Pdf 요약하는 코드 Troubleshooting 문서 2 – 금동현

[파일명] **admission\_pdf\_summarizer\_vFinal.py**

[목표]

1. 입시 요강 pdf 내용을 읽고 텍스트를 추출, 요약하여 요약\_pdf로 저장하기.

[코드의 기본 흐름]

1. 입시 요강 파일들이 있는 폴더 선택.
2. 파일 선택 및 텍스트 추출
   1. 표의 내용도 같이 추출
3. 추출된 텍스트를 요약
4. Pdf 파일 생성하여 요약 작성 및 저장

[코드 전략 수정]

1. 이전
   1. 추출한 텍스트를 한꺼번에 llm에 넘겨서 요약해달라고 함
2. 지금(개선)
   1. 섹션별로 적합한 문서를 검색하도록 함 (섹션 당 문서 7개, 유사도 0.75 이상)
   2. 중요 항목(섹션)별로 요약을 해서 그 섹션별 요약을 pdf문서로 만듦

[청크 사이즈 감소]

1. 전략을 수정함에 따라 청크 사이즈가 클 필요 없어짐
   1. 청크 사이즈 = 1100
   2. Overlap = 300

[retriever 개선]

1. 중요 항목(모집 인원, 수능 최저, 등록금 등)별로 문서를 검색함
   1. 검색 문서 수: 7
   2. 유사도 >= 0.75 이상
2. 쿼리 개선
   1. # 1. 섹션별 검색 쿼리 정의  
      self.section\_queries = {  
       "입학자격": [  
       "입학 지원 자격과 학력 조건은?",  
       "재직자 특별전형 지원 요건은?",  
       "외국인 특별전형 자격 기준은?",  
       "성인학습자 입학 자격은?",  
       "편입학 지원 자격 조건은?"  
       ],  
       "모집정원": [  
       "학과별 모집정원은 몇 명인가?",  
       "수시 정시 전형별 모집인원은?",  
       "정원내 정원외 모집 구분과 인원은?",  
       "특별전형별 선발 인원은?",  
       "총 모집정원과 학부별 배정 인원은?"  
       ],

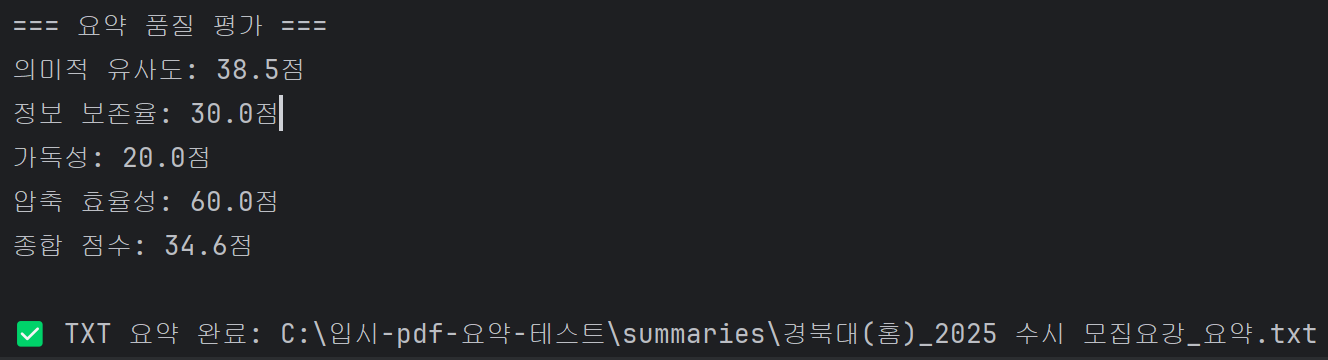
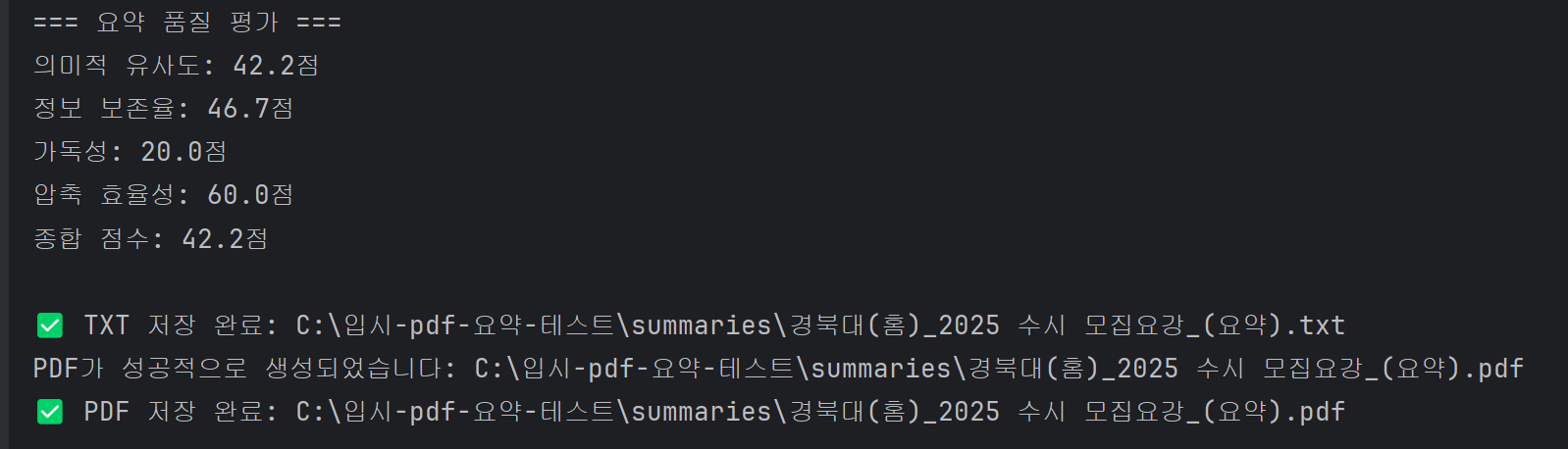
[프롬프트 개선]

