설문조사 등 지류조사 to 데이터

국립한밭대학교 리빙랩 3팀 운혜 (육종범, 옥주용, 채성수, 이동현)

연구배경

주민센터, 구청, 시청 같은 행정기관에서 주민 분들이 작성하는 각종 신청서, 신고서, 민원서류 설문조사 등의 지류 문서들을 창구에서 서비스하는 공무원들이 수동으로 처리하게 되어, 인력, 시간 낭비가 이루어지고 있으며, 일의 효율성이 낮아지게 된다. 또한, 수동 처리 방식은 그림1-3과 같이 인적 오류의 발생 가능성을 높이며, 문서 처리 도중 실수나 착오가 발생하게 된다. 이러한 오류는 주민들의 불만을 초래하며, 행정 서비스에 대한 신뢰도를 떨어뜨린다.

위와 같은 이유로, 공무원들의 행정력이 낭비되고, 주민들의 서비스 품질이 낮아질 수 있으며, 해당 문제를 해결하기 위해, 보다 정확하고 신속한 처리에 대한 방법을 고민하게 되었다.

문제 해결을 위해, OCR(Optical Character Recognition, 광학 문자 인식)을 이용한 데이터 스캔, 데이터 추출을 채택하였으며, 해당 기술을 이용하여 주민들이 지류 문서에 작성한 자필 내용을 AI의 딥러닝 기술을 통해 문자열 데이터를 얻어 문자열 데이터를 정리하여, 엑셀로 옮겨 자동화하는 로직을 고안하게 되었다.

그림 1-1 설문조사 대상

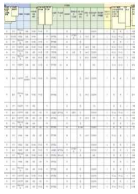


그림 1-2 인력 낭비

그림 1-3 인적 오류

AI기술을 도입하므로써, 담당자의 업무 부담이 축소되며 자필 지류들이 모두 데이터화 되므로 지류 데이터 축적이 용이하게 된다.

또한, 해당 프로그램을 개발하게 되어 온라인 사용이 어려운 장애인, 노인, 저소득층 등의 다양한 취약계층을 담당하는 부서에서 문서 처리를 자동화하여, 이들에 대한 서비스 제공 속도와 정확도를 높일 수 있다.

이는 취약계층이 신속하고 적절한 지원을 받을 수 있게 하여 사회적 약자를 보호하고 지원하는 역할을 강화하는 기대효과를 예상할 수 있다.

연구결과

OCR을 통해 문자를 성공적으로 인식할 수 있는 경우도 있지만, 이미지 전체를 OCR 할 때의 탐색 방식으로 인해 몇 가지 문제점들이 발생한다. 이 탐색 방식은 X축을 기준으로 시작과 끝을 탐색 후, Y축으로 이동하고 다시 X축 탐색 수행하는 방식이다. 이러한 탐색 방식 때문에 다음과 같은 문제점이 발생하게 된다.

이슈 사항

1. 그림 3-1과 같이 스마트폰 사용 여부, 주로 시간을 보내는 장소 등의 데이터에 오류가 나타남.
2. 그림 3-2과 같이 어디에 표시되어 있는지 알 수 없는 문제가 발생함.
3. 그림 3-3과 같이 문자가 없는 경우 인식이 불가능한 문제가 발생함.
4. 악필과 같은 문자는 잘못 인식 하는 문제가 발생함

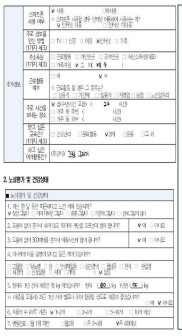


그림 3- 1. OCR 전/후(텍스트 오류)

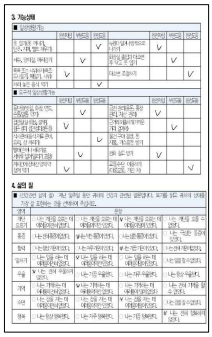


그림 3- 2. OCR 전/후(데이터 오류)

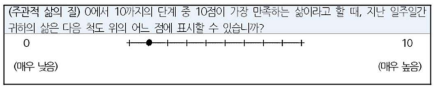
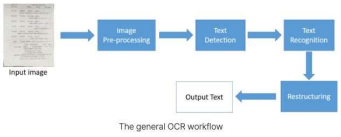


