sLLM을 활용한 투자정보 검색 서비스

2023 삼성증권 디지털/IT 학회 연계 프로그램
디지털기술팀 8조

삼성증권

권도영 김지윤 김나연 백성은 송경준 이동우 조수민

목차

과제 소개

sLLM을 활용한 투자정보 검색 서비스

- 제시된 문서의 내용에 한정해서 질문과 연관된 정보를 추출하고 답변을 생성하는 질의응답 서비스를 제작
- 즉, 주어진 문서를 학습해서 질문에 해당하는 내용을 찾고 답변하는 모델 구현 필요

참고 조건

- API 호출 방식이 아닌 로컬 모델 구현 필수, 모델 fine-tuning 불필요
- 총 5개의 pdf 파일 주어짐 (줄 글, 표, 그래프, 이미지 등으로 이루어짐)
- 질의응답을 통해 결과 확인이 가능한 환경(Dockerfile, docer image 주소) 제출 필요
- OS: Ubuntu / GPU 사용

프로젝트 진행 과정

10/30 ~ 12/7	1주차	2주차	3주차	4주차	5주차	6주차
데이터	전처리 방식 고안 및 조사	데이터 종류별 전처리 방법 리서치 및 실행	데이터 전처리 (줄, 글)	데이터 전처리 (도표, 그래프)	추가 데이터 전처리	데이터 전처리 검토
모델	사전학습 모델 조사 및 평가 모델 고안		사전 학습 모델 추가 리서치	모델 리서치 및 실험 (영어 모델)	모델 리서치 및 실험 (한글 모델)	모델 앙상블 조합, 최종 모델 선정
기타	프로젝트 파이프라인 고안	중간발표	전처리 2팀으로 나눠 진행	전처리 / 모델팀으로 나눠 진행	Docker 이용 방법 학습	발표 PPT 제작 발표 준비 Docker 연동

데이터 전처리

PDF 데이터를 텍스트 / 그래프 / 테이블 3가지 형태로 분리하여 전처리

텍스트 데이터

- PyMuPDF의 fitz를 이용해 특정 페이지 범위 내 텍스트를 추출하는 코드 활용

테이블 데이터

- tabula을 통해 자동화를 시도하였으나, 병합된 셀의 경우 제대로 파싱되지 않는 등 문제 발생
- 이에 PYPDF2의 PdfReader을 통해 테이블 데이터를 텍스트 형태로 파싱
- 이를 데이터프레임으로 만든 후 반복문을 활용해 문장으로 변환

그래프 데이터

- 그래프의 추세 및 대략적인 값과 관련된 정보를 데이터프레임으로 만들어 이를 문장으로 변환

전처리 결과

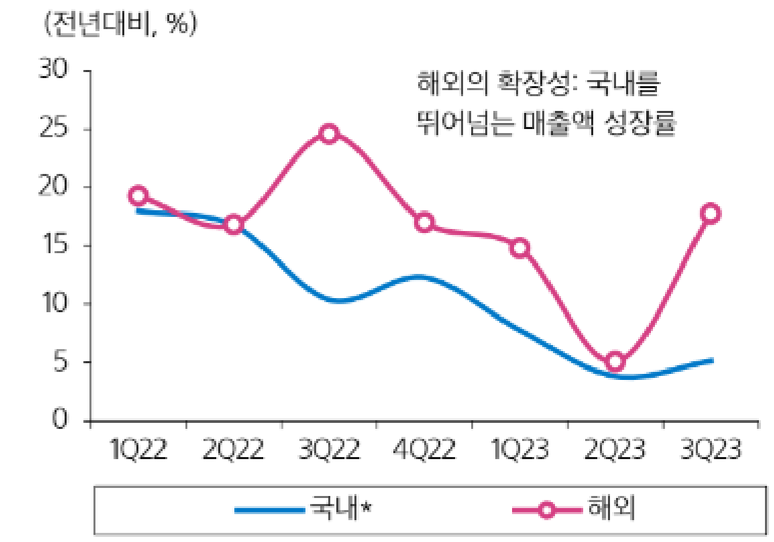
원본 데이터(pdf)

