CUAI 프로젝트 NLP 1팀

2024.05.21

발표자 : 권하연

스터디원 소개 및 만남 인증

스터디원 1 : 권하연
스터디원 2 : 송경준
스터디원 3 : 전용현
스터디원 4 : 최형용

목차

Review

성능향상 시도

- 데이터 재구성
- 리트리버 재탐색
- 최적청킹 :: overlap과 chunk_size
- 프롬프트 엔지니어링

향후 계획

Review

“QA 모델링을 통한 CUAI 챗봇 만들기”

선정 배경

: CUAI의 데이터 확인의 어려움

=> 이를 취합하여 정리 후 QA 챗봇을 생성

프로젝트 목표

- **RAG**의 활용으로 LLM의 한계를 극복한다
- QA 모델링을 활용해본다

구현 예시

Q : 2023년 하반기에 NLP 3팀이 한 프로젝트를 알려줘

A : Diffusion 기반 Text Generation

Review

“query에 질문을 넣어, ‘답변’ 과 ‘source documents’ 를 확인”

- source documents: 21,22,23 cuai 데이터

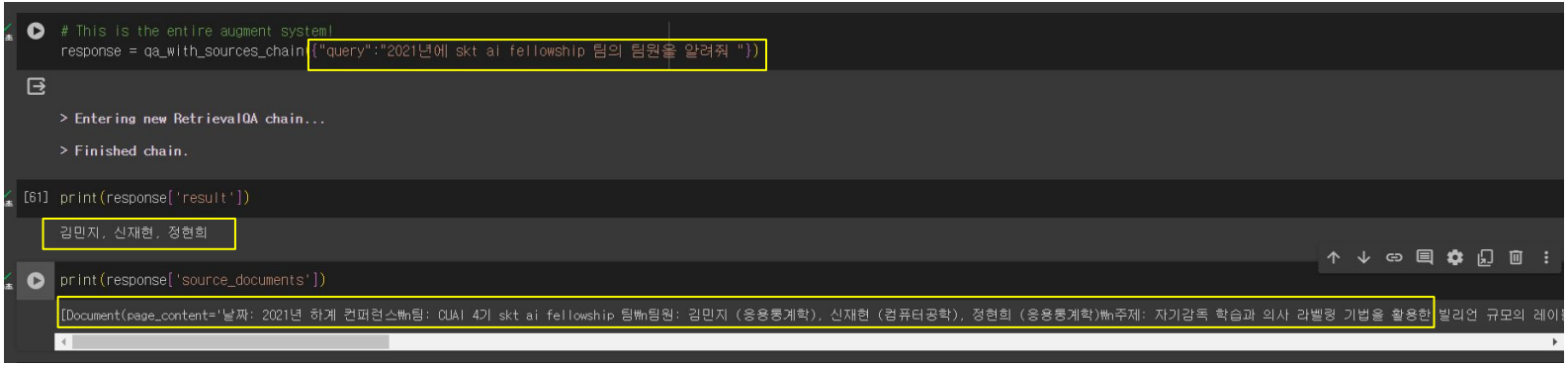
```
import pandas as pd

df_21= pd.read_excel('/content/drive/MyDrive/chatbot project/CUAI_2021.xlsx')
df_22= pd.read_excel('/content/drive/MyDrive/chatbot project/cuai 2022.xlsx')
df_23= pd.read_excel('/content/drive/MyDrive/chatbot project/cuai_2023.xlsx')
df_21.head()
```