1. # 2. 섹션별 상세 프롬프트 템플릿  
self.section\_prompts = {  
 "입학자격": """  
 다음 입시 요강 정보를 바탕으로 입학자격 정보를 정확하고 체계적으로 요약해주세요.  
  
 \*\*요약 기준:\*\*  
 - 일반전형, 특별전형별 학력 조건  
 - 재직자/외국인 등 특별 요건  
 - 편입학 자격 조건  
 - 지원 제한 사항  
  
 \*\*관련 정보:\*\*  
 {context}  
  
 \*\*입학자격 요약:\*\*  
 """,  
  
 "모집정원": """  
 다음 입시 요강 정보를 바탕으로 모집정원 정보를 정확하고 체계적으로 요약해주세요.  
  
 \*\*요약 기준:\*\*  
 - 계열별 모집인원  
 - 전형별 세부 인원  
 - 정원 내/외 구분  
 - 학과별 정확한 모집인원  
  
 \*\*관련 정보:\*\*  
 {context}  
  
 \*\*모집정원 요약:\*\*  
 """,

[그외 추가 및 개선사항]

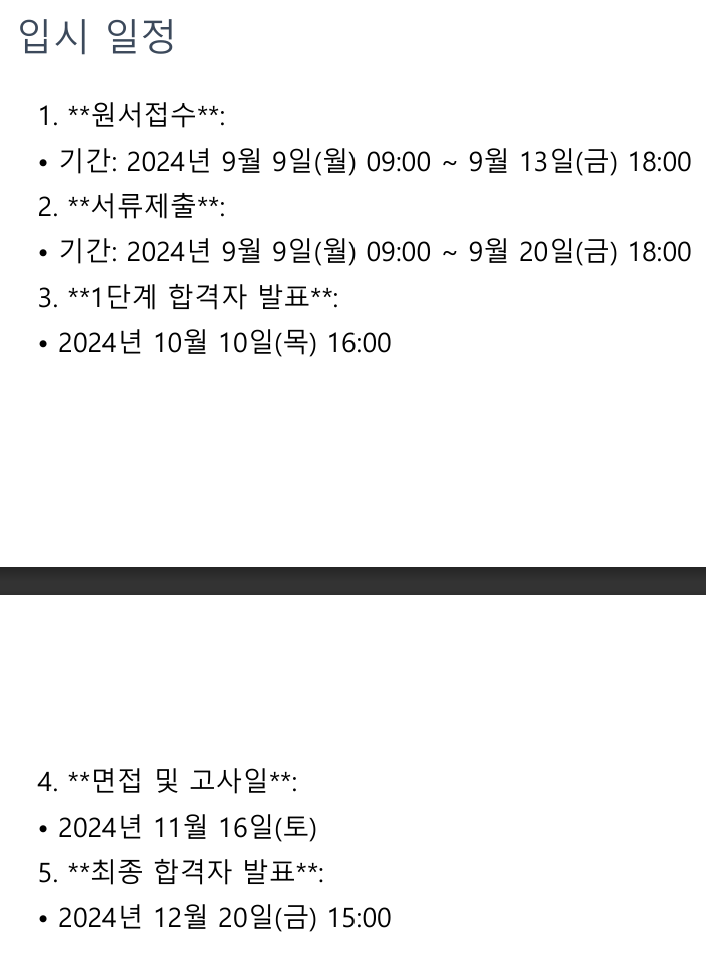
1. 평가지표 계산하는 코드 추가
   1. 의미적 유사도: TF-IDF 코사인 유사도
   2. 핵심 정보 보존율: TF-IDF 기반
   3. 가독성
      1. 평균 문장 길이: 적정 15~25자
      2. 복잡도: 접속사 사용률
   4. 압축 효율성
   5. weights = [0.25, 0.40, 0.20, 0.15] # 의미유사도, 정보보존, 가독성, 압축효율
2. 레이아웃을 섹션에 맞게끔 수정

[평가비교]

1. 이전:
   1. 
2. 지금(개선):
3. 

