캡스톤디자인 I 계획서

<table>
<tr><td rowspan="2">제 목</td><td>국문</td><td>문서 내 표 추출 및 변환 서비스</td></tr>
<tr><td>영문</td><td>Table extraction and conversion service within document</td></tr>
<tr><td>프로젝트 목표
(500자
내외)</td><td colspan="2">인터넷 상에서 문서를 다운로드 하거나 전달할 때 가장 많이 사용되는 방법을 찾으면 PDF 문서를 빼놓을 수 없다. 하지만 PDF 문서의 특성상 수정이 어려워 문서내 특정 부분만 필요로 하거나 다른 포맷의 문서로 추출하려는 경우 별도의 프로그램을 사용하거나 직접 입력하는 방법을 선택하게 된다. 이 프로젝트에서는 PDF 문서 내 다양한 자료 중 표 데이터를 엑셀로 변환하는 서비스를 목표로 하려고 한다. 또한 웹사이트 형태로 배포하여 다양한 플랫폼에서 사용할 수 있도록 하고, 비슷한 서비스들과 비교했을 때 더 나은 성능과 정확도를 제공하고자 한다.</td></tr>
<tr><td>프로젝트
내용</td><td colspan="2">

1. 캡스톤디자인의 배경 및 필요성

현대 사회에서 문서 작업은 대부분 디지털 문서 작성을 주로 이루고 있다. 디지털 문서에는 다양한 포맷이 존재하지만 PDF는 많은 업무나 인터넷 상에서 사용될 뿐만 아니라 국제 표준으로써 자리매김하고 있다.

장점	단점
강력한 크로스 플랫폼	편집하기 어려움
파일 무결성	텍스트 처리에 적합하지 않음
파일 압축	모든 유형의 문서에 적합하지 않음
좋은 조판 효과	

위 표의 단점을 좀 더 자세하게 설명하면 PDF의 열람에는 별도의 소프트웨어 없이 웹 브라우저를 사용할 수 있지만 편집이 필요한 경우 별도의 소프트웨어가 필요해진다. 또한 PDF 문서 내의 텍스트를 복사, 붙여넣기 또는 검색하는 작업을 위해서 OCR 기술을 사용 하여 텍스트를 변환해야 한다. 이러한 단점들로 인해 단순 작업 과정 중에서의 여러 문제가 표출되고 있다. 예를 들어 관세사 업무는 30%가 수출입 선적서류의 단순 입력 작업으로 업무에 피로도가 증가하고 있으며 오기재 등 잘못된 문서 처리로 인해 손해가 발생하여 업무에 가해지는 스트레스가 높다고 한다. 기존 OCR 기술의 경우 텍스트를 잘못 인식하기도 하고 복잡한 형태의 문장이나 표를 인식할 수 없다. 이러한 문제를 해결하기 위해 딥러닝을 사용한 OCR 기술들이 연구되고 있다. 이런 딥러닝 기술을 통해 PDF 문서 내 표를 엑셀로 추출하는 과정을 자동화하여 업무의 속도와 피로도가 개선하는 시스템을 개발하고 기존 모델들의 단점인 병합된 셀이나 외곽선이 없는 표를 정확하게 인식하지 못하는 문제를 개선하는 방법을 탐색하고자 한다.

Disability Category	Participants	Ballots Completed	Ballots Incomplete/ Terminated	Results	
				Accuracy	Time to complete
Blind	5	1	4	34.5%, n=1	1199 sec, n=1
Low Vision	5	2	3	98.3% n=2 (97.7%, n=3)	1716 sec, n=3 (1934 sec, n=2)
Dexterity	5	4	1	98.3%, n=4	1672.1 sec, n=4
Mobility	3	3	0	95.4%, n=3	1416 sec, n=3

Disability		Ballots	Ballots		Results
Category	Participant	Completed	Incomplet	Accuracy	Time tocomplete
Blind	5	1	4	34.5 %, n = 1	1199 sec, n = 1
Low Vision	5	2	3	98.3 % n = 2(97.7 %, n =	1716 sec, n = 3(1934 sec, n = 2)
Dexterity	5	4	1	98.3 %, n = 4	1672.1 sec, n = 4
Mobility	3	3	0	95.4 %, n = 3	1416 sec, n = 3

위 그림은 기존 시스템을 통해서 표를 엑셀로 변환한 예시로 병합된 셀이 분리된 셀로 변환된 것을 확인할 수 있다.