	날짜	팀	팀원	주제	요약
0	2021년 하계 컨퍼런스	CUAI 4기 NLP 2팀	유승욱 (중앙대학교 소프트웨어학부), 김상렬 (중앙대학교 컴퓨터공학부), 김중훈 (...	국내 MBTI 커뮤니티의 텍스트 데이터를 활용하여 다양한 머신러닝 및 딥러닝 모델로...	본 연구는 국내 MBTI 커뮤니티의 글을 분석하여 사람들의 MBTI 성격 유형을 예...
1	2021년 하계 컨퍼런스	CUAI 4기 skt ai fellowship 팀	김민지 (응용통계학), 신재현 (컴퓨터공학), 정현희 (응용통계학)	자기감독 학습과 의사 라벨링 기법을 활용한 빌리언 규모의 레이블되지 않은 이미지 ...	이 연구에서는 레이블되지 않은 대규모 이미지 데이터를 활용하여 적은 양의 레이블된 ...
2	2021년 하계 컨퍼런스	CUAI 4기 경영경제 A팀	강영훈 (경영학부), 김민주 (경영학부), 정욱준 (경영학부)	2019년 품목 나라 별 한국 수입액 예측을 위한 앙상블 기법 활용	이 연구에서는 공공 데이터를 기반으로 품목과 국가별 한국 수입액을 예측하는 모델을 ...
3	2021년 하계 컨퍼런스	CUAI 4기 금융 A팀	이재용 (응용통계), 이건이 (응용통계), 서준영 (AI), 김윤진 (소프트웨어)	뉴스 기사 제목을 활용한 주가 변동여부 예측	본 연구는 2021년 1월부터 7월까지의 뉴스 기사 제목을 활용하여 주가 변동 여부...
4	2021년 하계 컨퍼런스	CUAI 4기 금융B팀	윤다인 (소프트웨어학부), 최은서 (소프트웨어학부), 허인 (응용통계학과)	서울시 소비자의 특성에 따른 요식업종 선호도 파악 및 생활지역권 내의 상권 분석을 ...	이 프로젝트는 서울시 내에서 성별과 연령별로 다른 요식업에 대한 선호도 및 상권의 ...

Review

성공 사례


# This is the entire augment system!
response = qa_with_sources_chain({"query":"2021년에 skt ai fellowship 팀의 팀원을 알려줘 "})
> Entering new RetrievalQA chain...
> Finished chain.
[61] print(response['result'])
김민지, 신재현, 정현희
print(response['source_documents'])
[Document(page_content='날짜: 2021년 하계 컨퍼런스₩n팀: CUAI 4기 skt ai fellowship 팀₩n팀원: 김민지 (응용통계학), 신재현 (컴퓨터공학), 정현희 (응용통계학)₩n주제: 자기감독 학습과 의사 라벨링 기법을 활용한 빌리언 규모의 레이

Review

실패 사례

부족한 답변

```
response_2 = qa_with_sources_chain({"query":"2021년 하계 컨퍼런스의 팀들을 모두 알려줘 "})

> Entering new RetrievalQA chain...

> Finished chain.

[83] print(response_2['result'])

 CUAI 4기 문화콘텐츠 B팀, CUAI 4기 컴퓨터비전 1팀

[84] print(response_2['source_documents'])

[Document(page_content='팀 이름: 여름 컨퍼런스 데이터 분석 방법론 연구팀₩n날짜: 2022년 하계₩n팀원 및 발표자: 강연훈(경영학) , 김소은(통계학, 원민재(경영학)', metadata={'source': '/content/drive/MyDrive/chatbot proje
```

잘못된 답변

```
response_3 = qa_with_sources_chain({"query":"2021년 하계 컨퍼런스에서 NLP2팀의 주제가 뭐였어? "})

> Entering new RetrievalQA chain...

> Finished chain.

[92] print(response_3['result'])

 CUAI 4기의 NLP팀은 "한국 드라마 특성을 가진 챗봇 제작"이라는 주제로 활동했습니다.

[93] print(response_3['source_documents'])

[Document(page_content='날짜: 2021-05-04 00:00:00₩n팀: NLP 스터디₩n팀원: 서혜련, 김민주, 권예진, 이하은₩n주제: ML 및 DL의 기본사항, RNN, LSTM, GRU 네트워크와 같은 NLP 기술, 그리고 LSA와 LDA를 활용한 주제 모델링에 대한 심층
```

