영어분석을위한기계학습 – 6th Assignment Final Report

프로젝트명: **BonCahier AI — 수업 자료 자동 요약과 번역 및 핵심 정리 도구**
과목명: **영어분석을위한기계학습**
학번: **201903774**
이름: **한형준**
전공: **언어인지과학과**
제출일: **2025-12-17**

1. 프로젝트 개요

현대 대학 수업 환경에서 학생들은 여러 과목의 강의 자료(PPT, PDF)를 동시에 관리해야 하고, 과제·시험 일정이 몰리는 시기에는 **짧은 시간 안에 많은 슬라이드를 정리**해야 하는 문제가 있다. 특히 다음과 같은 어려움이 존재한다.

- 텍스트 양이 많은 슬라이드(특히 이론 설명 위주 과목)의 경우,
  직접 요약 노트를 만드는 데 시간이 과도하게 소요됨
- 수업 자료가 영어 등 **외국어로 작성된 경우**,
  번역과 이해를 동시에 해야 하므로 인지적 부담이 큼
- 한 학기 동안 받은 모든 파일(PPT, PDF)을 시험 직전에 다시 열어보며
  핵심만 정리하기가 매우 비효율적

이를 해결하기 위해, 본 프로젝트에서는 강의 자료(PPT, PDF)를 업로드하면 **슬라이드/페이지 단위로 텍스트를 추출하고, KoBART 기반 모델로 핵심 요약과 번역을 자동 생성**하는 도구 "BonCahier AI"를 설계·구현하였다.

2. 진행 과정 요약 (Assignment 3–5)

2.1 Assignment 3 – 프로젝트 제안서 작성

Assignment 3에서는 먼저 **문제 정의와 서비스 아이디어**를 구체화했다.

- 목표 서비스:
  - PPT/PDF 강의 자료 업로드 → **핵심 요약 + 번역 + 간단한 설명** 자동 생성
- 주요 기능:

    - 파일 업로드(PPT/PDF) 및 텍스트 추출(PyMuPDF, python-pptx)
    - 문단/슬라이드 단위 요약
    - 외국어 자료에 대한 한국어 번역
    - 결과를 Markdown/PDF 형식으로 저장
- 기대 효과:
    - 시험 대비, 과제 준비, 발표 준비 시
      "핵심 요약 노트"를 자동으로 얻어 **복습 효율 극대화**

이 단계에서는 실제 구현보다는 전체 그림(UX 시나리오, 모델 후보, 기술 스택)을 설계하는 데 집중했다.

2.2 Assignment 4 – 데이터 수집 및 분석

Assignment 4에서는 강의 슬라이드 요약 태스크와 유사한 구조를 갖는 **요약 데이터셋을 수집·분석(EDA)**하였다.

- 한국어 요약 데이터:
    - daekeun-ml/naver-news-summarization-ko
    - 약 2만 건 이상의 뉴스 기사(document)와 요약문(summary) 쌍
    - 결측치 거의 없고, 카테고리 분포·문장 길이 분포 등을 확인
- 영어 요약 데이터:
    - gopalkalpande/bbc-news-summary
    - BBC 뉴스 기사와 요약문 쌍

EDA 결과, 한국어 네이버 뉴스 데이터는 **요약 태스크의 대표적인 벤치마크**로 활용 가능했고, 문장 길이와 요약 길이 분포도 강의 슬라이드의 "설명형 텍스트"와 어느 정도 유사하다고 판단하였다. 따라서 Assignment 5에서는 이 데이터셋을 **슬라이드 요약 태스크의 proxy corpus**로 사용하기로 했다.

2.3 Assignment 5 – KoBART 요약 모델 학습 및 평가

Assignment 5에서는 HuggingFace의 gogamza/kobart-base-v2를 기반으로 네이버 뉴

스 요약 데이터에 대해 **KoBART 한국어 요약 모델을 미세조정(fine-tuning)** 했다.

- 데이터 전처리 및 분할:
    - Train/Validation/Test = 8: 1: 1 비율
    - 재현성을 위해 seed = 42 고정
- 학습 설정:
    - MAX_SOURCE_LENGTH = 512, MAX_TARGET_LENGTH = 128
    - BATCH_SIZE = 4, NUM_TRAIN_EPOCHS = 4, LEARNING_RATE = 3e-5
- 모델 저장:
    - 학습 완료 후 OUTPUT_DIR에 모델과 토크나이저 저장 →
      이후 evaluation.ipynb, inference.ipynb, 그리고 이번 Assignment 6의 boncahier_service.py에서 재사용 가능하도록 구조화

평가는 ROUGE-1, ROUGE-2, ROUGE-L 지표를 중심으로 이루어졌으며, 기본 KoBART 모델 대비 한국어 뉴스 요약 품질이 향상되는 것을 목표로 하였다.

지표	값
eval_loss	0.505
ROUGE-1	35.62
ROUGE-2	14.46
ROUGE-L	34.99
ROUGE-Lsum	44.90

3. Assignment 6 – 모델 서비스화 구조

Assignment 6의 목표는 **지금까지 학습한 모델을 실제로 사용할 수 있는 서비스 형태로 만들고**, 이를 최소 5회 이상 사용해 본 뒤, 전체 과정을 최종 보고서에 정리하는 것이다.

이번 단계에서 구현한 핵심은 **CLI 기반 요약·번역 도구**인 boncahier_service.py이다.

3.1 시스템 아키텍처 개요

1. **입력 단계**
    - 사용자로부터 .pdf 또는 .pptx 파일 경로를 인자로 받는다 (-i/--input).

2. **텍스트 추출 단계**
    - PDF: PyMuPDF(fitz)를 사용하여 각 페이지의 텍스트를 추출
    - PPTX: python-pptx를 사용하여 각 슬라이드의 텍스트를 추출
3. **언어 감지(간단 heuristic)**
    - 텍스트 내에 한글(유니코드 범위 ₩uac00–₩ud7a3)이 포함되었는지 확인하여 대략적으로 "한국어 vs 비한국어(주로 영어)"를 구분
4. **번역 단계 (옵션)**
    - 비한국어(주로 영어)로 추정되는 텍스트는 Helsinki-NLP/opus-mt-tc-big-en-ko 모델로 **영→한 번역** 후 요약
    - 한국어 요약 결과는 필요시 Helsinki-NLP/opus-mt-ko-en으로 **한→영 번역**
5. **요약 단계**
    - KoBART 기반 요약 모델(gogamza/kobart-base-v2 또는 fine-tuned 모델)로 문단 단위 한국어 요약 수행
6. **결과 저장**
    - 각 페이지/슬라이드별로
        - 원문 텍스트
        - 한국어 요약
        - (옵션) 영어 번역 요약
        - 비고(번역 사용 여부, 에러 등)
    - 위의 내용을 Markdown 파일(summary_{파일이름}_{타임스탬프}.md)로 저장

이 구조를 통해, **하나의 커맨드로 강의 슬라이드 전체에 대한 요약 노트**를 얻을 수 있다.