</td></tr>
</table>

2. 캡스톤디자인 목표 및 비전

PDF 문서는 디지털 문서의 표준으로써 자리 잡고 있다. 업무 자동화의 관점에서 이러한 PDF 문서를 처리하기 위해서는 문자 인식 기술, 즉 OCR이 중요한 기술로써 활용된다. 하지만 단순 문장 구조의 데이터가 아닌 표 형식의 데이터의 경우 단순 OCR로 처리하기 힘들 수가 있다. 이를 해결하기 위해 인공신경망을 이용한 영상 인식 기술이 사용되었고 이는 OCR의 인식 정확도를 크게 향상했다. 하지만 여전히 남아있는 문제점으로는

1) 같은 PDF 문서 내에 존재하는 표들의 구조, 데이터 타입, 문자의 위치가 일률적이지 않다.

2) 외곽선이 그려지지 않은 표나 병합된 셀이 존재하는 경우가 있다.

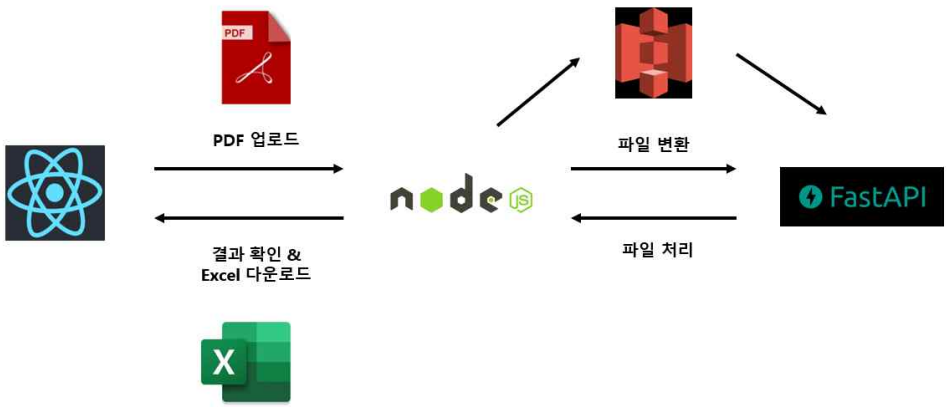
3) 표의 전체 내용이 아닌 일부 내용만 필요하거나 그래프와 같은 다른 형태의 표현 방법이 요구되는 경우가 있다.

따라서 이러한 세 가지 문제점의 해결과 개선을 목표로 하고 이를 웹으로 배포하여 실제 업무 자동화에 도움이 되는 서비스를 개발하려고 한다.