성능 향상 시도

Data

1. 데이터 수집
2. **데이터 가공**

INDEX

1. 데이터 로드
2. **데이터 청크**
3. 데이터 임베딩
4. 벡터 스토어

Retrieval

1. 유저 쿼리 입력
2. 쿼리 임베딩
3. **벡터 탐색**
4. **관련 문서 리턴**

Augment

1. **초기 프롬프트 생성**
2. **프롬프트 강화**
3. 프롬프트 전송
4. LLM 응답 수신

성능 향상 시도 - 데이터 가공

```
merge_data.head()
```

	날짜	팀	팀원	주제 및 내용	요약
0	2021년 하계 컨퍼런스	CUAI 4기 NLP 2팀	유승욱 (중앙대학교 소프트웨어학부), 김상렬 (중앙대학교 컴퓨터공학부), 김중훈 (...	국내 MBTI 커뮤니티의 텍스트 데이터를 활용하여 다양한 머신러닝 및 딥러닝 모델로...	본 연구는 국내 MBTI 커뮤니티의 글을 분석하여 사람들의 MBTI 성격 유형을 예...
1	2021년 하계 컨퍼런스	CUAI 4기 skt ai fellowship 팀	김민지 (응용통계학), 신재현 (컴퓨터공학), 정현희 (응용통계학)	자기감독 학습과 의사 라벨링 기법을 활용한 빌리언 규모의 레이블되지 않은 이미지 ...	이 연구에서는 레이블되지 않은 대규모 이미지 데이터를 활용하여 적은 양의 레이블된 ...
2	2021년 하계 컨퍼런스	CUAI 4기 경영경제 A팀	강영훈 (경영학부), 김민주 (경영학부), 정욱준 (경영학부)	2019년 품목 나라 별 한국 수입액 예측을 위한 양상블 기법 활용	이 연구에서는 공공 데이터를 기반으로 품목과 국가별 한국 수입액을 예측하는 모델을 ...
3	2021년 하계 컨퍼런스	CUAI 4기 금융 A팀	이재용 (응용통계), 이건이 (응용통계), 서준영 (AI), 김윤진 (소프트웨어)	뉴스 기사 제목을 활용한 주가 변동여부 예측	본 연구는 2021년 1월부터 7월까지의 뉴스 기사 제목을 활용하여 주가 변동 여부...
4	2021년 하계 컨퍼런스	CUAI 4기 금융B팀	윤다인 (소프트웨어학부), 최은서 (소프트웨어학부), 허인 (응용통계학과)	서울시 소비자의 특성에 따른 요식업종 선호도 파악 및 생활지역권 내의 상권 분석을 ...	이 프로젝트는 서울시 내에서 성별과 연령별로 다른 요식업에 대한 선호도 및 상권의 ...

QA From

```
[65] merge_qadata.head()
```

	Q	A
0	2021년에 CUAI 4기 NLP 2팀의 2021년 하계 컨퍼런스 발표에서의 팀원을...	유승욱 (중앙대학교 소프트웨어학부), 김상렬 (중앙대학교 컴퓨터공학부), 김중훈 (...
1	2021년에 CUAI 4기 NLP 2팀의 2021년 하계 컨퍼런스 발표에서의 주제를...	국내 MBTI 커뮤니티의 텍스트 데이터를 활용하여 다양한 머신러닝 및 딥러닝 모델로...
2	2021년에 CUAI 4기 NLP 2팀의 2021년 하계 컨퍼런스 발표를 요약해줘.	본 연구는 국내 MBTI 커뮤니티의 글을 분석하여 사람들의 MBTI 성격 유형을 예...
3	2021년에 CUAI 4기 skt ai fellowship 팀의 2021년 하계 컨...	김민지 (응용통계학), 신재현 (컴퓨터공학), 정현희 (응용통계학)
4	2021년에 CUAI 4기 skt ai fellowship 팀의 2021년 하계 컨...	자기감독 학습과 의사 라벨링 기법을 활용한 빌리언 규모의 레이블되지 않은 이미지 ...