3.2 주요 스크립트 설명 – boncahier_service.py

- extract_from_pdf(path)
    - PyMuPDF를 이용해 PDF 각 페이지의 텍스트를 추출하고, {"id": "page_1", "kind": "page", "text": "..."} 형태 리스트로 반환.

- extract_from_pptx(path)
    - python-pptx로 슬라이드별 텍스트를 모아 {"id": "slide_1", "kind": "slide", "text": "..."} 형태로 반환.
- load_summarization_model(model_dir, model_name)
    - model_dir가 존재하면 그 디렉토리(학습 완료된 KoBART)를 사용
    - 없으면 model_name (기본: gogamza/kobart-base-v2)에서 모델 로드
- summarize_text(...)
    - KoBART 요약 모델로 한국어 문단 요약 수행
- translate (text, direction, cache)
    - direction이 "en-ko" 또는 "ko-en"인 경우, 적절한 Helsinki-NLP 번역 모델을 로드하여 번역 수행
- process_file(...)
    - 입력 파일 하나를 처리해서 최종 Markdown 요약 파일을 생성하는 상위 함수
- main ()
    - CLI 인자를 파싱하고, process_file()을 호출하여 실제로 요약을 수행

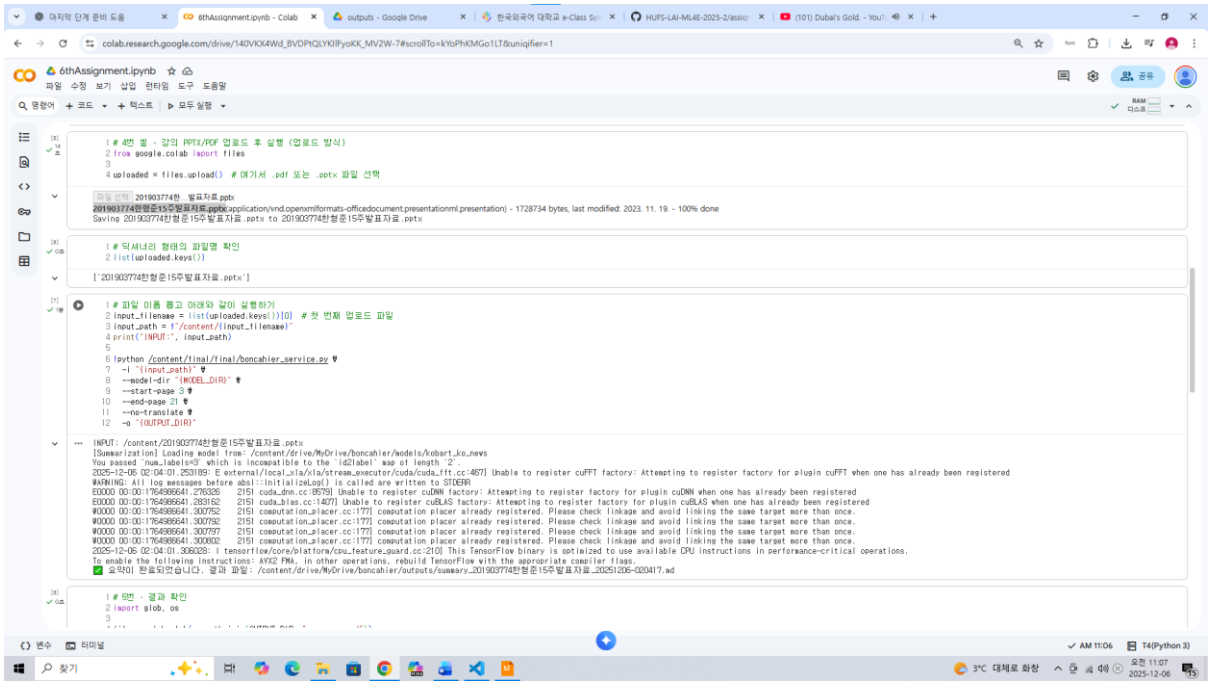
4. 실제 사용 결과 (5회 이상)

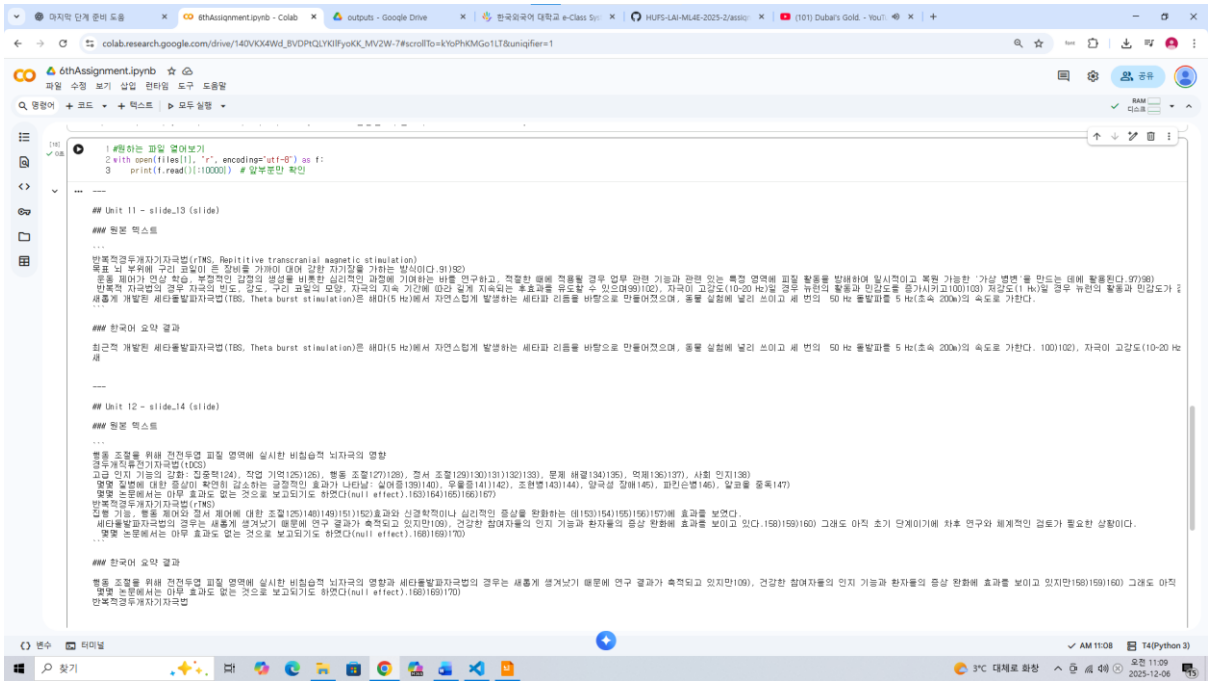
회차	사용 일시	입력 파일 이름	파일 유형	언어	주요 목표	소감
1	2025-12-06 11:00	201903774한형준15주발표자료.pptx	ppt	한국어	요약	핵심 키워드 중심으로 요약은 나쁘지 않다.
2	2025-12-06 11:50	ml4e-lecture-week13.pdf	pdf	영어	요약&번역	영어->한국어 번역에 심각한

