Assignment 6

프로젝트 최종 보고서

과목명	영어분석을 위한 기계학습
담당교수	최승택
제출일	2025년 12월 11일
전공	한국어교육과
학번	202302592
이름	임채운

1. 프로젝트 개요

1.1. 프로젝트명

: 교재 학습활동 추출 도구

1.2. 문제 상황

사범대 전공 특성상 교과서나 교재를 보아야 하는 경우가 매우 많다. 교과서의 종류가 과목별, 수준별, 학년별로 전부 다르고 출판사도 다양하여 교재 분석을 하거나 수업 시연 준비를 위해 교과서를 검토하는 데에 시간이 많이 소요된다. 이 수고를 덜기 위해 교과서 안에 어떤 글이 들어있는지, 어떤 질문이 들어있는지 등을 추출하여 보여주는 도구가 필요하다고 생각했다.

2. 진행 과정

2.1. 주제 선정 및 문제 정의

앞선 문제 상황을 해결하기 위해, 교과서 속 **학습활동 질문 추출 도구를 개발**하는 것을 이 프로젝트의 주제로 선정하였다. 처음에는 '성취 기준'이나 '작품'을 기준으로 학습활동을 추출하는 것을 목표로 하여 데이터 수집과 분석까지 진행하였으나, 실제 모델을 학습하는 과정이 너무 복잡하고 이 계획대로 프로젝트를 진행하기에는 개인 역량이 부족하다고 판단하여 기준을 설정하지 않고 단순히 지시문만 추출할 수 있는 도구를 개발하는 것으로 계획을 일부 수정하였다.

2.2. 데이터 수집 및 분석

- 최초 계획에 따른 데이터 수집 및 분석

데이터 수집과 분석 과정에서는 최초 계획에 따라 Assignment 4에서 '성취 기준'을 기준으로 설정하여 교과서 속 지시문을 수동으로 수집하였다. 수집한 데이터에 대한 개요는 다음과 같다.

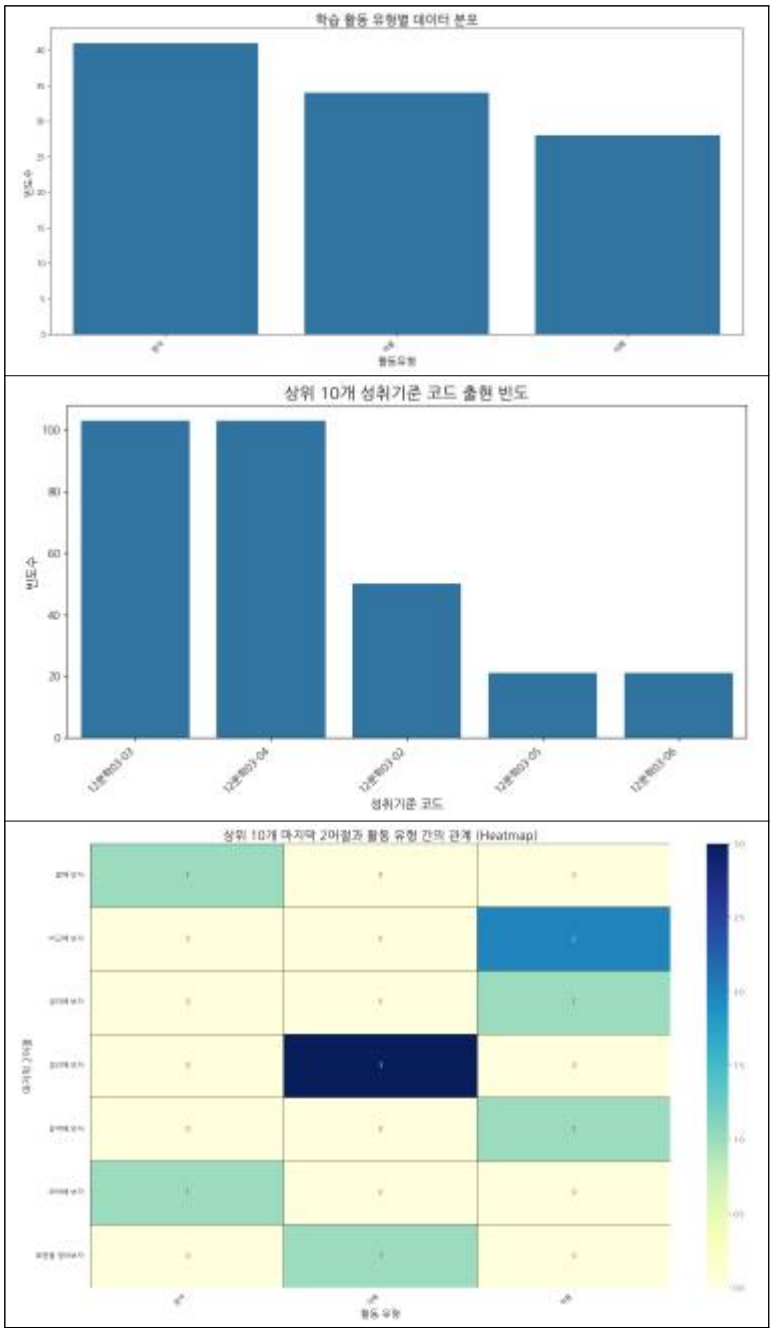
수집 데이터	2015 국어과 교육과정 성취기준, 국어 문학 교과서 (천재교과서, 동아출판, 비상교육, 천재교육)
개수	교육과정 성취기준 16개, 교과서 학습활동 지시문 103개

교육과정 성취기준은 '성취기준 코드 / 영역 / 학년 / 내용'으로 나누어 정리하였고, 교과서 학습활동 지시문은 '출판사 / 단원명 / 페이지 / 코드 / 글감제목 / 글쓴이 / 학습활동 지시문 / 활동 유형 / 질문 형식'으로 나누어 정리하였다. 수집한 자료들은 모두 csv 파일 형태로 저장하였다.