Text

- BUY 의견 제시하며, 목표주가 19만원으로 상향 조정 (12개월 선행 PER 11배 적용).
- 국내만 아니라, 해외에서도 점유율을 확대하며, 원재료 부담을 상쇄하고도 남는 수익성 개선이 예상됨. PCPPI 연결 편입이 더해지며, 24년 영업이익 24% y-y 성장 전망.
- 글로벌 확장성 보유한 동사의 밸류에이션 눈높이는 상향되어야 할 것으로 예상. 한편, 11월 21일 출시 예정인 맥주 신제품은 추가적인 밸류에이션 눈높이 상향의 열쇠.

Graph

롯데칠성 매출액 성장률 비교: 국내* vs 해외



전처리 후 데이터

• BUY 의견 제시하며, 목표주가 19만원으로 상향 조정 (12개월 선행 PER 11배 적용).
• 국내만 아니라, 해외에서도 점유율을 확대하며, 원재료 부담을 상쇄하고도 남는 수익
성 개선이 예상됨. PCPPI 연결 편입이 더해지며, 24년 영업이익 24% y-y 성장 전망.
• 글로벌 확장성 보유한 동사의 밸류에이션 눈높이는 상향되어야 할 것으로 예상.

롯데칠성 매출액 성장률 비교: 국내 vs 해외에서 4Q22의 전년대비 해외 매출액 성장률는 약 20%입니다.
롯데칠성 매출액 성장률 비교: 국내 vs 해외에서 1Q23의 전년대비 해외 매출액 성장률는 약 18%입니다.
롯데칠성 매출액 성장률 비교: 국내 vs 해외에서 2Q23의 전년대비 해외 매출액 성장률는 약 7%입니다.
롯데칠성 매출액 성장률 비교: 국내 vs 해외에서 3Q23의 전년대비 해외 매출액 성장률는 약 19%입니다.

전처리 결과

Table

원본 데이터(pdf)

연결 실적 추이와 전망

(십억원)	1Q22	2Q22	3Q22	4Q22	1Q23	2Q23	3Q23	4Q23E	2021	2022	2023E	2024E
매출액	626	762	784	669	680	796	830	961	2,506	2,842	3,268	4,121
별도법인	584	707	730	621	631	736	769	658	2,345	2,642	2,794	2,877
음료	390	519	537	422	423	538	568	447	1,673	1,868	1,976	2,043
주류	194	188	193	199	208	198	201	211	672	775	818	834
자회사	77	99	96	84	86	102	108	344	272	355	639	1,429
영업이익	60	64	75	24	59	59	84	39	182	223	241	300
별도법인	54	55	71	23				32	175	203	215	228
음료	33	45	64	24							176	186
주류	22	10	7	(1)							39	43
자회사	4	10	6	2							32	78
순이익	37	41	47	6							148	186
이익률 (%)												
영업이익	9.5	8.4	9.6	3.6							7.4	7.3

전처리 후 데이터

롯데칠성의 연결 실적 추이와 전망에서 2024E 순이익(십억원)은(는) 186(십억원)입니다.
롯데칠성의 연결 실적 추이와 전망에서 1Q22 영업이익(%)은(는) 9.5%입니다.
롯데칠성의 연결 실적 추이와 전망에서 2Q22 영업이익(%)은(는) 8.4%입니다.
롯데칠성의 연결 실적 추이와 전망에서 3Q22 영업이익(%)은(는) 9.6%입니다.

사전 학습 모델 선정 eng ver.