성능 향상 시도 - 데이터 청크

Chunk size : 문서를 작은 조각(chunks)으로 나눌 때 각 조각의 크기
Overlap: Overlap은 각 chunk가 겹치는 부분의 크기

4가지 경우)

1. **청크 크기 크고, 오버랩 크다**: 문맥과 정보 포괄성 유지, 하지만 메모리 사용량과 처리 시간 증가.
2. **청크 크기 크고, 오버랩 작다**: 문맥 포함, 메모리 절약, 하지만 정보 손실 가능성 증가.
3. **청크 크기 작고, 오버랩 크다**: 문맥 연결성 유지, 정보 누락 방지, 하지만 메모리 사용량과 처리 시간 증가.
4. **청크 크기 작고, 오버랩 작다**: 메모리 효율성 높고 처리 속도 빠름, 하지만 문맥과 정보 손실 가능성 높음.

•둘 다 크게 시작해서 작아지는 방향으로 실험 중

성능 향상 시도 - 데이터 청크

Chunker 변경

1. KoNLPy의 형태소 분석기 이용

: 문서에서 형태소를 기준으로 토큰을 설정한 뒤, chunk를 나눔
-> chunk size의 기준을 형태소로 둔 것

2. SemanticChunker 이용

: 텍스트를 의미론적 유사성에 기반하여 분할
open AI Embedding을 이용해 의미론적 기반으로 chunk를 나눔

성능 향상 시도 - 데이터 청크

baseline

```python
from langchain.text_splitter import RecursiveCharacterTextSplitter

text_splitter = RecursiveCharacterTextSplitter(
    chunk_size = 500,
    chunk_overlap = 50,
    length_function = len
)

cuai_data2021_chunks = text_splitter.transform_documents(cuai_data2021)

cuai_data2022_chunks = text_splitter.transform_documents(cuai_data2022)

cuai_data2023_chunks = text_splitter.transform_documents(cuai_data2023)
```

형태소 분석기 기반

```python
# 형태소 분석기 기반 청크 분할기
text_splitter = KoreanMorphemeTextSplitter(chunk_size=500, chunk_overlap=100)

cuai_data2021_chunks = text_splitter.transform_documents(cuai_data2021)

cuai_data2022_chunks = text_splitter.transform_documents(cuai_data2022)

cuai_data2023_chunks = text_splitter.transform_documents(cuai_data2023)
```

SemanticChunker

```python
text_splitter = SemanticChunker(OpenAIEmbeddings(api_key=API_KEY),
                                breakpoint_threshold_type="percentile",
                                breakpoint_threshold_amount=70,
                                )
##청크를 분할후 문서로 변환
cuai_data2021_chunks = text_splitter.split_text(cuai_data2021)
cuai_data2021_chunks = [Document(page_content=chunk) for chunk in cuai_data2021_chunks]
```

결론)
지금까지 소개한 방법을 비교해 나가며, 적절한 chunk 분할 방법을 찾고 있음

성능 향상 시도 - Retrieval

```
#질문하기
questions = ["2021년에 skt ai fellowship팀의 팀원을 알려줘",
             "2022년 하계 컨퍼런스의 팀들을 모두 알려줘",
             "2021년 하계 컨퍼런스에서 NLP 2팀의 주제가 뭐였어?",
             "2023년 하계 컨퍼런스에서 NLP 2팀의 주제가 뭐였어?",
             # 인공지능을 활용하여 회의 녹취록에서 중요한 정보를 추출하고 요약하는 기술을 개발하고자 하였습니다
             "24년 Advanced NLP 1팀의 두번째 발표 주제를 알려줘."]
```

	ensemble	MultiQuery	ParentDocs
A	O	O	O
B	X	X	X
C	O	O	O
D	X	X	X
E	X	X	X

-> Retrieval 의 문제가 아닌듯함

성능 향상 시도 - Retrieval

Chunking?

