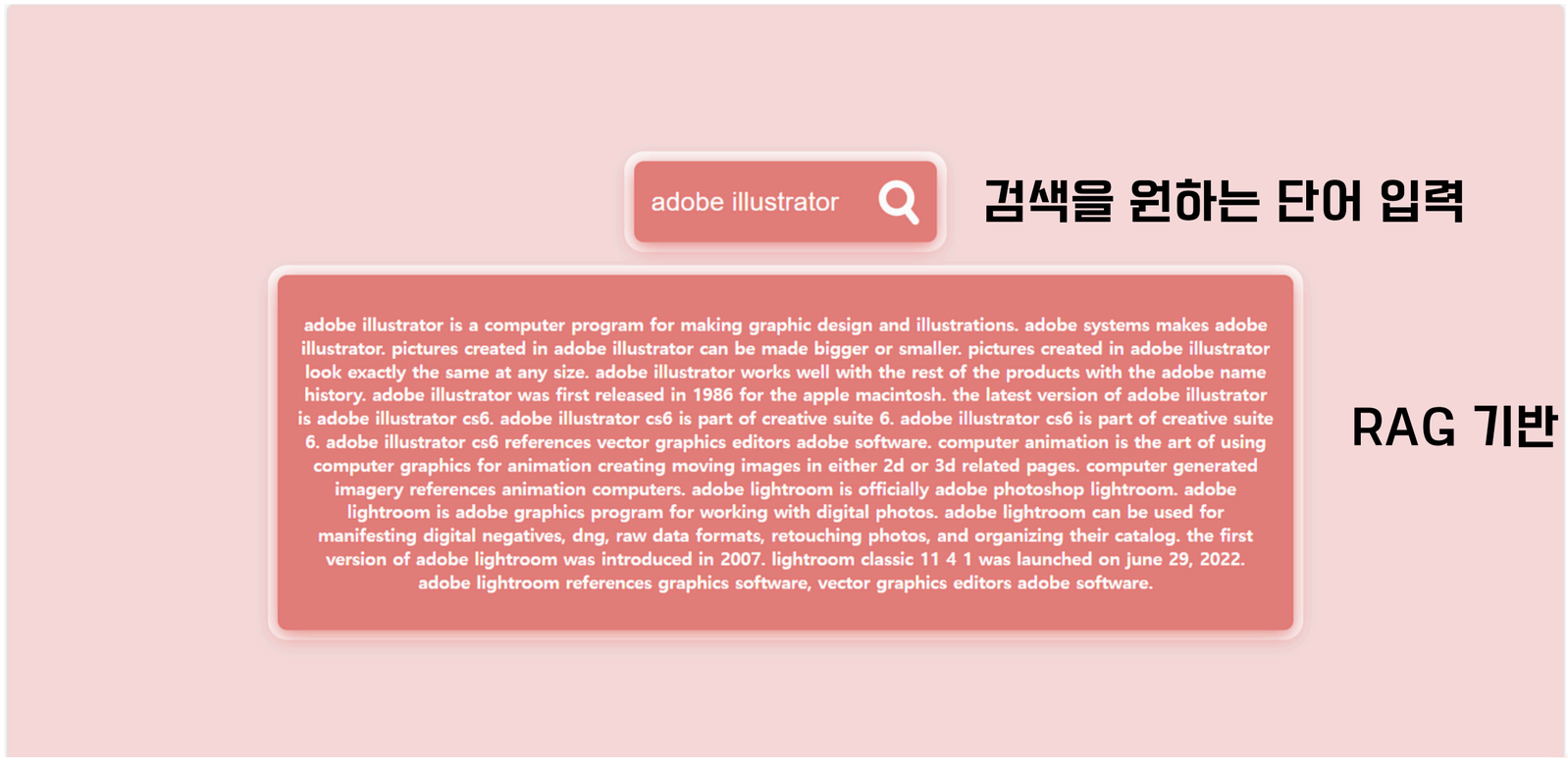
GPT를 이용한 답변 품질 평가 시, propositionizer-wiki-flan-t5-large 모델이 가장 우수한 것으로 나타남.

02. 설명 생성

비교 요소	프롬프트 튜닝 전	프롬프트 튜닝 후
입력 정보	Title + Text	Title + Section + Text
출력 구조	일반적인 설명	"Overview" 중심 설명
설명 스타일	문맥에 따라 달라질 가능성이 큼	개요 위주의 설명을 생성할 가능성이 높음
실험 결과	adobe illustrator is a computer program for making graphic design and illustrations. adobe illustrator is made by Adobe.	adobe illustrator is a computer program for making graphic design and illustrations. adobe illustrator is made by Adobe.

프롬프트 튜닝 전후 결과에 차이가 없음을 확인함.

02. 서비스 구현


adobe illustrator
검색을 원하는 단어 입력
adobe illustrator is a computer program for making graphic design and illustrations. adobe systems makes adobe illustrator. pictures created in adobe illustrator can be made bigger or smaller. pictures created in adobe illustrator look exactly the same at any size. adobe illustrator works well with the rest of the products with the adobe name history. adobe illustrator was first released in 1986 for the apple macintosh. the latest version of adobe illustrator is adobe illustrator cs6. adobe illustrator cs6 is part of creative suite 6. adobe illustrator cs6 is part of creative suite 6. adobe illustrator cs6 references vector graphics editors adobe software. computer animation is the art of using computer graphics for animation creating moving images in either 2d or 3d related pages. computer generated imagery references animation computers. adobe lightroom is officially adobe photoshop lightroom. adobe lightroom is adobe graphics program for working with digital photos. adobe lightroom can be used for manifesting digital negatives, dng, raw data formats, retouching photos, and organizing their catalog. the first version of adobe lightroom was introduced in 2007. lightroom classic 11 4 1 was launched on june 29, 2022. adobe lightroom references graphics software, vector graphics editors adobe software.
RAG 기반 단어 설명 생성

KUBIG
DATA SCIENCE & AI

03. 결론

03. 발전 방안

1. 크롬 확장프로그램으로 개발 시 더 유용하게 사용 가능할 것.
2. 생성된 설명의 품질을 높일 수 있는 추가적인 방안 고려.
    - 위키피디아 외 추가 데이터 활용
    - 설명의 신뢰도 점검
        - 원본 문장과 비교
        - 문서 자체의 신뢰도 점수화
3. 추가적인 질문에 대한 답변도 생성할 수 있도록 기능 개발.
4. 배경지식을 고려한 맞춤형 답변 생성이 가능하도록 기능 개발.

KUBIG
DATA SCIENCE & AI

Thank You