HuggingFace의 Question Answering Model 중 높은 성능을 보인 모델을 1인당 7-10개 정도 선정하여 실험

- **deepset/tinyroberta-squad2**
- **deepset/roberta-base-squad2**
- **ntel/dynamic_tinybert**
- **Deepset/tiny_roberta**
- **distilbert-base-cased-distilled-squad**
- **SRDdev**
- **MobileBert**

데이터 전처리 완료된 데이터를 넣고 질의응답을 진행했을 때의 응답을 보고 선정

```
What did Lotte Chilsung pay attention to? Answer: 'domestic and overseas market share expansion', score: 0.9549, start: 90, end: 134
```

ISSUE

영어 사전 학습 모델 이용시 번역과정 필요
질문 > 한영 번역 모델 > qa모델 > 영한 번역 모델 > 답변 출력

번역 모델의 문제로 한글 모델보다 정확성이 떨어질 가능성 존재
-> **한글 사전 모델 이용 결정**

사전 학습 모델 선정 kor ver.

HuggingFace의 Question Answering Model 중 한글 데이터를 기반으로 학습된 모델 10개 선정 및 실험 진행

- timpal0l/mdeberta-v3-base-squad2
- ~~monologg/koelectra-base-v3-finetuned-korquad~~
- arogyaGurkha/koelectra-base-discriminator-finetuned-squad_kor_v1
- ~~eliza-dukim/bert-base-multilingual-cased_korquad-v1~~
- ~~monologg/koelectra-small-v3-finetuned-korquad~~
- ~~arogyaGurkha/kobert-finetuned-squad_kor_v1~~
- Kdogs/klue-finetuned-squad_kor_v1
- ~~ainize/klue-bert-base-mrc~~
- HieuLV3/QA_UIT_xlm_roberta_large
- yjgwak/klue-bert-base-finetuned-squard-kor-v1

- 대부분 KorQuAD 데이터를 학습한 모델

- 데이터 전처리 완료된 데이터(context)를 넣고 질의응답을 진행했을 때의 응답을 보고 선정

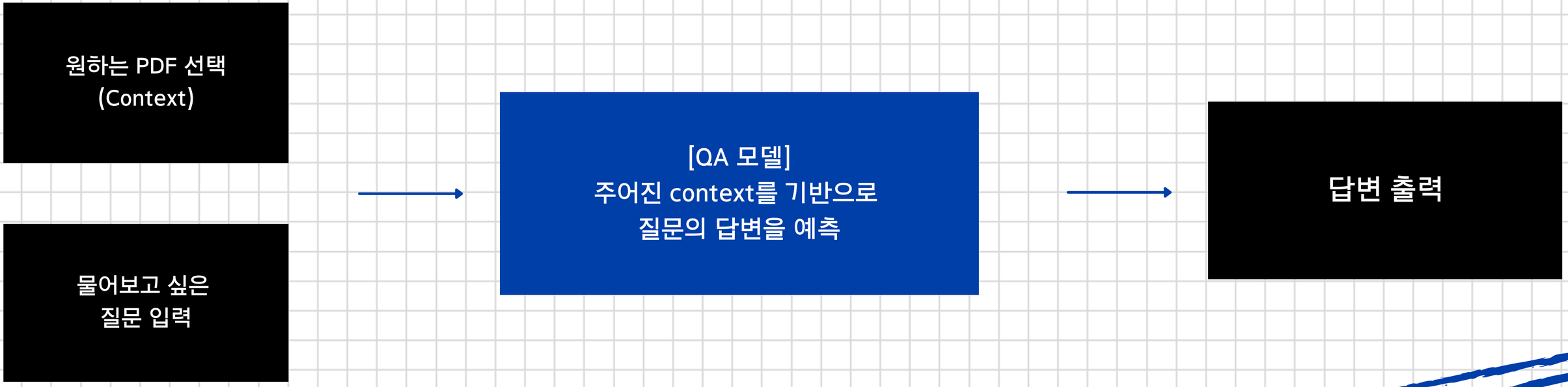
- 약 15개의 질문 - 답변 set을 제작하여 실험
- 모델의 답변이 정답과 다를 경우 > 후보에서 제외
- 올바른 답변이 나오는 경우 > 후보에 포함(5개)

```
S&P100 수익률 상위 종목 2위 종목명은?
Answer (score) : U.S.뱅코프입니다.,
Answer (freq) : U.S.뱅코프입니다.
```