데이터 분석 과정에서 결측값은 존재하지 않았고, 활동 유형별 분포 확인 결과 '분석', '적용', '이해' 순으로 많았던 것을 확인하였다. 이 유형은 학습활동 순서대로 이해, 분석, 적용 영역으로 임의로 부여한 것으로, 라벨링하는 주체에 따라 그 범위가 조금씩 달라질 수 있다. 성취 기준 코드의 출현 빈도 확인 결과 출현 빈도로는 '12문학03-03'과 '12문학03-04'가 가장 많았고, 이 두 항목은 수집된 모든 데이터에 적용된 코드임을 확인하였다. 문학 교과서에

한정되고, 특히 '문학사' 소단원에 한정하여 데이터를 수집하고 정리하였기 때문에 이 결과가 나타난 것으로 보인다. 이에 관해서는 향후 불균형한 데이터를 보완할 필요가 있다. 마지막으로는 학습활동 질문과 활동 유형과의 관계를 찾아보고자 하였다. 이에 따라, '학습활동 질문'의 마지막 2어절을 추출하여 활동 유형 '이해, 분석, 적용'과의 관계를 파악해 보았다. (예시: '말해 보자'로 끝나는 질문은 '이해' 유형에 해당할 것이다) 그러나 분석 결과 두 항목 간의 관계는 나타나지 않았고, 이 질문을 통해 활동 유형을 파악하기는 어려울 것으로 예상되었다.

다음은 데이터 분석과 관련된 그래프다.



수집한 데이터가 문학 교과서의 문학사 부분에만 해당하여 데이터의 불균형이 나타난 것으로 보이므로, 이를 보완하기 위해 국어의 다른 세부 과목 데이터나 다른 소단원의 데이터를 더 수집하고자 하였다.

- 수정된 계획에 따른 데이터 수집

계획이 수정됨에 따라 불필요한 성취 기준 데이터는 폐기하고, 기존에 수집한 학습활동 지시문만 채택하여 다시 정리하였다. 이 과정에서 데이터를 조금 더 추가하여 지시문 데이터를 총 115개를 수집하였다. 지시문이 아닌 일반 문장의 경우, 매우 정제된 패턴을 보이는 교과서의 문장에서 벗어나서 더 다양한 형태의 문장을 수집하기 위해 한국어 위키피디아에서 193개의 문장을 수집하였다. 지시문은 '1'을 할당하였고, 일반 문장은 '0'을 할당하여 정리한 자료는 csv 파일 형태로 저장하였다. 다음은 라벨링한 데이터의 예시다.

sentence	label
단군의 탄생 과정을 중심으로 '변화'라는 관점에서 살펴보고 빈칸을 채워 보자.	1
이 대회의 예선 방식은 조별 리그와 플레이오프로 되어 있다.	0

계획 수정에 따라 시간 관계상 이 데이터에 대해서는 분석 절차를 생략하였다.

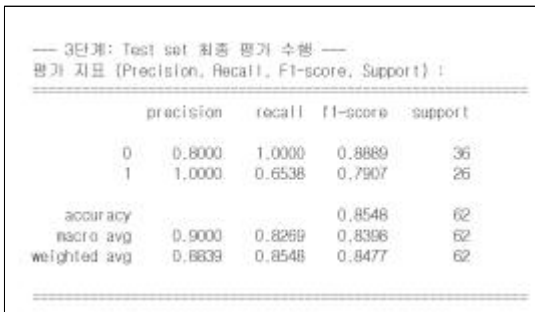
2.3. ML 모델 학습 및 평가

한국어 데이터 처리를 위해 KoBERT 기반 정보추출 모델(skt/kobert-base-v1)을 활용하여 Assignment 4에서 수집한 데이터를 바탕으로 모델을 학습하였다. 다음은 학습 과정에서 분할한 데이터의 개수다.

분류	개수(개)
Train	196
Validation	50
Test	62

이를 바탕으로 학습을 시작하였고, 최초 epoch을 3으로 하여 실행하였으나, evaluation과 inference 성능이 너무 좋지 않아 epoch의 수를 10으로 늘렸다. epoch을 3에서 10으로 점차 늘린 결과, epoch의 수를 늘릴수록 validation accuracy가 항상 높아지는 것은 아님을 발견하였다. 그러나 전반적으로 우상향하는 추세를 보이고 있고, 더불어 training loss가 함께 줄어들고 있는 모습을 볼 수 있었다. epoch 10일 때의 주요 성능지표를 살펴보면, 지시문 추출 기준으로 Accuracy가 0.8, Precision 1.0, Recall 0.65, F1-score 0.79였다.