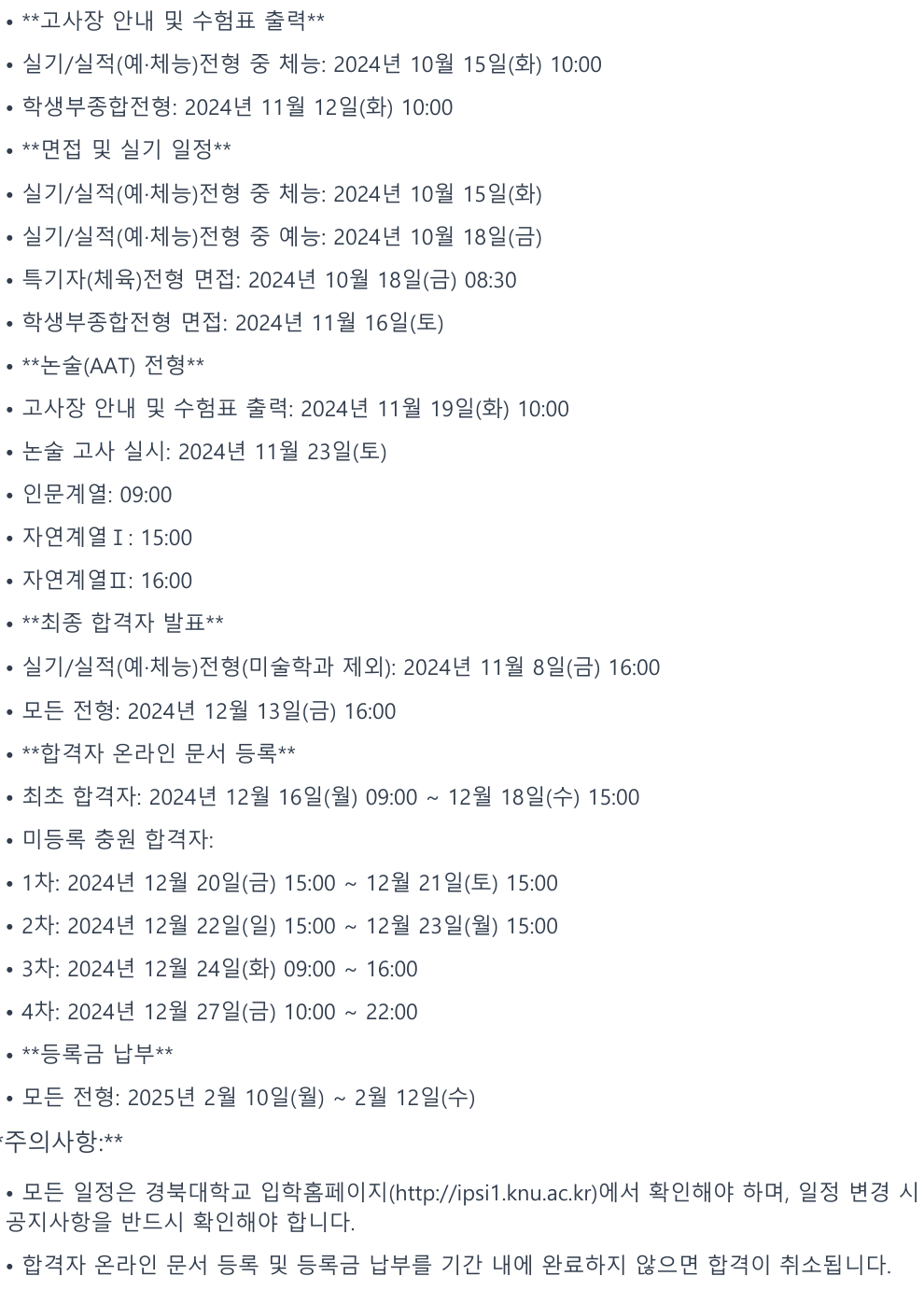
[실제 결과]

1. 이전



1. 지금(개선):





[문제점]

1. 이전 버전과 지금 버전의 평가지표 체감과 다르게 큰 차이 없이 나왔다

[코드 링크]

<https://github.com/reko1999/summarize_admission_pdfs_codes/blob/main/admission_pdf_summarizer_vFinal.py>