최종 질의응답 모델 파이프라인

Requirement : Python == 3.11 / transformers == 4.35.2 / torch == 2.1.1

transformers의 ‘Question-Answering’ pipeline 사용
+
HuggingFace의 ‘text-generation & Question-answering’ 기반 모델 사용



Q. 기아의 주가 걸림돌이 뭐야?

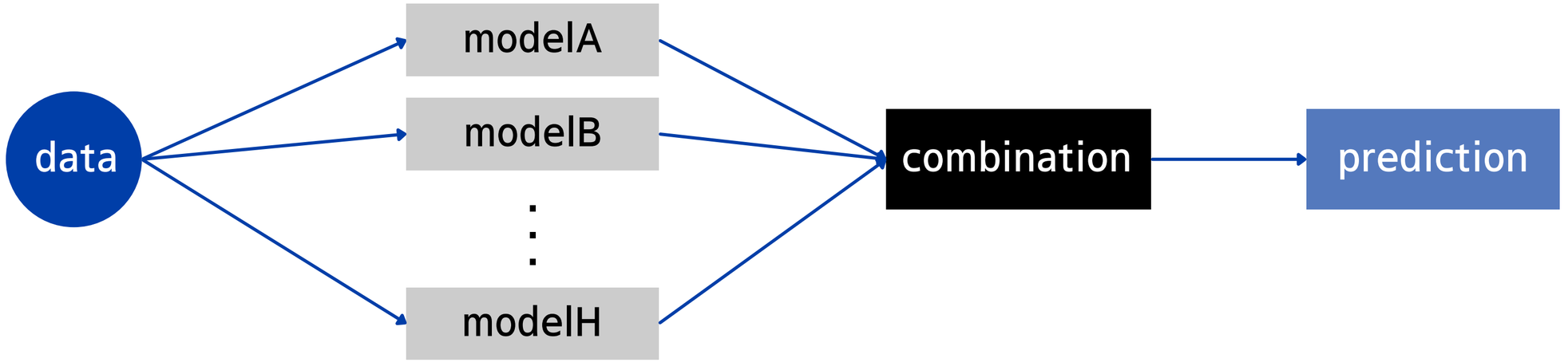
모델A의 Answer: 'None', score: 0.0324, start: 113680, end: 113684	X
모델B의 Answer: '2Q23', score: 0.0084, start: 119672, end: 119676	X
모델C의 Answer: '23.7%', score: 0.8489, start: 105256, end: 105261	X
모델D의 Answer: ' 전기차 판매 둔화가', score: 0.9437, start: 1832, end: 1843	O
모델E의 Answer: '202310', score: 0.1257, start: 5747, end: 5753	X
모델F의 Answer: '1,569', score: 0.0618, start: 28066, end: 28071	X
모델G의 Answer: '전기차 판매 둔화가', score: 0.9996, start: 1833, end: 1843	O