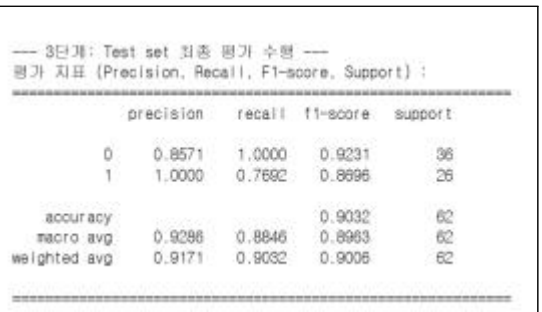
epoch을 늘릴수록 성능은 개선되었지만, Recall 수치가 너무 낮았다. 이 도구의 활용 목적이 '지시문 추출 후 검토'이므로 지시문이 아닌 문장을 추출하더라도 지시문인 문장을 빠짐없이 추출하도록 하고자 Recall 수치를 높이는 방식으로 조정하였다. 이에 따라, learning rate를 2e-5에서 3e-5로 조정하여 모델 학습을 진행하였고, recall 수치가 0.77, F1-score가 0.87로 개선되었다. 다음 표는 모델 학습 및 평가에 관한 요약이다.



--- 3단계: Test set 최종 평가 수행 ---
평가 지표 (Precision, Recall, F1-score, Support) :

	precision	recall	f1-score	support
0	0.8000	1.0000	0.8889	36
1	1.0000	0.6538	0.7907	26
accuracy			0.8548	62
macro avg	0.9000	0.8269	0.8398	62
weighted avg	0.8839	0.8548	0.8477	62

<learning rate 조정 전>



--- 3단계: Test set 최종 평가 수행 ---
평가 지표 (Precision, Recall, F1-score, Support) :

	precision	recall	f1-score	support
0	0.8571	1.0000	0.9231	36
1	1.0000	0.7692	0.8696	26
accuracy			0.9032	62
macro avg	0.9286	0.8846	0.8963	62
weighted avg	0.9171	0.9032	0.9006	62

<learning rate 조정 후 최종>

3. 모델을 서비스로 만든 구조

3.1. 서비스화

학습을 완료하고 실제 사용이 가능하도록 CLI 기반 도구로 만들었다. txb_inference.py 파일이 핵심 로직이며, 모델 가중치는 instruction_classifier_model 폴더에 저장하였고, 모델 서비스에 필요한 라이브러리 목록을 requirements.txt에 저장하였다.

python 기반 서비스이기 때문에 이 서비스를 사용하기 위해서는 python이 설치된 환경이 필요하다. 또한, txt 파일을 입력으로 받기 때문에, txt 형태가 아닌 다른 형태의 파일이라면 미리 변환이 필요하다.

3.2. 시스템 구조

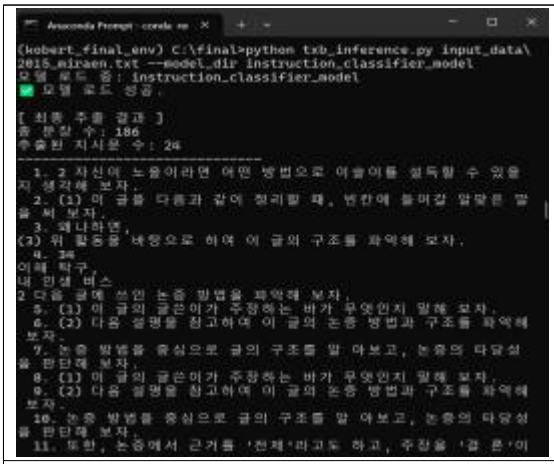
1) CMD나 Anaconda prompt에서 사용자로부터 txt 파일의 경로를 입력받는다.
실행형식:
python txb_inference.py input_data\{파일명}.txt –model-dir instruction_classifier_model
2) 입력받은 파일에서 모델이 사전 학습된 '지시문'만 추출한다.
3) 추출된 지시문들이 콘솔 창을 통해 사용자에게 제시된다.

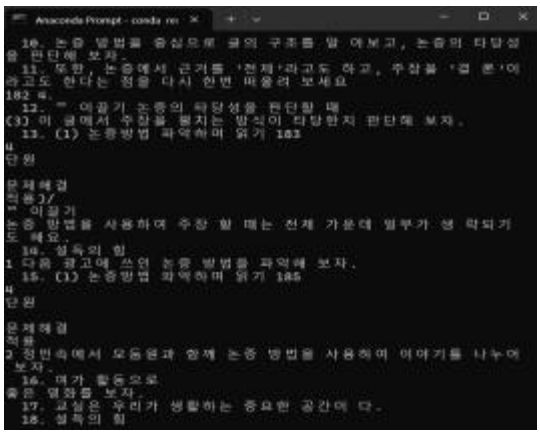
4. 실제 사용 결과

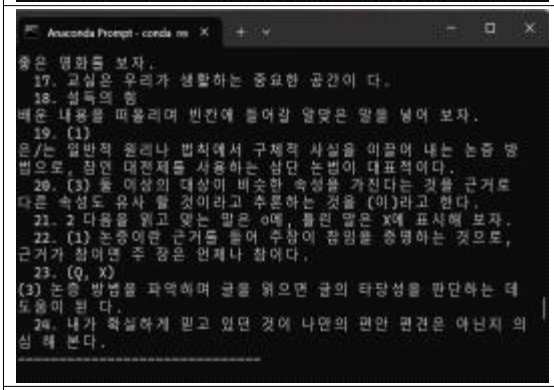
4.1. 5회 이상 사용 기록 (스크린샷)