3. 캡스톤디자인 내용

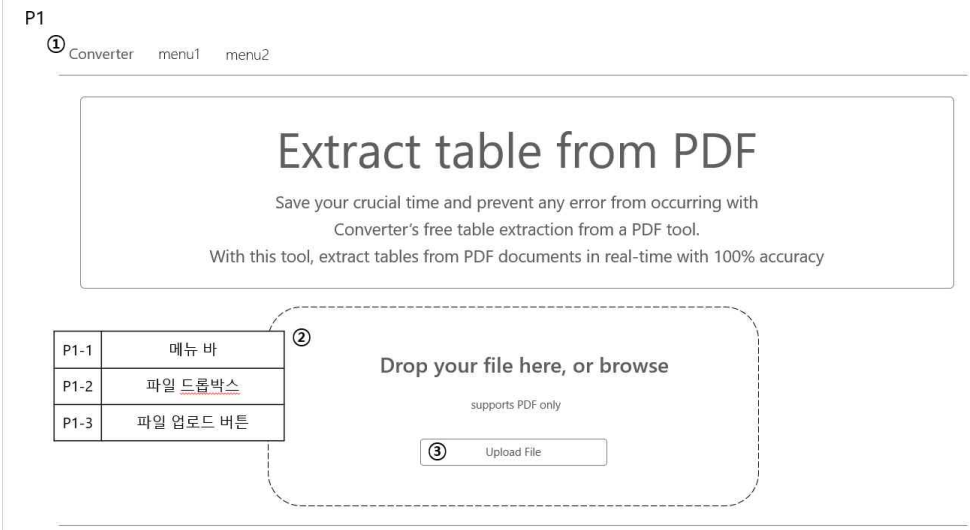
1) 캡스톤디자인의 범위

프론트엔드는 React.js로 구성하여 동적인 웹 페이지를 개발하고자 한다. 메인 백엔드 서버는 Node.js로 구성하여 업로드된 PDF를 모델에서 사용하기 위해 이미지 파일 형식으로 처리하는 기능과 모델에서 처리된 데이터를 프론트엔드로 넘겨 사용자가 결과를 확인하고 엑셀로 다운로드 할 수 있는 기능을 포함한다. 모델은 메인 백엔드 서버와 분리하여 FastAPI로 별도 모델 서버를 구성하여 모델의 오류로 전체 웹 서비스가 마비되지 않도록 예방하고 기능을 분리하여 개발과 배포의 편의성을 높이려고 한다. 모델 서버에서는 페이지 내에 있는 표의 위치를 인식하고 표 내의 텍스트를 추출하여 메인 백엔드 서버로 전달한다. 또한 백엔드 서버간의 이미지 파일을 AWS S3에 업로드하도록 분리하였다.

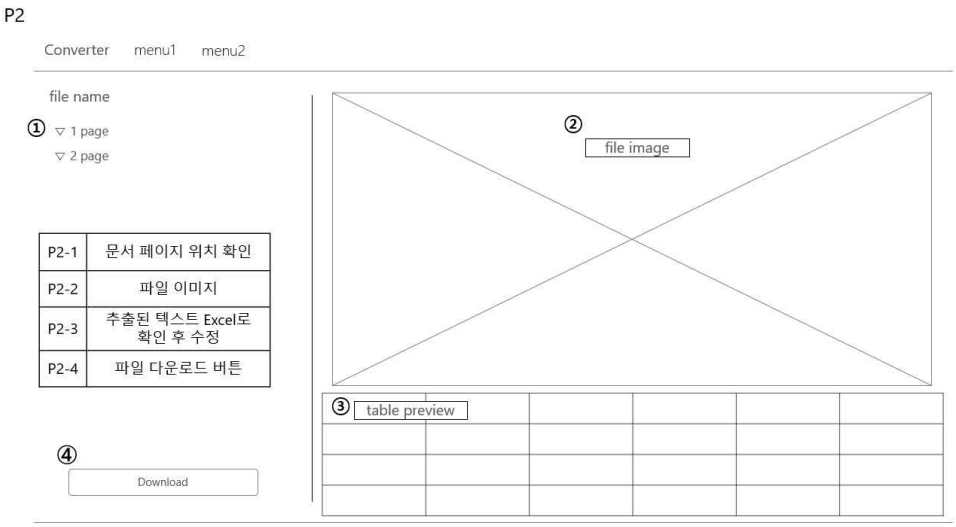


2) 주요 기능

처음 접속할 화면의 구성을 P1 그림에서 나타냈다. 사용자가 엑셀로 추출할 PDF 문서를 드롭박스에 드래그 앤 드롭하거나 파일 업로드 버튼을 클릭하여 파일관리자 창을 통해 직접 업로드 할 수 있는 기능을 제공한다.



P2 그림은 업로드된 PDF 문서내 표가 추출 된 후 결과를 보여주는 화면이다. 문서에 여러 페이지가 있을 경우 각 페이지별로 토글 메뉴가 존재하고 메뉴를 클릭하면 페이지 내의 표를 각각 확인할 수 있도록 서브메뉴를 구성한다. 원본 표를 P2-2에서 확인하고 만약 잘못 추출된 텍스트가 있을 경우 P2-3에서 수정할 수 있는 기능을 지원한다. 이후 다운로드 버튼을 통해 사용자가 원하는 표를 엑셀로 추출할 수 있다.



3) 비 기능적 요구사항

1. PDF 문서 내의 표를 엑셀로 변환시키는데 까지 걸리는 시간을 최소화한다.
2. 표를 인식하고 엑셀로 변환하는 모델들을 비교하고 성능을 개선하여 정확성을 높인다.
3. 사용자가 업로드한 파일명을 암호화하여 개인정보의 유출을 최소화한다.
4. AWS 클라우드에 업로드 된 파일은 관리자 계정 외에 열람할 수 없도록 설정하여 기밀성을 유지한다.
5. Git을 통한 형상관리와 디자인 패턴의 적용으로 유지보수 및 변경이 쉽도록 돕는다.