Q. 2018년도에 현대차에서 두 번째로 많이 판매된 차는 뭐야?

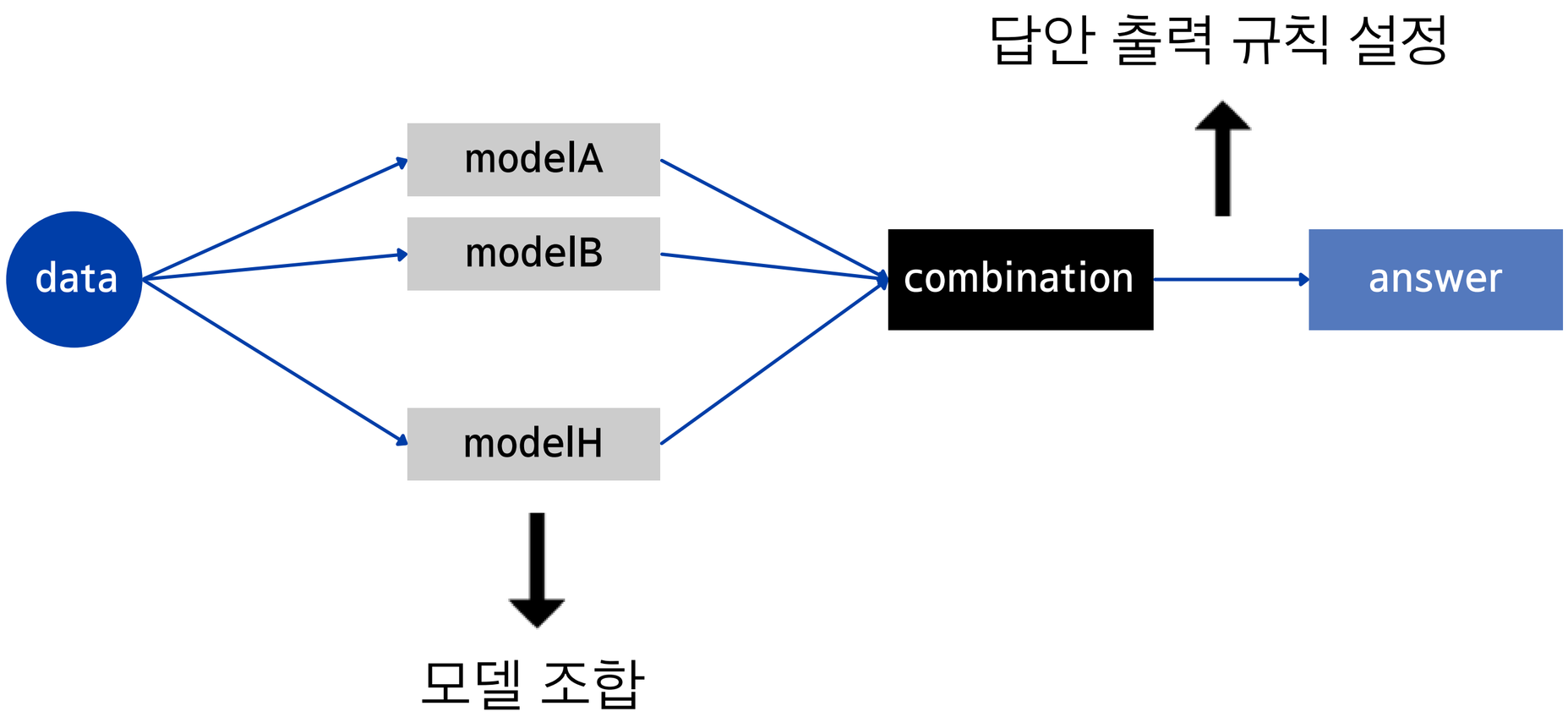
모델A의 Answer: '120,000원', score: 0.1341, start: 113276, end: 113284	X
모델B의 Answer: '산타페입니다', score: 0.0213, start: 120418, end: 120424	O
모델C의 Answer: '펠리세이드입니다', score: 0.6979, start: 121262, end: 121270	X
모델D의 Answer: ' 기아,', score: 0.8371, start: 1739, end: 1743	X
모델E의 Answer: '202310', score: 0.2183, start: 9270, end: 9276	X
모델F의 Answer: '코로나19', score: 0.0198, start: 122281, end: 122286	X
모델G의 Answer: 'EV6의', score: 0.9937, start: 2565, end: 2569	X

앙상블(Ensemble)기법

- 여러 개별 예측 모델을 결합하여 하나의 모델보다 더 나은 예측 성능을 달성하는 방법
- 각 모델이 개별적으로 가진 약점을 서로 보완하고 강점을 강화하는 것이 목표