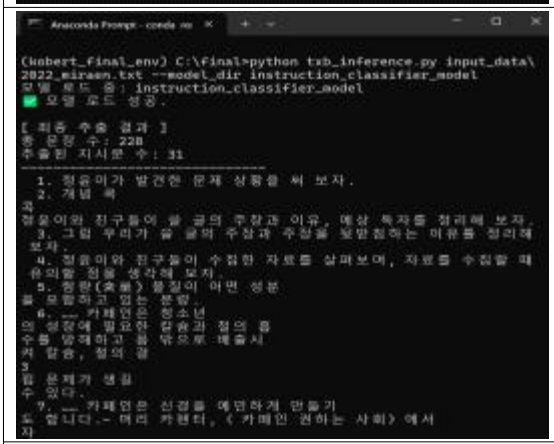
1) <국어논리및논술> 국어 교과서 제작에 필요한 교재 분석

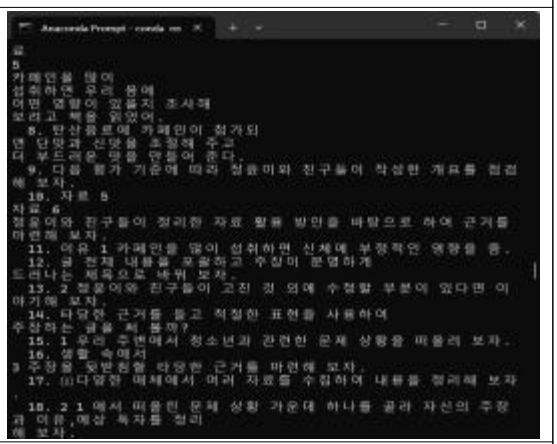
<국어논리및논술> 과목에서 논리 교과서 제작을 하였다. 이를 위해 중학교 국어 교과서 속 논리 논술 단원을 분석하였고, 그 과정에서 이 도구를 활용하여 교과서 5종의 학습활동 속 지시문을 추출하여 교과서 제작에 참고하였다. OCR을 활용하여 미리 교과서에서 수동 추출한 텍스트를 txt 파일로 저장하여 도구에 제공하였고, 그 파일을 바탕으로 도구가 추출한 문장들을 출력하였다. 전반적으로 지시문을 거의 빠짐없이 추출하는 편이었으나, 불필요한 단어나 지시문이 아닌 문장들을 함께 추출하는 모습을 보였다. 다음은 실제 도구를 사용한 스크린샷 기록이다.

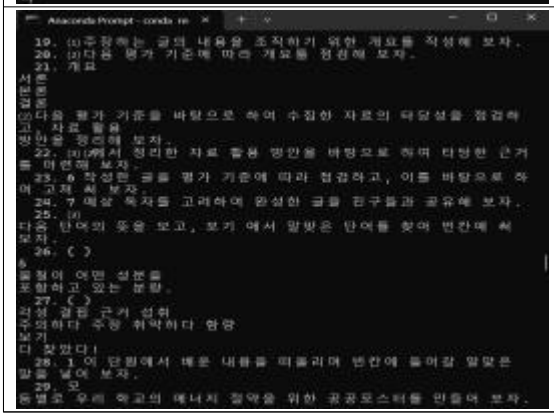

Anaconda Prompt - conda rn
(kobert_final_env) C:\final>python txb_inference.py input_data\
2015_miraen.txt --model_dir instruction_classifier_model
모델 로드 중: instruction_classifier_model
✅ 모델 로드 성공.
[최종 추출 결과]
총 문장 수: 186
추출된 지시문 수: 24
1. 2 자신이 노을이라면 어떤 방법으로 이슬이를 설득할 수 있을
지 생각해 보자.
2. (1) 이 글을 다음과 같이 정리할 때, 빈칸에 들어갈 알맞은 말
을 써 보자.
3. 왜냐하면,
(3) 위 활동을 바탕으로 하여 이 글의 구조를 파악해 보자.
4. 38
이해 탐구
내 인생 버스
2 다음 글에 쓰인 논증 방법을 파악해 보자.
5. (1) 이 글의 글쓴이가 주장하는 바가 무엇인지 말해 보자.
6. (2) 다음 설명을 참고하여 이 글의 논증 방법과 구조를 파악해
보자.
7. 논증 방법을 중심으로 글의 구조를 알 아보고, 논증의 타당성
을 판단해 보자.
8. (1) 이 글의 글쓴이가 주장하는 바가 무엇인지 말해 보자.
9. (2) 다음 설명을 참고하여 이 글의 논증 방법과 구조를 파악해
보자.
10. 논증 방법을 중심으로 글의 구조를 알 아보고, 논증의 타당성
을 판단해 보자.
11. 또한, 논증에서 근거를 '전제'라고도 하고, 주장을 '결 론'이