						오류가 있다.
3	2025-12-06 12:03	2025fall_7.DNN과 RNN.pdf	pdf	한국어	요약&번역	한국어->영어 번역에도 다소 오류가 있으나 영어->한국어 번역에 비해 심각하지 않음
4	2025-12-06 12:30	NLP_Lec3&4_NLP_TM_slides.pdf	pdf	영어	요약	번역에 비해 핵심 키워드 중심을 찾지만 전처리 과정이 필요함
5	2025-12-06 12:40	PythonReview.pdf	pdf	영어	요약	프로그래밍 언어가 포함된 슬라이드는 오류가 다수 검출됨

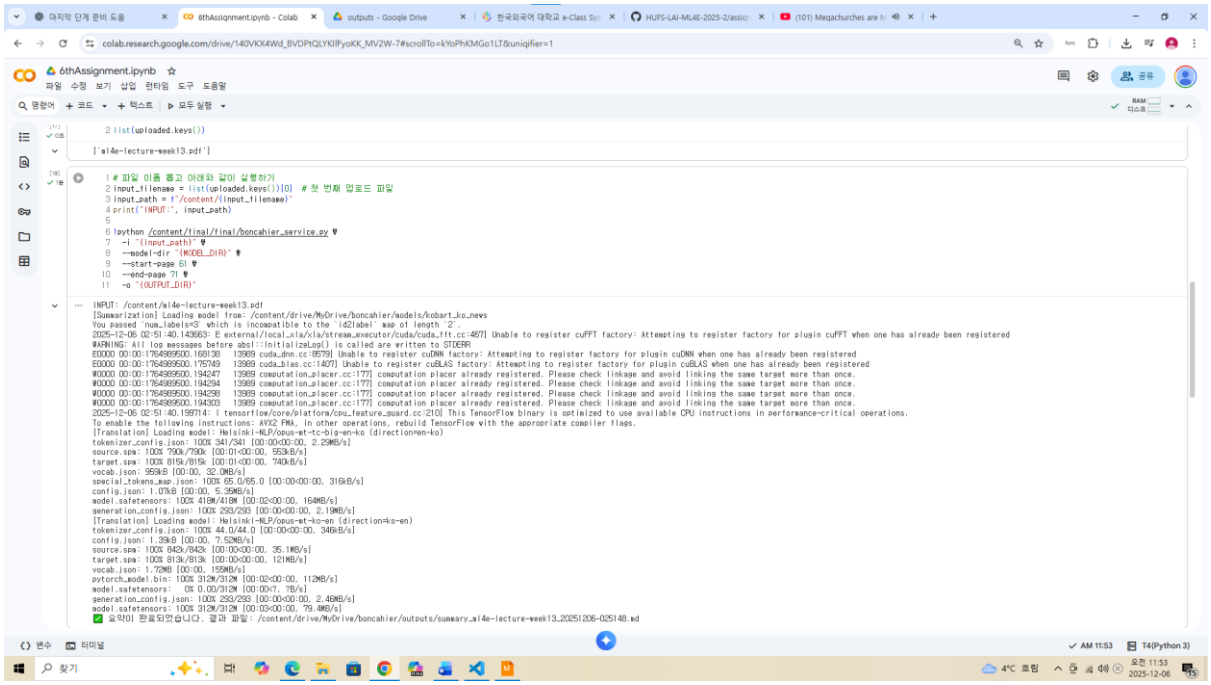
스크린샷:

1회차:



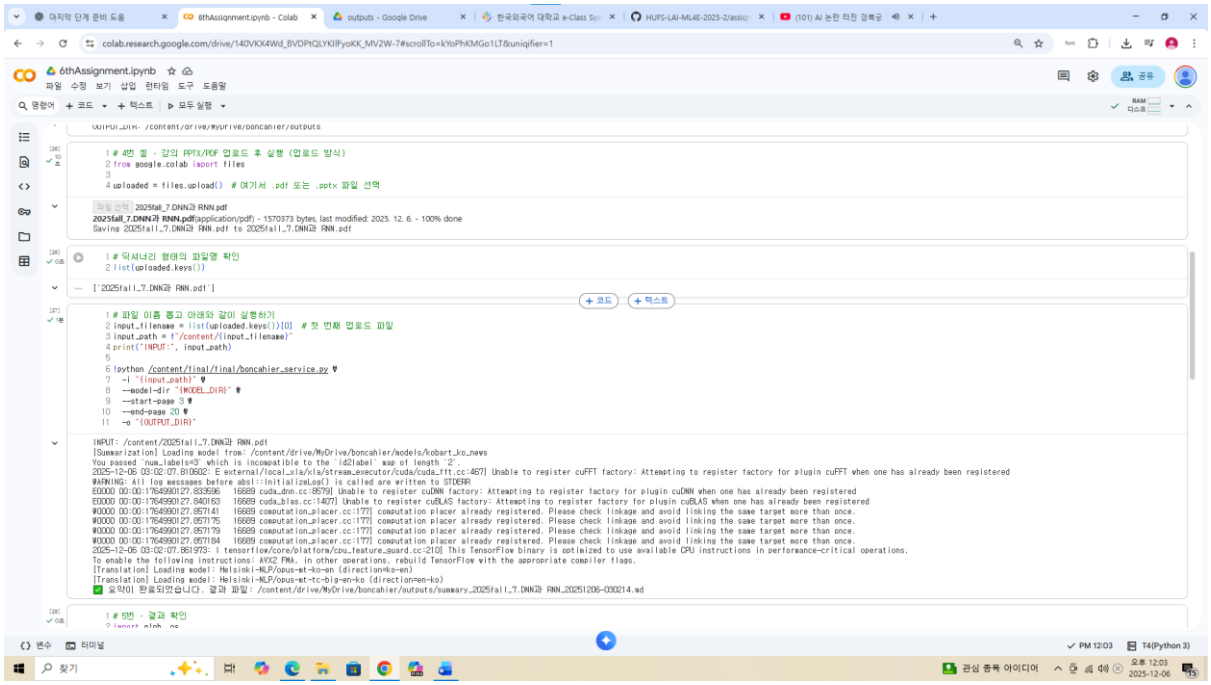


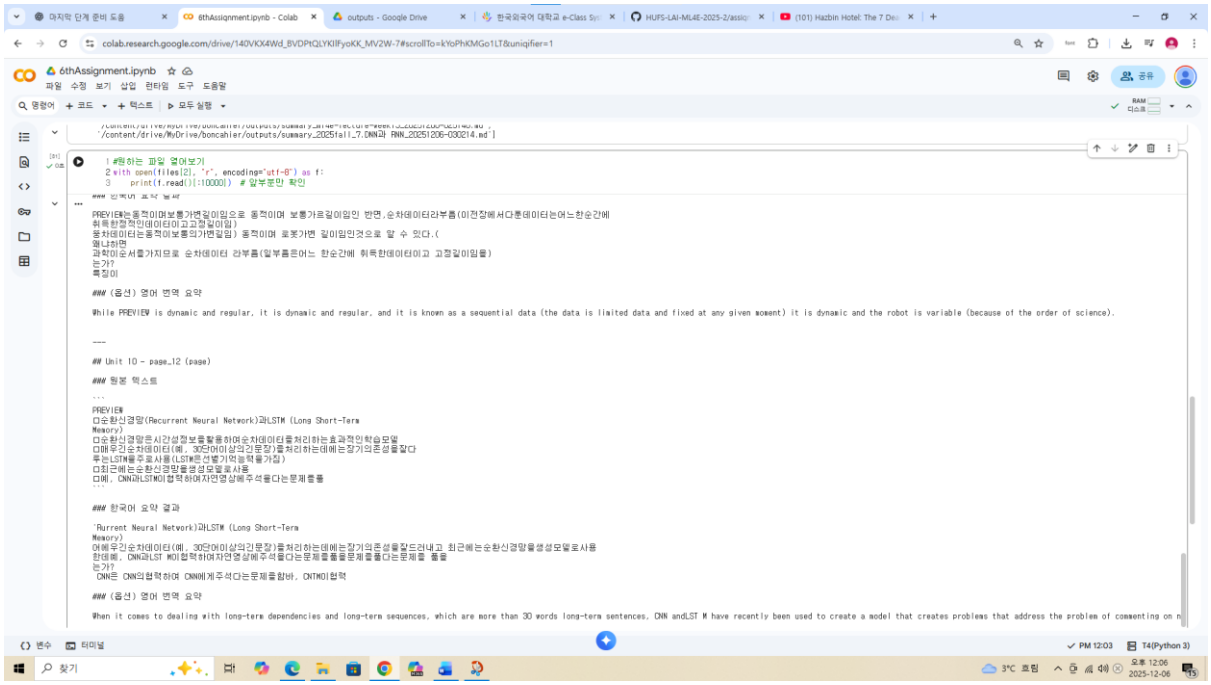
2회차:


6thAssignment.ipynb
list(uploaded.keys())
['wl4e-lecture-week13.pdf']
# 파일 이름 복고 아래와 같이 실행하기
input_filename = list(uploaded.keys())[0] # 첫 번째 업로드 파일
input_path = f"/content/{input_filename}"
print("INPUT:", input_path)
!python /content/final/final/boncahier_service.py
-i "{input_path}"
--model-dir "{MODEL_DIR}"
--start-page 61
--end-page 71
-o "{OUTPUT_DIR}"
INPUT: /content/wl4e-lecture-week13.pdf
[Summarization] Loading model from: /content/drive/MyDrive/boncahier/models/kobart_ko_news
[Translation] Loading model: Helsinki-NLP/opus-mt-tc-big-en-ko (direction=en-ko)
[Translation] Loading model: Helsinki-NLP/opus-mt-ko-en (direction=ko-en)
요약이 완료되었습니다. 결과 파일: /content/drive/MyDrive/boncahier/outputs/summary_wl4e-lecture-week13_20251206-025148.md


6thAssignment.ipynb
#원하는 파일 열어보기
with open(files[1], "r", encoding="utf-8") as f:
print(f.read()[:10000]) # 앞부분만 확인
한국어 요약 결과
(옵션) 영어 번역 요약
비고
- ※ 원문이 비한국어로 추정되어 `en→ko` 번역 후 요약을 수행했습니다.
Unit 7 - page_67 (page)
원본 텍스트
한국어 요약 결과
(옵션) 영어 번역 요약
비고
- ※ 원문이 비한국어로 추정되어 `en→ko` 번역 후 요약을 수행했습니다.

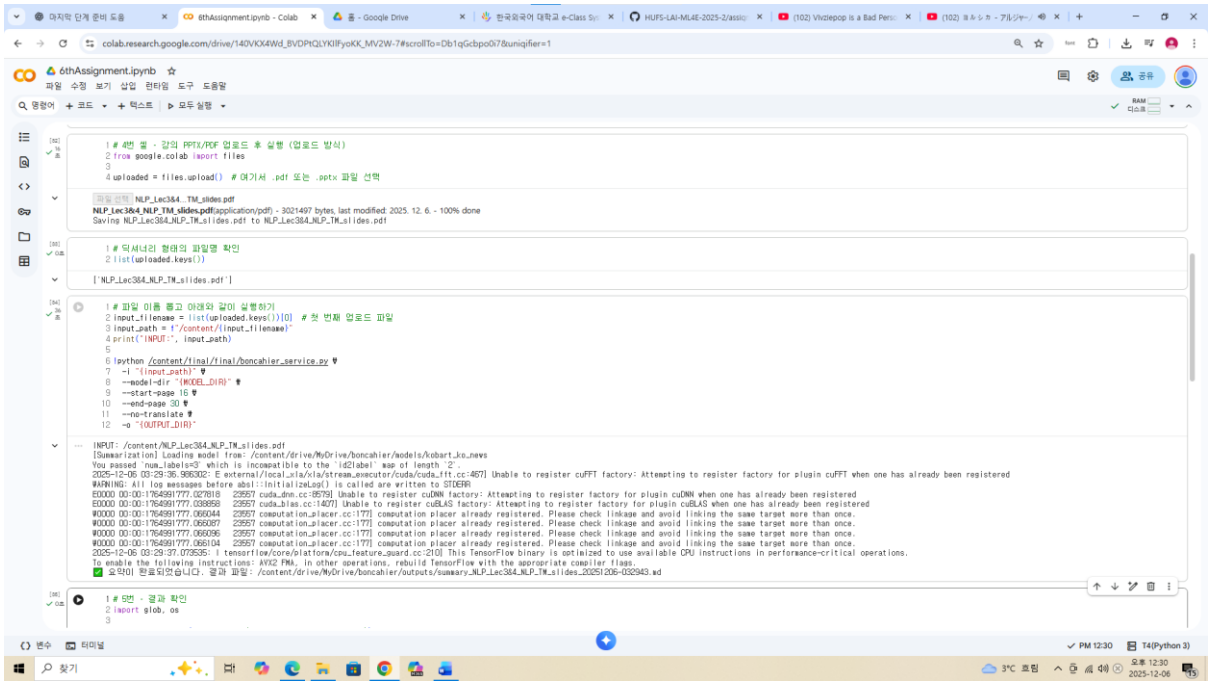
3회차:

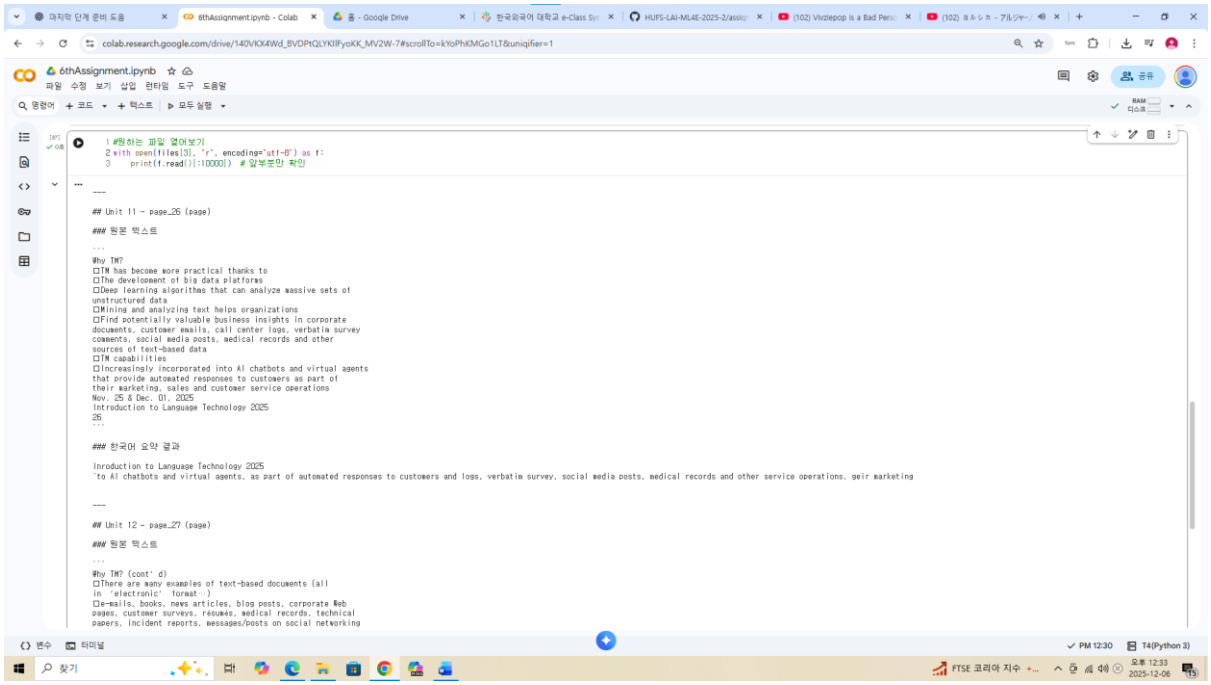

6thAssignment.ipynb
파일 수정 보기 삽입 런타임 도구 도움말
# 4번 셀 - 강의 PPTX/PDF 업로드 후 실행 (업로드 방식)
from google.colab import files
uploaded = files.upload() # 여기서 .pdf 또는 .pptx 파일 선택
2025fall_7.DNN과 RNN.pdf
Saving 2025fall_7.DNN과 RNN.pdf to 2025fall_7.DNN과 RNN.pdf
# 덕셔너리 형태의 파일명 확인
list(uploaded.keys())
['2025fall_7.DNN과 RNN.pdf']
# 파일 이름 동고 아래와 같이 실행하기
input_filename = list(uploaded.keys())[0] # 첫 번째 업로드 파일
input_path = f"/content/{input_filename}"
print("INPUT:", input_path)
!python /content/final/final/boncahier_service.py \
--input "{input_path}" \
--model-dir "{MODEL_DIR}" \
--start-page 3 \
--end-page 20 \
-o "{OUTPUT_DIR}"
INPUT: /content/2025fall_7.DNN과 RNN.pdf
[Summarization] Loading model from: /content/drive/MyDrive/boncahier/models/kobart_ko_news
You passed 'num_labels=5' which is incompatible to the 'id2label' map of length '2'.
2025-12-06 03:02:07.810502: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register cuFFT factory: Attempting to register factory for plugin cuFFT when one has already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1764990127.823596 16689 cuda_dnn.cc:8579] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
E0000 00:00:1764990127.840163 16689 cuda_blas.cc:1407] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
W0000 00:00:1764990127.857141 16689 computation_placer.cc:177] computation placer already registered. Please check linkage and avoid linking the same target more than once.
W0000 00:00:1764990127.857175 16689 computation_placer.cc:177] computation placer already registered. Please check linkage and avoid linking the same target more than once.
W0000 00:00:1764990127.857179 16689 computation_placer.cc:177] computation placer already registered. Please check linkage and avoid linking the same target more than once.
W0000 00:00:1764990127.857184 16689 computation_placer.cc:177] computation placer already registered. Please check linkage and avoid linking the same target more than once.
2025-12-06 03:02:07.861973: I tensorflow/core/platform/cpu_feature_guard.cc:210] This TensorFlow binary is optimized to use available CPU instructions in performance-critical operations.
To enable the following instructions: AVX2 FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
[Translation] Loading model: Helsinki-NLP/opus-mt-ko-en (direction=ko-en)
[Translation] Loading model: Helsinki-NLP/opus-mt-tc-big-en-ko (direction=en-ko)
✅ 요약이 완료되었습니다. 결과 파일: /content/drive/MyDrive/boncahier/outputs/summary_2025fall_7.DNN과 RNN_20251206-030214.md
5번 - 결과 확인
PM 12:03
T4(Python 3)
오후 12:03
2025-12-06