6. 코드에 충분한 주석을 추가하고, 시스템 아키텍처 및 기능에 대한 문서를 작성하여 변경사항을 반영할 때 참고 자료로 사용한다.

4. 캡스톤디자인 추진전략 및 방법

1) 캡스톤디자인에 대한 이해

1. 기존 모델들을 비교하여 프로젝트에 적합한 모델을 선택한다.
2. 기존 모델의 하이퍼 파라미터 또는 레이어 수정을 통해 성능개선을 시도한다.
3. 한정된 컴퓨팅 자원을 사용자들이 많이 사용되는 시간대를 분석하여 분배한다.

2) 캡스톤디자인 경험

저학년에 필수적으로 배웠던 객체지향언어와 소프트웨어 설계 이론을 기반으로 웹 서비스를 구현하고, 인공지능 지식을 활용하여 프로젝트의 주요 부분인 모델을 생성할 때 도움이 될 것이다. 또한 그동안의 교양 및 전공에서 진행했었던 팀/개인 프로젝트 경험을 보고서 작성, ppt, 발표 등의 능력을 프로젝트에 녹여낼 수 있을 것이다. 이러한 팀원들의 경험을 토대로 캡스톤 디자인 프로젝트에서도 유용한 기술과 노하우를 발휘하여 프로젝트를 수행할 예정이다.

3) 검증된 멘토 활용

1. 한정된 기간내에 프로젝트에서 구현하고자 하는 범위와 일정 조절에 도움을 받는다.
2. 변환 모델의 정확도 개선 방향 조언을 구한다.
3. 현재 기존 서비스를 이용하는 사용자들의 니즈 파악에 도움을 받는다.

4) 프로젝트 관리체계 수립

1) 역할 분담

- 이영호: 팀 컨버터의 조장 역할, 웹 서비스의 전반적인 서버 기능 구현
- 이태윤: 문서 내 표 추출 및 텍스트 변환 모델을 생성하는 비즈니스 로직 구현
- 박민이: 사용자에게 보여지는 웹 페이지의 UI, UX 서비스 구현

2) 분업 및 개발과정 공유

온라인 서비스 도구들을 활용하여 프로젝트 과정이 원활하게 진행되도록 할 계획이다. 팀원들과 진행되고 있는 모든 프로젝트 내용을 공유하기 위해 문서 작성을 위한 노션과 소스 코드들을 효과적으로 관리하기 위해 Git을 선택했다.

프로젝트 수행물에 대한 강제성을 부여하기 위해 주 1회 회의를 통해 그동한 했던 작업물 및 해결하기 어려운 점 등을 공유하는 시간을 가질 예정이다. 비대면으로 회의를 진행할 경우 Discord와 zoom을 활용하고, 대면으로 진행할 경우 학교 도서관의 스터디룸 및 미디어실을 예약하여 사용할 예정이다. 이러한 과정으로 우리팀은 캡스톤의 방향성을 올바르게 확립해나갈 것이다.

5. 참고문헌

1) 이동석, 권순각.(2021).딥러닝을 통한 문서 내 표 항목 분류 및 인식 방법.멀티미디어학회논문지,24(5),651-658.

중심어(국문)	웹 서비스	광학식 문자 인식	딥러닝	문서 변환

Keywords (english)	web service	OCR	deep learning	convert document

멘토	소속	튜터러스랩스	이름	경민영 (연구원)

팀 구성원	학년/반	학 번	이 름	연락처(전화번호/이메일)
	4	20197128	이영호	010-8313-3747
	4	20217134	박민이	010-9672-1996
	4	20181597	이태윤	010-8486-5684

컴퓨터공학과의 캡스톤디자인 관리규정과 모든 지시사항을 준수하면서 본 캡스톤디자인을 성실히 수행하고자 아래와 같이 계획서를 제출합니다.

2024 년 3 월 8 일

책 임 자 : 이영호 (인)

희망 지도교수 : 이현빈 교수님(인)