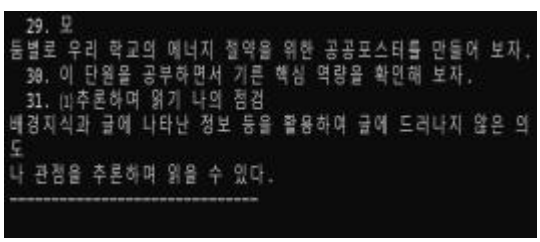

Anaconda Prompt - conda rn
10. 논증 방법을 중심으로 글의 구조를 알 아보고, 논증의 타당성
을 판단해 보자.
11. 또한, 논증에서 근거를 '전제'라고도 하고, 주장을 '결 론'이
라고도 한다는 점을 다시 한번 떠올려 보세요
182 4.
12. " 이끌기 논증의 타당성을 판단할 때
(3) 이 글에서 주장을 펼치는 방식이 타당한지 판단해 보자.
13. (1) 논증방법 파악하며 읽기 183
4
단원
문제해결
적용)/
" 이끌기
논증 방법을 사용하여 주장 할 때는 전제 가운데 일부가 생 략되기
도 해요.
14. 설득의 힘
1 다음 글고에 쓰인 논증 방법을 파악해 보자.
15. (1) 논증방법 파악하며 읽기 185
4
단원
문제해결
적용
2 정반의에서 모둠원과 함께 논증 방법을 사용하여 이야기를 나누어
보자.
16. 여가 활동으로
좋은 영화를 보자.
17. 교실은 우리가 생활하는 중요한 공간이 다.
18. 설득의 힘

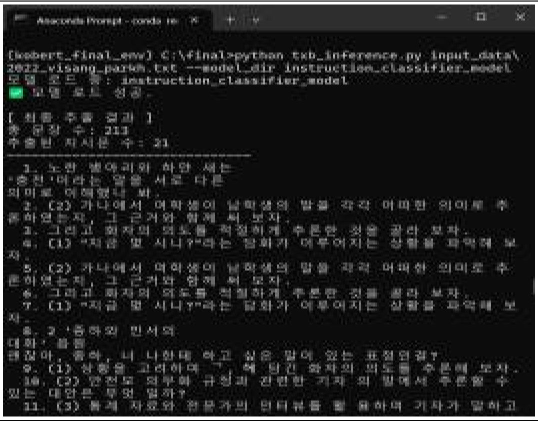

Anaconda Prompt - conda rn
좋은 영화를 보자.
17. 교실은 우리가 생활하는 중요한 공간이 다.
18. 설득의 힘
배운 내용을 떠올리며 빈칸에 들어갈 알맞은 말을 넣어 보자.
19. (1)
은/는 법반적 원리나 법칙에서 구체적 사실을 이끌어 내는 논증 방
법으로, 참인 대전제를 사용하는 삼단 논법이 대표적이다.
20. (3) 둘 이상의 대상이 비슷한 속성을 가진다는 것을 근거로
다른 속성도 유사 할 것이라고 추론하는 것을 (이)라고 한다.
21. 2 다음을 읽고 맞는 말은 o에, 틀린 말은 X에 표시해 보자.
22. (1) 논증이란 근거를 들어 주장이 참임을 증명하는 것으로,
근거가 참이면 주 장은 언제나 참이다.
23. (Q, X)
(3) 논증 방법을 파악하며 글을 읽으면 글의 타당성을 판단하는 데
도움이 된 다.
24. 내가 확실하게 믿고 있던 것에 나만의 편안 편견은 아닌지 의
심 해 본다.

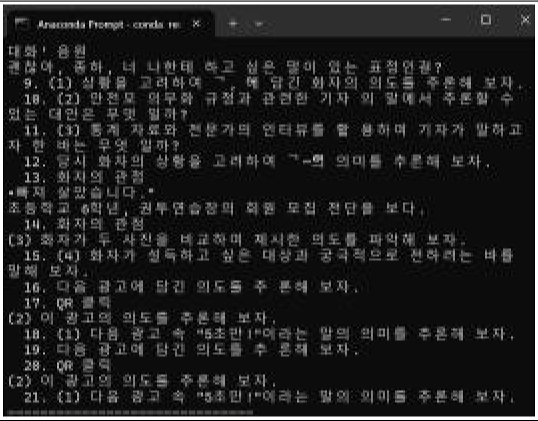

Anaconda Prompt - conda rn
(kobert_final_env) C:\final>python txb_inference.py input_data\
2022_miraen.txt --model_dir instruction_classifier_model
모델 로드 중: instruction_classifier_model
✅ 모델 로드 성공.
[최종 추출 결과]
총 문장 수: 228
추출된 지시문 수: 31
1. 철윤이가 발견한 문제 상황을 써 보자.
2. 개념 목
적
철윤이와 친구들이 쓴 글의 주장과 이유, 예상 독자를 정리해 보자.
3. 그럼 우리가 쓴 글의 주장과 주장을 뒷받침하는 이유를 정리해
보자.
4. 철윤이와 친구들이 수집한 자료를 살펴보며, 자료를 수집할 때
유의할 점을 생각해 보자.
5. 형량(호봉) 월급이 어떤 성분
을 모임하고 있는 분량.
6. — 카페인은 청소년
의 성장에 필요한 칼슘과 철의 흡
수를 방해하고 뼈 약으로 배출시
켜 칼증, 철의 결
3
립 문제가 생길
수 있다.
7. — 카페인은 신경을 예민하게 만들기
도 합니다.- 머리 카펜티, 〈 카페인 권하는 사회〉 에서
자

