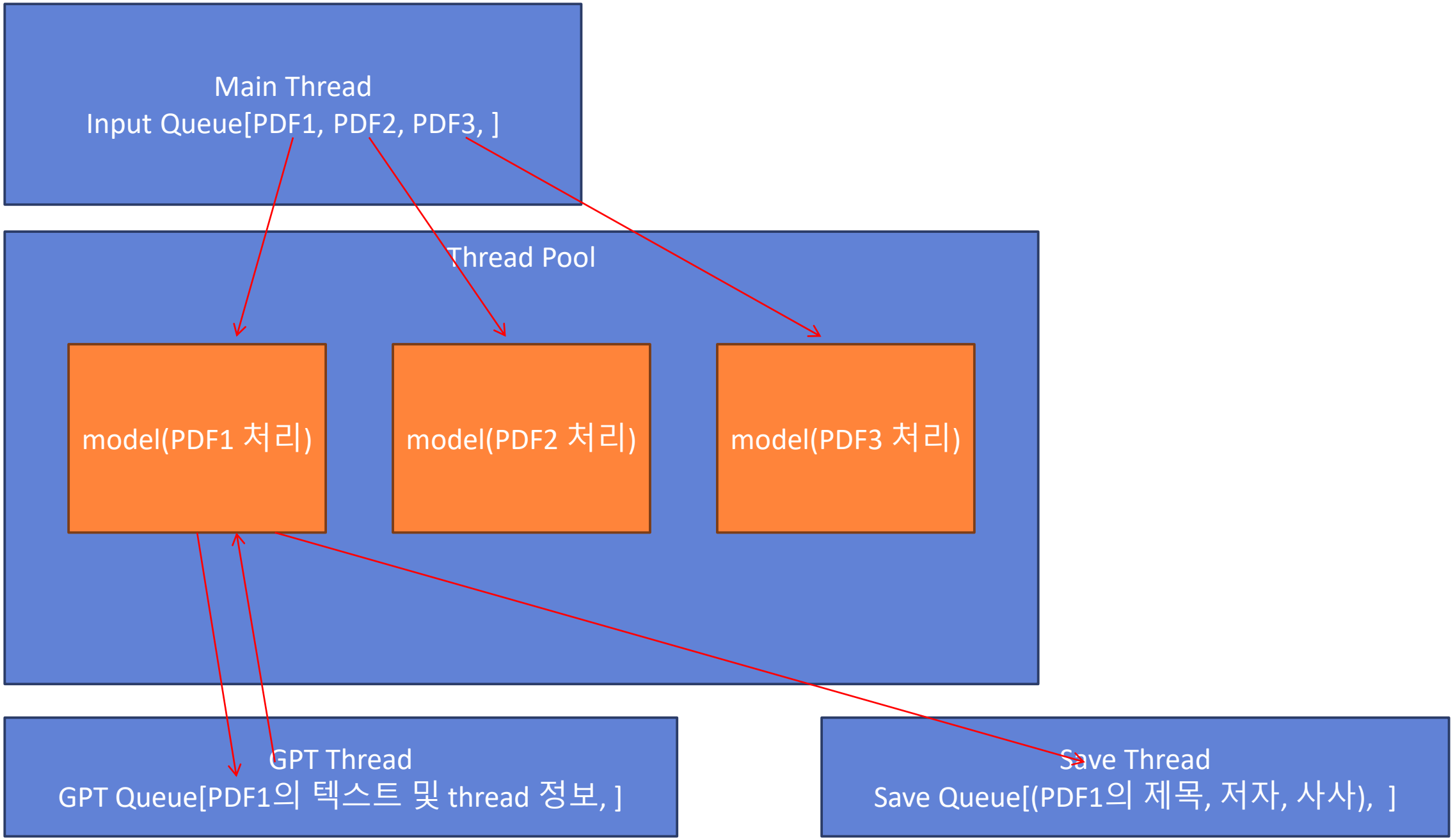
Main Thread
Input Queue[PDF1, PDF2, PDF3,]
Thread Pool
model(PDF1 처리)
model(PDF2 처리)
model(PDF3 처리)
GPT Thread
GPT Queue[PDF1의 텍스트 및 thread 정보,]
Save Thread
Save Queue[(PDF1의 제목, 저자, 사사),]

전체 동작

1. Input Queue에 PDF1이 들어오면, thread pool에게 thread 할당 요청
2. Thread Pool에서 할당된 thread는 각 PDF의 텍스트 추출
3. 추출한 텍스트 및 thread 정보를 GPT Queue에 전달
4. GPT Thread는 GPT Queue의 데이터 가공(질문 및 반환)
5. 요청한 thread에게 가공된 데이터 전달
6. 분류 성공 시, Save Queue에 데이터 넣고 반환
7. 분류 실패 시, GPT Queue에 나머지 페이지(분량 조절 가능) 재전송
8. Save Thread는 Save Queue의 데이터를 excel에 저장

모델 기능

1. Input Queue에 있는 PDF 가져오기
2. PDF 첫 장과, 마지막(2장)의 텍스트 추출
3. 추출 실패 시, Image 텍스트 추출 방식으로 2번 진행
4. GPT Queue에 추출한 텍스트 전달
5. GPT Thread의 반환 대기
6. 분류 성공 시 Save Queue에 분류 데이터 넣고 return(Thread 해제)
7. 분류 실패 시, 나머지 PDF Page의 텍스트를 추출하여 GPT Queue에 다시 전달

1. 제목, 저자는 가장 첫 장에 존재하기 때문에 첫 장의 텍스트를 추출
2. 사사의 경우 마지막 장(혹은 그 앞장)에 존재한다고 가정하고 추출
3. 2번의 가설이 틀렸을 경우 남은 장들의 내용을 일정 단위로 쪼개서 다시금 사사 추출

장점: 2번 가설이 맞을 경우에는 아주 적은 텍스트를 바탕으로 제목, 저자, 사사를 추출할 수 있음

단점: 2번 가설이 틀릴 경우 여러번에 거쳐 사사를 탐색해야 함