앙상블 모델 구축 과정


답안 출력 규칙 설정
data
modelA
modelB
modelH
combination
answer
모델 조합

앙상블 모델 구축 과정

1. 모델 조합

1단계: 동일한 질문에 대한 각 모델의 답변을 비교하며 성능이 우수한 모델들을 선택

Q. 뷔 솔로 1집 앨범 이름이 뭐야?

모델A의 Answer: 'None장입니다', score: 0.0448, start: 7753, end: 7761
모델B의 Answer: '%YoY', score: 0.0219, start: 32181, end: 32185
모델C의 Answer: 'Layover', score: 0.5976, start: 6742, end: 6749 (O)
모델D의 Answer: ' Layover', score: 0.9937, start: 6741, end: 6749 (O)
모델E의 Answer: '2023E', score: 0.0495, start: 40402, end: 40407
모델F의 Answer: '257.1', score: 0.0218, start: 30362, end: 30367
모델G의 Answer: 'Layover', score: 0.9995, start: 6742, end: 6749 (O)

앙상블 모델 구축 과정

1. 모델 조합

2단계: 1단계에서 선택한 모델들을 바탕으로 여러 조합의 성능을 테스트해보며 정확도가 가장 높은 조합 선택

Q. 기아의 2022년 국내 판매량은 2021년에 비해 어때?

모델C & 모델D & 모델G 조합
Answer: 감소하였습니다. **(O)**

모델C & 모델D 조합
Answer: K8입니다 **(X)**

앙상블을 통한 성능 향상

2. 답변 추출 규칙 설정

context: 민수는 달콤한 사과와 새콤한 귤을 좋아해
question: 민수가 좋아하는 과일은 뭐야?

1단계: context 길이의 0으로 채워진 리스트 ‘context_score_rank‘생성
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

2단계: 개별 모델로부터 정답 텍스트, 정답의 시작/끝 인덱스, 해당 정답에 대한 점수를 각각 출력
모델A의 출력값: {'score': 0.3, 'start': 8, 'end': 18, 'answer': '사과와 새콤한 귤을'}
모델B의 출력값: {'score': 0.4, 'start': 3, 'end': 18, 'answer': ' 달콤한 사과와 새콤한 귤을'}
모델C의 출력값: {'score': 0.9, 'start': 8, 'end': 18, 'answer': '사과와 새콤한 귤을'}

앙상블을 통한 성능 향상

3단계: 각 개별 모델은 정답의 시작과 끝 사이 범위에 해당하는 context_score_rank의 인덱스 위치에 출력한 점수를 더함

모델A: [0, 0, 0, 0, 0, 0, 0, 0, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0, 0, 0, 0]

+

모델B: [0, 0, 0, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0, 0, 0, 0]

+

모델C: [0, 0, 0, 0, 0, 0, 0, 0, 0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0, 0, 0, 0]

⇓

[0, 0, 0, 0.4, 0.4, 0.4, 0.4, 0.4, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 0, 0, 0, 0]

4단계: context_score_rank 리스트 중 Maximum score 구하기 ➔ 1.7

[0, 0, 0, 0.4, 0.4, 0.4, 0.4, 0.4, **1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7,** 0, 0, 0, 0]

앙상블을 통한 성능 향상

5단계: Maximum score에 임계값(ex. 0.8)을 곱한 값을 구한 후, context_score_rank에서 해당 수보다 큰 값을 갖는 모든 인덱스를 찾아냄

Maximum score = 1.7

Maximum score X 0.8 = 1.36

[0, 0, 0, 0.4, 0.4, 0.4, 0.4, 0.4, **1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7**, 0, 0, 0, 0]

6단계: 인덱스들에 해당하는 context 텍스트를 추출하여 최종 정답으로 출력

[0, 0, 0, 0.4, 0.4, 0.4, 0.4, 0.4, **1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7**, 0, 0, 0, 0]

⇓

context = "민수는 달콤한 **사과와 새콤한 귤을** 좋아해."