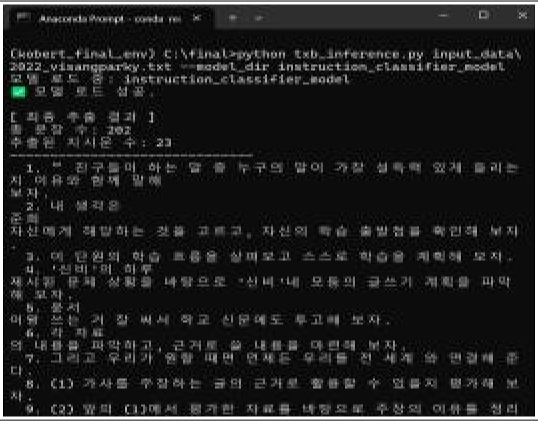

료
5
카페인을 많이
섭취하면 우리 몸에
어떤 영향이 있을지 조사해
보라고 책을 읽었어.
8. 탄산음료에 카페인이 첨가되
면 단맛과 신맛을 조절해 주고
더 부드러운 맛을 만들어 준다.
9. 다음 평가 기준에 따라 철윤이와 친구들이 작성한 개요를 점검
해 보자.
10. 자료 5
자료 6
철윤이와 친구들이 정리한 자료 활용 방안을 바탕으로 하여 근거를
마련해 보자.
11. 이유 1 카페인을 많이 섭취하면 신체에 부정적인 영향을 줌.
12. 글 전체 내용을 모음하고 주장이 분명하게
드러나는 제목으로 바꿔 보자.
13. 2 철윤이와 친구들이 고친 것 외에 수정할 부분이 있다면 이
야기해 보자.
14. 타당한 근거를 들고 적절한 표현을 사용하여
주장하는 글을 써 봅시다.
15. 1 윤아 그면에서 청소년과 관련한 문제 상황을 떠올려 보자.
16. 생활 속에서
3 주장할 뒷받침할 타당한 근거를 마련해 보자.
17. (1)다양한 매체에서 여러 자료를 수집하여 내용을 정리해 보자
.
18. 2 1 에서 떠올린 문제 상황 가운데 하나를 골라 자신의 주장
과 이유,예상 독자를 정리
해 보자.

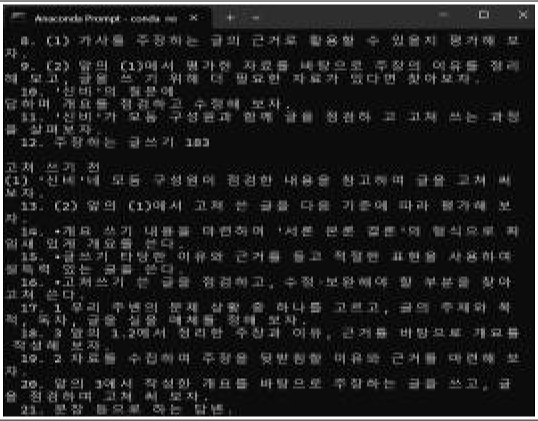

Anaconda Prompt - conda rn
19. (1)주장하는 글의 내용을 조직하기 위한 개요를 작성해 보자.
20. (2)다음 평가 기준에 따라 개요를 점검해 보자.
21. 개요
서론
본론
결론
(2)다음 평가 기준을 바탕으로 하여 수집한 자료의 타당성을 점검하
고 자료를 활용
방안을 정리해 보자.
22. (1)에서 정리한 자료 활용 방안을 바탕으로 하여 타당한 근거
를 마련해 보자.
23. 6 작성한 글을 평가 기준에 따라 점검하고, 이를 바탕으로 하
여 고쳐 써 보자.
24. 7 예상 독자를 고려하여 완성한 글을 친구들과 공유해 보자.
25. (2)
다음 단어의 뜻을 보고, 보기 에서 알맞은 단어를 찾아 빈칸에 써
보자.
26. ( )
6
철윤이 어떤 성분을
포함하고 있는 분량.
27. ( )
적절 결점 근거 설취
주의하다 주장 취약하다 함유
보기
다 찾았다!
28. 1 이 단원에서 배운 내용을 떠올리며 빈칸에 들어갈 알맞은
말을 넣어 보자.
29. 또
둠별로 우리 학교의 에너지 절약을 위한 공공포스터를 만들어 보자.

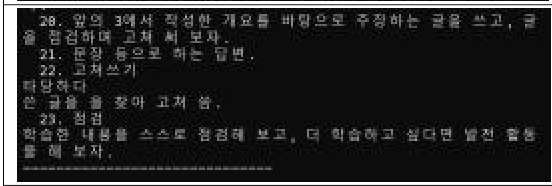

29. 또
둠별로 우리 학교의 에너지 절약을 위한 공공포스터를 만들어 보자.
30. 이 단원을 공부하면서 기른 핵심 역량을 확인해 보자.
31. (1)추론하며 읽기 나의 점검
배경지식과 글에 나타난 정보 등을 활용하여 글에 드러나지 않은 의
도
나 관점을 추론하며 읽을 수 있다.