6thAssignment.ipynb
파일 수정 보기 삽입 런타임 도구 도움말
'/content/drive/MyDrive/boncahier/outputs/summary_2025fall_7.DNN과 RNN_20251206-030214.md']
#원하는 파일 열어보기
with open(files[2], "r", encoding="utf-8") as f:
print(f.read()[:10000]) # 앞부분만 확인
PREVIEW는동적이면보통가변길이임으로 동적이며 보통가르길이임인 반면,순차데이터라부름(이전장베서다룬데이터는어느한순간에
허록한정적인데이터이고고정길이임)
동차데이터는동적이보통의가변길임) 동적이며 로봇가변 길이임인것으로 알 수 있다.(
왜냐하면
과적이순서를가지므로 순차데이터 라부름(일부름른어는 한순간에 허록한데이터이고 고정길이임물)
는가?
특징이
(줌선) 영어 번역 요약
While PREVIEW is dynamic and regular, it is dynamic and regular, and it is known as a sequential data (the data is limited data and fixed at any given moment) it is dynamic and the robot is variable (because of the order of science).

Unit 10 - page_12 (page)
원본 텍스트
PREVIEW
다순환신경망(Recurrent Neural Network)과LSTM (Long Short-Term
Memory)
다순환신경망은시간성정보를활용하여순차데이터를처리하는효과적인학습모델
다배우간순차데이터(예, 30단어이상의긴문장)를처리하는데에는장기의존성을잘다
루는LSTM을주로사용(LSTM은선별기억능력을가짐)
다최근에는순환신경망을생성모델로사용
다예, CNN과LSTM이협력하여자연영상에주석을다는문제를풀
한국어 요약 결과
(Rurrent Neural Network)과LSTM (Long Short-Term
Memory)
어배우간순차데이터(예, 30단어이상의긴문장)를처리하는데에는장기의존성을잘드러내고 최근에는순환신경망을생성모델로사용
한데예, CNN과LST MOI협력하여자연영상에주석을다는문제를풀을문제를풀을다는문제를 풀을
는가?
CNN은 CNN의협력하여 CNN에게주석다는문제를함바, CNTNOI협력
(줌선) 영어 번역 요약
When it comes to dealing with long-term dependencies and long-term sequences, which are more than 30 words long-term sentences, CNN andLST M have recently been used to create a model that creates problems that address the problem of commenting on n
PM 12:03
T4(Python 3)
3°C 흐림
오후 12:06
2025-12-06

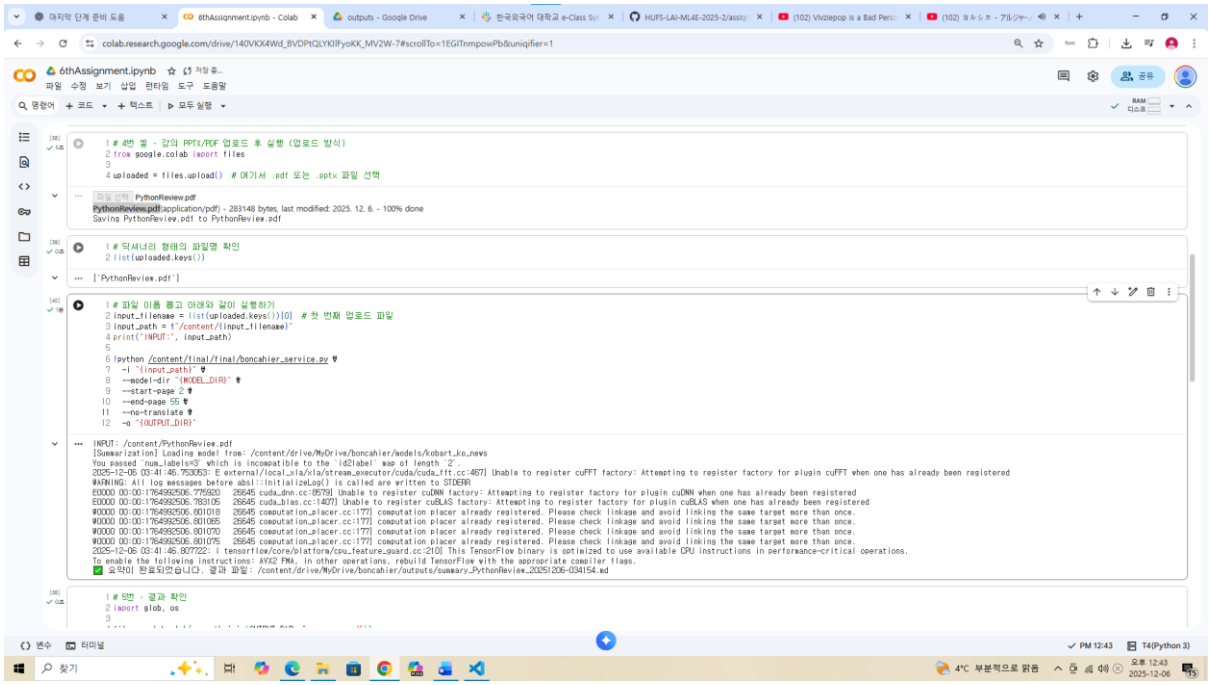
4회차:


6thAssignment.ipynb
# 4번 셀 - 강의 PPTX/PDF 업로드 후 실행 (업로드 방식)
from google.colab import files
uploaded = files.upload() # 여기서 .pdf 또는 .pptx 파일 선택
NLP_Lec3&4_NLP_TM_slides.pdf
NLP_Lec3&4_NLP_TM_slides.pdf(application/pdf) - 3021497 bytes, last modified: 2025. 12. 6. - 100% done
Saving NLP_Lec3&4_NLP_TM_slides.pdf to NLP_Lec3&4_NLP_TM_slides.pdf
# 딕셔너리 형태의 파일명 확인
list(uploaded.keys())
['NLP_Lec3&4_NLP_TM_slides.pdf']
# 파일 이름 복고 아래와 같이 실행하기
input_filename = list(uploaded.keys())[0] # 첫 번째 업로드 파일
input_path = f"/content/{input_filename}"
print("INPUT:", input_path)
!python /content/final/final/boncahier_service.py \
  -i "{input_path}" \
  --model-dir "{MODEL_DIR}" \
  --start-page 16 \
  --end-page 30 \
  --no-translate \
  -o "{OUTPUT_DIR}"
INPUT: /content/NLP_Lec3&4_NLP_TM_slides.pdf
[Summarization] Loading model from: /content/drive/MyDrive/boncahier/models/kobart_ko_news
You passed 'num_labels=3' which is incompatible to the 'id2label' map of length '2'.
2025-12-06 03:29:35.986302: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register cuFFT factory: Attempting to register factory for plugin cuFFT when one has already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1764991777.027818 23557 cuda_dnn.cc:8579] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
E0000 00:00:1764991777.038858 23557 cuda_blas.cc:1407] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
W0000 00:00:1764991777.066044 23557 computation_placer.cc:177] computation placer already registered. Please check linkage and avoid linking the same target more than once.
W0000 00:00:1764991777.066087 23557 computation_placer.cc:177] computation placer already registered. Please check linkage and avoid linking the same target more than once.
W0000 00:00:1764991777.066096 23557 computation_placer.cc:177] computation placer already registered. Please check linkage and avoid linking the same target more than once.
W0000 00:00:1764991777.066104 23557 computation_placer.cc:177] computation placer already registered. Please check linkage and avoid linking the same target more than once.
2025-12-06 03:29:37.073535: I tensorflow/core/platform/cpu_feature_guard.cc:210] This TensorFlow binary is optimized to use available CPU instructions in performance-critical operations.
To enable the following instructions: AVX2 FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
✅ 요약이 완료되었습니다. 결과 파일: /content/drive/MyDrive/boncahier/outputs/summary_NLP_Lec3&4_NLP_TM_slides_20251206-032943.md
# 5번 - 결과 확인
import glob, os


6thAssignment.ipynb
#원하는 파일 열어보기
with open(files[3], 'r', encoding="utf-8") as f:
    print(f.read()[:10000]) # 앞부분만 확인

Unit 11 - page_26 (page)
원본 텍스트
```
Why TM?
□TM has become more practical thanks to
□The development of big data platforms
□Deep learning algorithms that can analyze massive sets of
unstructured data
□Mining and analyzing text helps organizations
□Find potentially valuable business insights in corporate
documents, customer emails, call center logs, verbatim survey
comments, social media posts, medical records and other
sources of text-based data
□TM capabilities
□Increasingly incorporated into AI chatbots and virtual agents
that provide automated responses to customers as part of
their marketing, sales and customer service operations
Nov. 25 & Dec. 01, 2025
Introduction to Language Technology 2025
26
```
한국어 요약 결과
Introduction to Language Technology 2025
'to AI chatbots and virtual agents, as part of automated responses to customers and logs, verbatim survey, social media posts, medical records and other service operations, geir marketing
---
Unit 12 - page_27 (page)
원본 텍스트
```
Why TM? (cont' d)
□There are many examples of text-based documents (all
in 'electronic' format.)
□E-mails, books, news articles, blog posts, corporate Web
pages, customer surveys, résumés, medical records, technical
papers, incident reports, messages/posts on social networking

5회차:
```



로그 발췌 예시:

1회차:

Unit 8 — slide_10 (slide)

원본 텍스트

```

인간 대상 연구에서 발견된 공격적인 행동과 전전두엽 전부 기능 장애의 연관성: 반사회적인 개개인에게는 회백질과 혈역학 반응의 감소, 배외측 전전두엽 피질의 왼쪽, 안와 전두 피질 오른쪽, 전방 대상 피질 오른쪽에 지역적인 뇌 혈류가 관찰되었다.63)

쥐 대상 연구: 내측 전전두엽 피질의 기능을 억제하면 수컷 쥐 간의 공격성이 증가하나 활성화하면 공격성이 감소하고69), 세로토닌이 감소할 수록 공격성이 증가한다.70)

전두엽 손상이 목표에 대한 공격적인 반응과 관련이 있고, 측두엽 손상이 격렬하지만 집중되지 않은 분노와 공격성과 관련이 있다는 가정도 있다.72)

배외측 전전두엽 피질의 손상이 자기 제어를 실패하게 하여 분노 조절에 결함을 일으키고 도발 당하는 경우 공격적인 충동으로 이어진다.73)

공격적인 행동과 전전두엽 피질 장애의 상호 연관성

```

한국어 요약 결과

반사회적인 개개인에게는 회백질과 혈역학 반응의 감소, 배외측 전전두엽 피질의 왼쪽, 안와 전두 피질 오른쪽, 전방 대상 피질 오른쪽에 지역적인 뇌 혈류가 관찰되었다.73)

공격적인 행동과 전전 두엽 피질 장애의 상호 연관성: 내측 전 전두엽피질의 기능을 억제하면 수컷 쥐 간의 공격성이 증가하나 활성화하면 공격성이 감소하고69), 세로토닌이 감소할 수록 공격성이 증가한다.70)

측두엽 손상이 목표에 대한 공격적인 반응과 관련이

2회차:

Unit 7 — page_67 (page)

원본 텍스트

```

Adding nonlinearities in self-attention

• Note that there are no element-wise

nonlinearities in self-attention;

stacking more self-attention layers just

re-averages value vectors

• Easy fix: add a feed-forward network to

post-process each output vector

• $mi= MLPoutputi$

•

$= W2 *ReLUW1\ outputi+ b1 + b2$

https://web.stanford.edu/class/cs224n/

Attention Is All You Need (https://arxiv.org/pdf/1706.03762)

```

한국어 요약 결과

웹 사이트는 귀하가 웹 사이트를 탐색하는 동안 귀하의 경험을 향상시키기 위해 쿠키를 사용하는데 이 쿠키들 중에서 필요에 따라 분류 된 쿠키는 웹 사이트의 기본적인 기능을 수행하는 데 필수적이므로 브라우저에 저장되며 또한 웹 사이트가 사용 방식을 분석하고 이해하는 데 도움이되는 제 3 자 쿠키를 사용한다. 이러한 쿠키를 거부 할 수도 있다. 그러나 이러한 쿠키 중 일부를 선택 해제하면 검색 환경에 영향을 미칠 수 있다. 이러한 쿠키들을 선택 해제 하면 검색 환경에 영향 미칠 수도 있다. 이런 쿠키들 중 일부 선택 해제된 쿠키와 제 3자 쿠키를 사용할 수 있다. 또한 이러한 쿠키들은 웹 사이

(옵션) 영어 번역 요약

The web site uses cookies to improve your experience while you are browsing for web sites, which are required to perform the basic features of web sites, so that cookies are stored in the browser, as well as using my third cookie, which helps the site to analyze and understand how they are used.