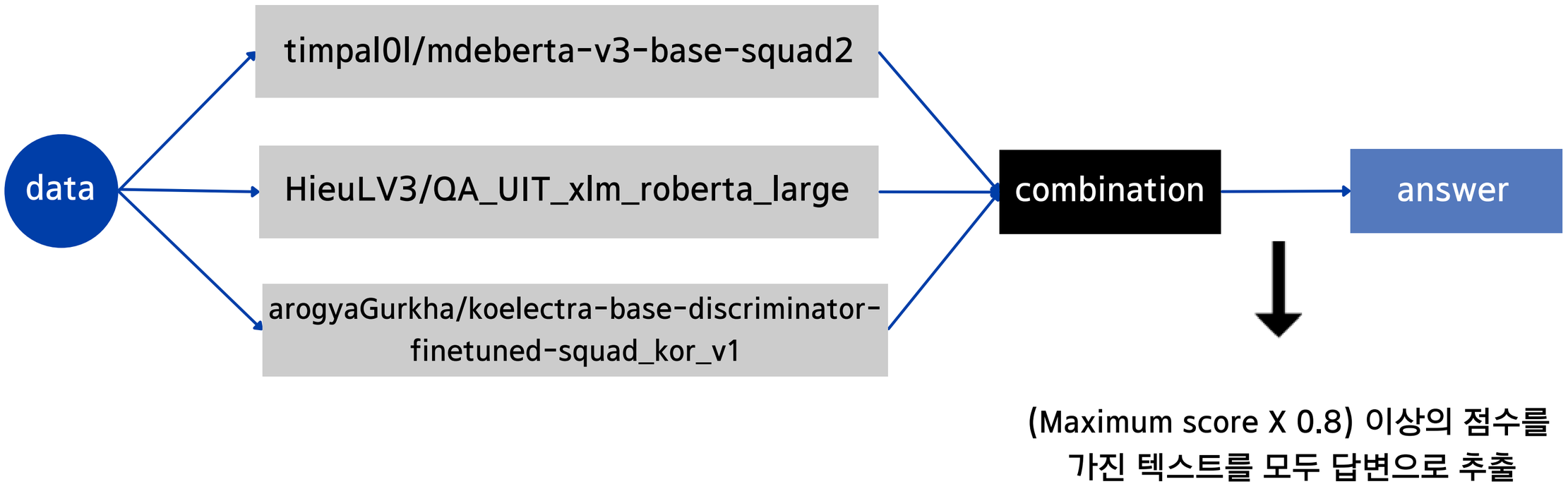
Output

사과와 새콤한 귤을

앙상블을 통한 성능 향상

질문	앙상블 모델	단독 모델
1. 현대차와 기아의 10월 글로벌 도매 판매 기록은 각각 어때?	X	X
2. 자동차 수요에 대한 불안감이 증폭된 이유는 뭐야?	O	O
3. 노사협상은 끝났어? 아니면 진행 중이야?	X	X
4. 기아의 주가 걸림돌은 뭐야?	O	X
5. RV 라인업은 얼마나 판매되었어?	O	O
6. 내수 친환경차는 얼마나 판매됐어?	X	X
7. 현대차 대비 노사협상 길어지면서 어떤 일이 일어났어?	X	O
8. 기아의 2022년 국내 판매량은 2021년에 비해 어때?	O	X
9. 내수 시장 판매에서 세단은 얼마나 판매됐어?	X	X
10. 2018년도에 두 번째로 많이 판매된 차량은 뭐야?	O	X
11. 쏘나타의 2022년 10월 판매실적은 어때?	O	X
	6개 정답	**3개 정답**