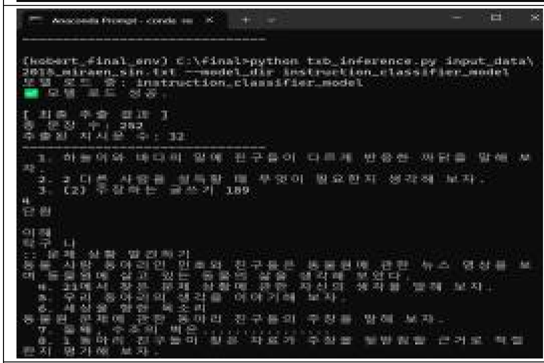

Anaconda Prompt - conda
(kobert_final_env) C:\final>python txb_inference.py input_data\2022_visang_park4h.txt --model_dir instruction_classifier_model
모델 로드 중: instruction_classifier_model
모델 로드 성공.
[최종 추출 결과]
총 문장 수: 213
추출된 지시문 수: 21

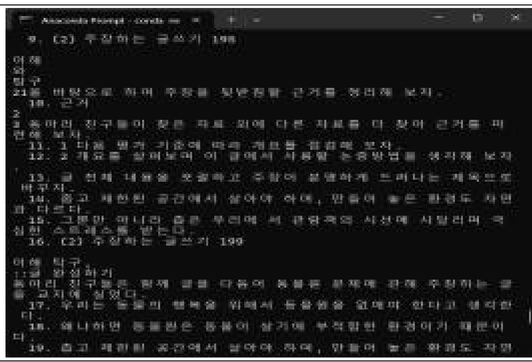

Anaconda Prompt - conda
대화' 응원
괜찮아, 좋아, 너 나한테 하고 싶은 말이 있는 표정인걸?
9. (1) 상황을 고려하여 ㄱ, ㅌ에 담긴 화자의 의도를 추론해 보자.
10. (2) 안전모 의무와 규정과 관련한 기자 의 말에서 추론할 수 있는 대안은 무엇 일까?
11. (3) 통계 자료와 전문가의 인터뷰를 활 용하며 기자가 말하고자 한 바는 무엇 일까?
12. 당시 화자의 상황을 고려하여 ㄱ~ㅁ의 의미를 추론해 보자.
13. 화자의 관점
14. 화자의 관점
(3) 화자가 두 사진을 비교하며 제시한 의도를 파악해 보자.
15. (4) 화자가 설득하고 싶은 대상과 궁극적으로 전하려는 바를 말해 보자.
16. 다음 광고에 담긴 의도를 추 론해 보자.
17. QR 코드
(2) 이 광고의 의도를 추론해 보자.
18. (1) 다음 광고 속 "5초만!"이라는 말의 의미를 추론해 보자.
19. 다음 광고에 담긴 의도를 추 론해 보자.
20. QR 코드
(2) 이 광고의 의도를 추론해 보자.
21. (1) 다음 광고 속 "5초만!"이라는 말의 의미를 추론해 보자.

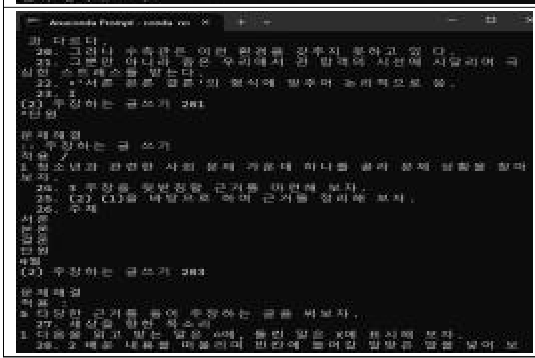

Anaconda Prompt - conda
(kobert_final_env) C:\final>python txb_inference.py input_data\2022_visangparky.txt --model_dir instruction_classifier_model
모델 로드 중: instruction_classifier_model
모델 로드 성공.
[최종 추출 결과]
총 문장 수: 202
추출된 지시문 수: 23
3. 이 단원의 학습 흐름을 살펴보고 스스로 학습을 계획해 보자.
8. (1) 가사를 주장하는 글의 근거로 활용할 수 있을지 평가해 보자.

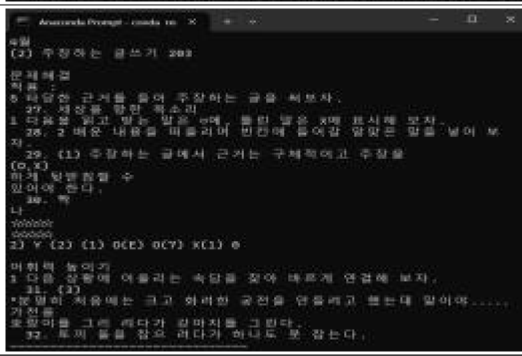

Anaconda Prompt - conda
8. (1) 가사를 주장하는 글의 근거로 활용할 수 있을지 평가해 보자.
12. 주장하는 글쓰기 183
고쳐 쓰기 전
20. 앞의 3에서 작성한 개요를 바탕으로 주장하는 글을 쓰고, 글을 점검하며 고쳐 써 보자.
21. 문장 등으로 하는 답변.