그림 3- 3.OCR 평가 이미지

연구방법



The general OCR workflow

그림 2-1 OCR 흐름도

설문조사의 지류 데이터를 받아오기 위한 여러 방법론을 제시하였다.
수기 작성한 설문지의 데이터를 로컬환경에 가져오는 방법에 대한 고민
작성된 설문지의 수기 작성 데이터를 가져오기 위한 방법론 제시

2-1. 영상처리를 이용한 수기 작성 정보 추출

2-2. 원본 설문지와 작성설문지를 비교하여 비교되는 부분의 추출

2-3. 좌표값을 설정하여 텍스트 데이터 추출

글씨체 탓에 깨지는 텍스트에 대한 대처
체크박스에 관련하여 텍스트와 체크표시의 정보 일치 여부 판단
데이터를 모두 추출하여 엑셀 양식에 맞춰 데이터 자동 등록

PDF 형식으로 설문지 데이터를 로컬 환경으로 받아와서 작성된 설문지를 Colab에서 Python을 통한 여러 라이브러리 활용하였고, OCR API를 활용해 텍스트를 추출한다.

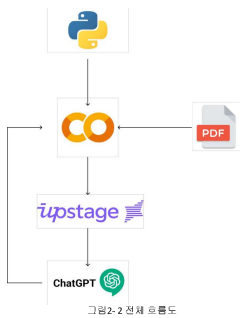


그림2- 2 전체 흐름도

여러 라이브러리와 API 등을 토대로 정확도(신뢰도)가 가장 높았던 API인 업스테이지를 채택했다.

이를 확인하고 Colab에서 Python을 통해 예시 데이터와 원본데이터를 토대로 진행 이후 AI Open API를 활용하여 추출된 데이터를 보다 효과적으로 관리하고, Excel에 저장하기 위해 텍스트 데이터를 넘겨 원하는 형식으로 데이터를 재정리하는 부분까지 완성하였으며, 해당 부분을 Excel로 옮기는 것까지 테스트를 진행했다.

이후 텍스트 추출의 예외처리적인 글씨체에 관련된 문제는 아직 방법을 고안 중이다.

결론

이전에는 수많은 설문조사, 각종 신청서, 신고서, 민원서류 설문조사 등의 데이터를 담당자 등이 직접 수기로 작성하여 데이터를 저장하였지만, 이러한 방식에는 인적오류가 발생하며, 잘못된 데이터 입력은 이후의 행정 처리 과정에서 추가적인 확인 작업과 수정 작업을 요구하게 되어, 시간과 자원이 낭비가 되었다. 또한, 수기로 데이터를 작성하고 저장하는 과정은 담당자들에게 막대한 업무 부담을 가중시켰으며, 업무 효율은 자연스럽게 저하가 되는 상황이었다.

그러나 최근 OCR 기술의 도입으로 인해 해당 기술을 사용하여, 자동으로 데이터를 추출하여 엑셀에 옮기는 프로그램을 만들어, 담당자들이 더 이상 모든 문서를 일일이 수기로 입력할 필요가 없어지고, 데이터 입력과정에서 효율성을 극대화 할 수 있게 된다. 이는 담당자들이 중요한 업무에 집중할 수 있게 해주고, 더 나아가, 데이터 검색이나 분석, 통계를 내는데에 더욱 편리해지므로, 축적된 데이터를 분석하여 주민들의 요구사항을 보다 정확하게 파악할 수 있는 순기능을 가지게 된다.

키워드

Python, Colab, Open API (Chat GPT, Upstage), Library (pandas)

정보 출처

[OCR/AI] 2023년 최신판 OCR 8가지 API 비교평가 테스트 (sk.com)

이미지 출처

https://velog.io/@rmlstia123/%ED%8C%8C%EC%9D%B4%EC%8D%AC-%EC%86%8C%EA%B0%9C
https://everydaywoogi.tistory.com/61
https://kor.pngtree.com/free-png-vectors/pdf
https://m.saramin.co.kr/job-search/company-info-view?csn=S2RCTlh2UkpuMlNVYWZXWU9hRmltZz09
https://chatgpt.com/