완성된 앙상블 모델



(Maximum score X 0.8) 이상의 점수를
가진 텍스트를 모두 답변으로 추출

모델 결과 및 예시

Q. 롯데칠성 주식을 살까 말까?
A. BUY입니다.

Q. 롯데칠성의 현재 주가는 얼마야?
A. 149,300원입니다.

Q. 롯데칠성의 해외법인 영업이익률은 어느정도야?
A. 10.9%로

Q. 롯데칠성의 제로음료 매출액은 2021년도에 얼마야?
A. 4.12조원 8,304억원 89

Q. 롯데칠성의 소주 실적이 어때?
A. 28%

context 기반 추출형 응답에 중점을 둔 앙상블 모델
-> 주어진 맥락에 일치하거나 유사한 단어를 사용해 질문을 하는 경우 정확한 답변 추출

-> 주어진 질문에 대한 답만 추출되는 것이 아닌 조사, 구둣점 등도 함께 출력되는 경우 존재

-> Max Score기반 답안이 모두 join되도록 하여 올바른 답이 포함된 오답이 생성되는 경우 존재

-> 추출형이 아닌 의견을 요구하는 질문을 다루는 경우, 생성형 qa모델로서의 기능은 구현하지 못하고 오답 출력

모델 결과 및 예시

```
docker (com.docker.cli)

lotte
car
qualcom
hive
us

Selected: lotte
Enter your question:
2. 롯데칠성의 현재주가는 얼마야?
2. 롯데칠성의 현재주가는 얼마야?
                                    Answer(score): 149,300원입니다.
```

▶ AT A GLANCE

투자의견	BUY	
목표주가	190,000원	27.3%
현재주가	149,300원	
시가총액	1.4조원	
Shares (float)	9,278,884주 (37.6%)	
52주 최저/최고	118,800원/184,500원	
60일-평균거래대금	38.2억원	

모델 결과 및 예시

```
docker (com.docker.cli)

lotte
car
qualcom
hive
us

Selected: lotte
Enter your question:
 롯데칠성 주식을 살까 말까?
롯데칠성 주식을 살까 말까?
                          Answer(score):  BUY입니다.
```

롯데칠성 (005300)

국내외 점유율 확대에 주목: 24년 24%의 증익을 예상

- BUY 의견 제시하며, 목표주가 19만원으로 상향 조정 (12개월 선행 PER 11배 적용).
- 국내뿐 아니라, 해외에서도 점유율을 확대하며, 원재료 부담을 상쇄하고도 남는 수익성 개선이 예상됨. PCPPI 연결 편입이 더해지며, 24년 영업이익 24% y-y 성장 전망.
- 글로벌 확장성 보유한 동사의 밸류에이션 눈높이는 상향되어야 할 것으로 예상. 한편, 11월 21일 출시 예정인 맥주 신제품은 추가적인 밸류에이션 눈높이 상향의 열쇠.

모델 결과 및 예시

```
Selected: lotte
Enter your question:
롯 데 칠 성 의  해 외  법 인 의  24년  연 결  영 업 이 익  성 장 에  대 한  기 여 도 는 ?
롯 데 칠 성 의  해 외  법 인 의  24년  연 결  영 업 이 익  성 장 에  대 한  기 여 도 는 ?
                                                    Answer(score): 77.5%에
```

4Q23 PCPPI 연결 편입으로, 해외의 실적 기여도가 큰 폭으로 상승할 전망: 관계 기업으로 분류되었던 PCPPI가 (기존 지분법), 9월 실질적 지배력 확보 완료로, 4Q23부로 연결 매출액과 영업이익 실적에 반영될 예정. 연결 편입으로, 동사 해외 매출액 비중은 23년 15%에서 24년 30.5%까지 확대될 것으로 예상되며, 해외 법인의 24년 연결 영업이익 성장에 대한 기여도는 77.5%에 달할 전망.

의의 및 한계

프로젝트 의의

- 사전훈련된 small-LM을 앙상블하여 기존 모델보다 향상된 성능을 이끌어 냄
- 전처리 자동화를 위한 PDF parsing, OCR, table detection 등의 여러가지 알고리즘 모델 공부 및 시도
- 한국어 언어 모델을 사용하여 번역 과정 없이 처리할 수 있는 파이프라인 형성

프로젝트 한계

컴퓨팅 리소스 부족

RAM 용량의 부족으로 다양한 LLM방법을 시도하는 데에 어려움

API 호출 불가

open API 사용 제한으로 좀 더 효율적인 방법 고안의 한계

번역모델의 한계

영한번역에서 정보의 손실을 우려해 정확도가 낮은 한국어 모델을 차선으로 택함

전처리 자동화의 어려움

일반적인 문서에도 적용 가능한 자동화 구현의 어려움 (시간 상의 문제)

감사합니다.

2023 삼성증권 디지털/IT 학회 연계 프로그램
디지털기술팀 8조