20. 앞의 3에서 작성한 개요를 바탕으로 주장하는 글을 쓰고, 글을 점검하며 고쳐 써 보자.
21. 문장 등으로 하는 답변.
22. 고쳐쓰기
타당하다
쓴 글을 을 찾아 고쳐 씀.
23. 점검
학습한 내용을 스스로 점검해 보고, 더 학습하고 싶다면 발전 활동을 해 보자.


Anaconda Prompt - conda
(kobert_final_env) C:\final>python txb_inference.py input_data\2015_miraen_sin.txt --model_dir instruction_classifier_model
모델 로드 중: instruction_classifier_model
모델 로드 성공.
[최종 추출 결과]
총 문장 수: 252
추출된 지시문 수: 32


Anaconda Prompt - conda
9. (2) 주장하는 글쓰기 195


Anaconda Prompt - conda
(2) 주장하는 글쓰기 201
(2) 주장하는 글쓰기 203


Anaconda Prompt - conda
(2) 주장하는 글쓰기 203

2) 한국어 문화 수업 자료 만들기를 위한 교재 분석

위의 과정에서 학습 데이터에 등장했거나 그와 비슷한 표현을 찾아냈던 것을 볼 수 있었는데, 눈여겨볼 것은 '-하자' 형태의 지시문이 아닌 '-무엇일까?'로 끝나는 지시문을 찾아낸 것이다. 혹시 '무엇'과 같은 다른 질문 패턴을 학습하였는지 궁금증이 생겨 다른 교재에도 적용해 보았다. 학습 데이터의 거의 모든 지시문이 '-자' 형태였기 때문에 추출할 수 있는 문장은 없을 것이라 예상했지만, 확인된 결과에 따라 모델을 어떻게 개선하면 좋을지 파악할 수 있을 것 같아 시도해 보았다. <한국문화의 이해> 과목에서 한국어 문화 자료를 만들어야 했고, 이때 한국어 교재의 문화 파트를 분석하면서 도구를 사용해 보았다.

예상대로 추출된 지시문은 없었음을 확인하였다. 한국어 교재 속 지시문을 수집하여 학습 데이터로 활용한다면 한국어 교재에도 활용할 수 있을 것으로 보인다.

3) 그 외 활용 가능성

이 도구를 주로 이번 학기에 수강한 강의에서 사용하기는 했으나, 앞으로도 매 학기 전공 수업에서도 계속해서 활용할 수 있을 것으로 예상되며, 성능을 더 개선한다면 실제 교육 현장에서도 유용하게 사용할 수 있을 것으로 예상된다.

5. 결론

5.1. 배운 점

- 프로젝트에 활용할 모델을 선정하는 과정에서 다양한 모델의 역할과 목적을 알게 되었다. 이전까지는 BERT라는 모델에 대해 기본적인 내용만 알고 있었고, 실제로 사용해 본 적이 없었는데, 이번 프로젝트를 진행하면서 BERT뿐만 아니라 KoBERT라는 모델이 존재한다는 것을 알게 되었고, 실제로 활용해 보면서 모델의 기능을 알게 되었다. 또한, 어떤 모델을 활용할지 찾아보고 선정하는 과정에서 그 밖의 다양한 모델에 대해서도 많이 알게 된 시간이었다.
- 간단한 도구를 만든다고 해도 수집해야 할 데이터의 규모나 형식, 전처리 등 고려해야 할 것이 매우 많고 복잡하다고 느꼈다. 특히, 전처리만 잘 되어도 모델의 성능에 크게 영향을 준다는 것을 깨달았고, 그만큼 그 중요성을 크게 체감할 수 있었다.

5.2. 개선 방향

1) pdf 파일을 txt파일로 수동으로 변환하는 과정을 자동화한다.

프로젝트를 진행하면서 가장 생각지도 못했던 난관은 바로 모델에 입력하는 자료의 형태였다. 교과서 파일이 일반적인 pdf로 구성되어 있다고 생각하고 프로젝트를 계획하였지만, 모델

학습 과정에서 교과서 파일은 사진을 pdf로 변환한 형태로 제공되거나, e-book 형태로 배포되고 있었다는 것을 깨닫게 되었다. 이 때문에 OCR을 활용하여 직접 수동으로 텍스트를 추출하여 txt 파일로 저장해서 모델에 입력할 수밖에 없었는데, 이 과정 때문에 오히려 시간이 더 많이 소요되기도 했다. 이 도구가 계속해서 사용되려면 이 번거로운 과정에 대한 개선이 꼭 필요하다. OCR 과정을 모델에 넣거나, 그와 비슷한 수준의 자동화 과정이 필요하다.

2) 데이터를 더 정교하게 전처리한다.

실행 결과를 보면, 띄어쓰기의 문제로 출력된 결과가 정돈되지 않은 형태를 보이고 있고, 단원의 제목과 같은 불필요한 텍스트가 함께 출력되고 있다. 띄어쓰기나 문장부호를 처리하고 불필요한 표현은 사전에 없애서 정제된 데이터를 만든다.

3) 학습 데이터의 양과 종류를 확대한다.

학습 데이터의 양을 늘려 국어 교과서에서 한국어 교재나 다른 교재로 확장한다면 더 광범위한 활용이 가능할 것으로 예상된다.