비고

- ※ 원문이 비한국어로 추정되어 `en→ko` 번역 후 요약을 수행했습니다.

3회차:

Unit 10 — page_12 (page)

원본 텍스트

```
PREVIEW
ν순환신경망(Recurrent Neural Network)과LSTM (Long Short-Term
Memory)
♣순환신경망은시간성정보를활용하여순차데이터를처리하는효과적인학습모델
♣매우긴순차데이터(예, 30단어이상의긴문장)를처리하는데에는장기의존성을잘다
루는LSTM을주로사용(LSTM은선별기억능력을가짐)
ν최근에는순환신경망을생성모델로사용
♣예, CNN과LSTM이협력하여자연영상에주석을다는문제를풂
```

한국어 요약 결과

'Rurrent Neural Network)과LSTM (Long Short-Term

Memory)

어에우긴순차데이터(예, 30단어이상의긴문장)를처리하는데에는장기의존성을잘드러내고 최근에는순환신경망을생성모델로사용

한데예, CNN과LST M이협력하여자연영상에주석을다는문제를풂을문제를풂다는문제를 풂을

는가?

CNN은 CNN의협력하여 CNN에게주석다는문제를함바, CNTM이협력

(옵션) 영어 번역 요약

When it comes to dealing with long-term dependencies and long-term sequences, which are more than 30 words long-term sentences, CNN andLST M have recently been used to create a model that creates problems that address the problem of commenting on natural images in cooperation with CNN?

4회차:

Unit 9 — page_24 (page)

원본 텍스트

```
What is TM? (cont'd)

ϖText Mining vs. Data Mining

¬Text Mining

⎫Dealing with unstructured textual data
```

– Free-form text

» No pre-defined or organized in any way

⎫Requiring an extra step

– The unstructured data has to be organized and structured in a

way that allows the data modeling and analytics to occur

⎫Concerned with the detection of patterns in NL texts

¬Data Mining

⎫Handling structured (highly formatted) data

⎫Combining disciplines including statistics, AI and machine

learning to apply directly to structured data

⎫Concerned with the detection of patterns in database

Nov. 25 & Dec. 01, 2025

Introduction to Language Technology 2025

24

```

**### 한국어 요약 결과**

'Tith unstructured data modeling and analytics to occur

Concerned with the detection of patterns in NL texts

Data Mining disciplines including statistics, AI and machine and apply directly patted daa

'Introd
```

5회차:

Unit 16 — page_17 (page)

원본 텍스트

```
§ Tuples(are(defined(with(parentheses: (1,(2,(3),(for(example.
• In(essence,(you(can(think(of(a tuple as(a list that(is(not(mutable.
• Thus,(indexing(works(as(with list objects.
§ But(assignment(fails:
• t(=((1,(2,(3)
t[1](=(100
• This(won't(work(because(it(would(change t.
§ If(you(know(that(you(aren't(going(to(change(your(object(at(all,(
then tuple is(probably(a(good(choice(over list.
§ The(built-in tuple() function(will(convert(its(argument(to(a tuple,(
and list() will(covert(its(argument(to list.
§ Examples:
§ x(=((1,(2,(3)
y(=(list(x)
x[1](=(100(##(Throws(an(exception.
y[1](=(100(##(Works(fine;(changes(y.
tuple(list(x))(==(x(##(True!
The$basics$of$dicts,$and$tuples
```

tuple

```

### 한국어 요약 결과

**Thus,(indexing(works)as(with parth list objects)) But(assignment(fails, (1,(2,(3),(for(example,))) function(will(convert(its(argument(to list)) will(covert, its, _ If(won't(wou(that(you(know(thou(**

## 5. 배운 점 및 한계, 향후 개선 방향

### 5.1 배운 점

- **데이터 수집–EDA–모델 학습–서비스화**까지 하나의 흐름을 직접 경험하면서, "단순한 Colab 실습"이 아니라 실제로 **사용 가능한 도구를 만드는 과정**을 익혔다.
- 한국어 요약에 특화된 KoBART 모델을 사용해 보면서, **인퍼런스 속도·토큰 길이·Beam Search 파라미터**가 결과 품질과 시간에 어떤 영향을 미치는지 체감할 수 있었다.
- PyMuPDF, python-pptx, transformers, HuggingFace Datasets 등을 함께 사용하면서 실제 ML 프로젝트에서 흔히 쓰이는 라이브러리들의 역할을 이해하게 되었다.

### 5.2 한계

- 학습 데이터가 **뉴스 기사**이기 때문에, 교수님 슬라이드의 레이아웃/요약 스타일과는 약간의 차이가 있다.
- 영어 슬라이드의 경우, en→ko → 한국어 요약 → ko→en 순서로 파이프라인이 구성되어 en→ko 번역이 ko→en 번역보다 오류 가능성이 심각하게 높다.
- 텍스트 추출 품질은 슬라이드의 특징에 따라 달라지므로, 표/이미지로 이루어지거나 수식, 프로그래밍 언어, 웹사이트 링크가 다수 포함된 슬라이드는 요약이 어렵다.

### 5.3 향후 개선 방향

- 강의 슬라이드 전용 데이터(예: 실제 PPT 텍스트 + 요약 노트)를 구축하여
```

KoBART를 **슬라이드 도메인에 특화된 모델로 재학습**해 볼 수 있다.

- 웹 UI(예: Streamlit, Flask)를 얹어서, 명령어에 익숙하지 않은 사용자도 **드래그 앤 드롭으로 사용할 수 있는 서비스**로 확장 가능하다.
- 번역 부분은 HuggingFace의 다른 번역 모델이나, 필요하다면 외부 API(예: DeepL, Google Translate)를 조합해 품질을 끌어올릴 수 있다.
- 더욱 효율적이고 성능 높은 영어 데이터 및 모델을 다시 선정하여 en→ko 번역의 품질을 향상시킬 수 있다.

6. 결론

본 프로젝트는 **강의 자료 요약·번역 도구 BonCahier AI**를 기획(Assignment 3), 데이터 수집 및 분석(Assignment 4), KoBART 모델 학습 및 평가(Assignment 5), 그리고 실제 사용 가능한 CLI 서비스로 구현(Assignment 6)하는 전 과정을 다루었다.

그 결과, PDF/PPTX 형식의 수업 자료를 입력으로 받아 **슬라이드/페이지 단위로 핵심 내용을 요약하고 번역해 주는 실질적인 도구**를 완성하였고, 이를 실제 강의 자료에 적용하여 **시험 대비용 요약 노트 자동 생성**이라는 초기 목표를 달성할 수 있었다.