2021년 하계 컨퍼런스에서 NLP 2팀의 주제가 뭐였어?
CUAI 4기 NLP 2팀은 2021년 하계 컨퍼런스에서 국내 MBTI 커뮤니티의 텍스트 데이터를 활용하여

2023년 하계 컨퍼런스에서 NLP 2팀의 주제가 뭐였어?
2023년 하계 컨퍼런스에서 NLP 2팀의 주제는 "국내 MBTI 커뮤니티의 텍스트 데이터를 활용하여 (

24년 Advanced NLP 1팀의 두번째 발표 주제를 알려줘.
24년 Advanced NLP 1팀의 두번째 발표 주제는 'CUAI 6기 NLP 1팀'으로 한국어 특화된 B

24년 Advanced NLP 1팀의 두번째 발표 주제를 알려줘.
24년 Advanced NLP 1팀의 두번째 발표 주제는 "Stanford NLP group의 Chris Manning 교수의 CS224n 강의를 바탕

성능 향상 시도 - Retrieval

2022년 하계 컨퍼런스의 팀들을 모두 알려줘
2022년 하계 컨퍼런스의 팀들을 알려드리겠습니다.
1. 여름 컨퍼런스 데이터 분석 방법론 연구팀: 강연훈, 김소은, 원민재
2. Machine Learning Travelers 핸즈온 1팀: 김정우, 이하은, 정승욱, 최윤한
3. 여름 컨퍼런스 서울시산업스파이팀: 김지민, 김진재, 정재희, 최서희
4. CUAI 4기 문화콘텐츠 B팀: 안익균, 이효근, 이민규, 음호준
5. 여름 컨퍼런스 CUAI 5기 NLP 15팀: 김민기, 이강민, 오창준
6. Kaggle/Dacon 스터디: 박정현, 김민주, 김준엽, 박정현
7. CUAI NLP 프로젝트 3팀: 곽수민, 양현우, 배현규, 정달민
8. 2022 CUAI PYTORCH 스터디 1팀: 김민기, 김중훈, 이주호

Ensemble

2022년 하계 컨퍼런스의 팀들을 모두 알려줘
2022년 하계 컨퍼런스의 팀들은 다음과 같습니다:
1. 여름 컨퍼런스 데이터 분석 방법론 연구팀
2. Machine Learning Travelers 핸즈온 1팀
3. 여름 컨퍼런스 서울시산업스파이팀
4. CUAI 4기 문화콘텐츠 B팀
5. 여름 컨퍼런스 CUAI 5기 NLP 15팀
6. Kaggle/Dacon 스터디
7. CUAI NLP 프로젝트 3팀
8. CUAI PYTORCH 스터디 1팀

ParentDocs

1. 여름 컨퍼런스 데이터 분석 방법론 연구팀
- 팀원: 강연훈(경영학), 김소은(통계학), 원민재(경영학)
- 주제: Gradient-based One-Side Sampling (GOSS) 및 leaf-wise 성장 전략을 활용한 모델 훈련 효율성 향상
- 내용: GOSS 및 leaf-wise 성장 전략을 통해 모델 훈련 효율성 개선

2. Machine Learning Travelers 핸즈온 1팀
- 팀원: 김정우, 이하은, 정승욱, 최윤한
- 주제: "한눈에 보는 머신러닝" 챕터 학습 및 퀴즈 진행

3. 여름 컨퍼런스 서울시산업스파이팀
- 팀원: 김지민(응용통계학과), 김진재(응용통계학과), 정재희(응용통계학과), 최서희(생명과학과)
- 주제: 서울시 핫플레이스 젠트리피케이션 분석 및 음식점업 생존율 요인 조사

4. CUAI 4기 문화콘텐츠 B팀
- 팀원: 안익균(응용통계), 이효근(소프트웨어), 이민규(소프트웨어), 음호준(물리학)
- 주제: 뉴 노멀 시대의 문화 관광 실태 분석 및 정책 제언

5. 여름 컨퍼런스 CUAI 5기 NLP 15팀
- 팀원: 김민기(경영학부), 이강민(산업보안학과), 오창준(AI학과)
- 주제: AI-hub 감성 대화 말뭉치를 활용한 강의 평가 감성 분석

6. Kaggle/Dacon 스터디
- 팀원: 발표자 - 박정현; 스터디원 - 김민주(경영학과), 김준엽(소프트웨어학과), 박정현(응용통계학과)
- 주제: DACON 경진대회 코드 리뷰 및 진행 계획

7. CUAI NLP 프로젝트 3팀
- 팀원: 곽수민, 양현우, 배현규, 정달민(응용통계학과 소속)
- 주제: Fake News Detection을 위한 LSTM 및 BERT 모델 사용

8. 2022 CUAI PYTORCH 스터디 1팀
- 팀원: 김민기(발표자), 김중훈, 이주호
- 주제: "모두를 위한 딥러닝 시즌 2 - PYTORCH" 학습 및 코드 실습

multiquery

성능 향상 시도 - Prompt Engineering

```
prefix = "이제부터 langchain에 있는 데이터셋에 대한 질문을 해볼거야. 몇 가지 예시를 줄 테니, 아래 예시처럼 답변해줘. ₩
        ₩
        예시1. ₩
        Q: 24년에 푸리에 해석학 스터디 1팀의 첫 번째 발표에서의 내용을 알려줘. ₩
        A: 푸리에 해석학은 함수를 삼각함수들의 선형 결합으로 나타내는 방법을 통해 다양한 신호 처리 문제를 해결하는데 사용됩니다. 이 발표에서는 푸리에 변환의 기본적인 성
        ₩
        예시2. ₩
        Q: 2021년에 CUAI 4기 NLP 2팀의 2021년 하계 컨퍼런스 발표를 요약해줘. ₩
        A: 본 연구는 국내 MBTI 커뮤니티의 글을 분석하여 사람들의 MBTI 성격 유형을 예측하는 것에 초점을 맞추었습니다. 다양한 머신러닝과 딥러닝 기법을 사용해 데이터를 분
        ₩
        예시3. ₩
        Q: 2023년에 CUAI 6기 CV 1팀의 팀원을 알려줘. ₩
        A: 이하윤(소프트웨어학부), 강민기(소프트웨어학부), 박도영(기계공학부) ₩
        ₩
        예시4. ₩
        Q: 2022년에 NLP 논문 구현 3팀 팀의 4번째 발표에서의 주제를 알려줘. ₩
        A: 스터디는 BERT 모델에 대한 심층 분석을 진행했습니다. BERT의 주요 개념인 Masked Language Model (MLM)과 Next Sentence Prediction (NSP)에 대해 설명하고, BERT가
```

성능 향상 시도 - Prompt Engineering

```
response_2 = qa_with_sources_chain({"query":prefix + "Q: 2022년 하계 컨퍼런스의 팀들을 모두 알려줘 "})
```

```
> Entering new RetrievalQA chain...

> Finished chain.
```

```
print(response_2['result'])
```

```
A: CUAI 4기 금융B팀, CUAI 4기 모빌리티 A팀
```

```
response_3 = qa_with_sources_chain({"query": prefix + "Q: 2021년 하계 컨퍼런스에서 NLP2팀의 주제가 뭐였어? "})
```

```
> Entering new RetrievalQA chain...

> Finished chain.
```

```
print(response_3['result'])
```

```
A: 본 연구는 국내 MBTI 커뮤니티의 글을 분석하여 사람들의 MBTI 성격 유형을 예측하는 것에 초점을 맞추었습니다. 다양한 머신러닝과 딥
```

```
response_4 = qa_with_sources_chain({"query":"Q: 2022년 하계 컨퍼런스에 참여한 팀이름을 알려줘"})
print(response_4['result'])
print(response_4['source_documents'])
```

```
> Entering new RetrievalQA chain...

> Finished chain.
A: 2022년 하계 컨퍼런스에 참여한 팀은 CUAI 5기 CV T4 Blueberry팀이었습니다.
```

향후 계획

Data 처리 전/후 데이터 모두 활용

INDEX	Retrieval	Augment
- 청크와 **overlap** 사이즈 조정 - **chunker** 변경	- 리트리버 모델 탐색 - 하이퍼파라미터 조정	- **Few shot promptinig**

병합진행

+ Fine tuning