

# OCR을 이용한 서류 데이터 추출

---



Team.윤 혜 (육종범, 채성수, 옥주용, 이동현)

# 목차

---

1. 개발 배경
2. 개발 방법(시연및 착오)
3. 결과
4. 기대 효과
5. 향후 계획
6. 참고 문헌

# 개발 배경

---

## 효율성 문제

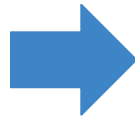
지류 문서에 대해서  
담당자들 직접  
수동처리

인력, 시간 낭비

## 인적오류

수동처리에 대해  
실수나 착오 발생

확인, 수정 시간 발생



공무원들의 행정력, 주민들의 서비스  
품질 향상 도모

정확하고 신속한 처리 가능

# 개발 방법 (시행착오)

보훈예우수당 지급신청서				처리기간 30일				
신청인	성명	육종범	지급대상유형					
	생년월일	2000.09.29 (남) 여						
	주소	대전 유성구 가나다 311						
	연락처	010-1234-5678						
예금계좌	예금주	금융기관명	계좌번호					
	홍길동	국민은행	1234-571234					
<p>「서울특별시 성동구 국가보훈대상자 예우 및 지원에 관한 조례」 제11조에 따라 보훈예우수당의 지급을 신청합니다.</p> <p>2011년 9월 22일</p> <p>신청인 육종범 (서명인)</p> <p>성동구청장 귀하</p> <table><tr><td>구비서류</td><td>수수료</td></tr><tr><td>1. 국가유공자증 또는 국가유공자 유족증. 고엽제 후유의증 환자는 지방보훈청장의 확인서 사본 1부. 2. 본인 통장사본 1부(계좌번호가 표시되어 있는 면을 말함).</td><td>없음</td></tr></table> <p>본인은 이 건 업무처리와 관련하여 「개인정보 보호법」 제15조제1항 및 「전자정부법」 제38조제1항에 따라 본인의 주민등록자료 등을 담당 공무원이 열람하는 것에 동의합니다.</p> <p>신청인: 육종범 (서명인)</p>					구비서류	수수료	1. 국가유공자증 또는 국가유공자 유족증. 고엽제 후유의증 환자는 지방보훈청장의 확인서 사본 1부. 2. 본인 통장사본 1부(계좌번호가 표시되어 있는 면을 말함).	없음
구비서류	수수료							
1. 국가유공자증 또는 국가유공자 유족증. 고엽제 후유의증 환자는 지방보훈청장의 확인서 사본 1부. 2. 본인 통장사본 1부(계좌번호가 표시되어 있는 면을 말함).	없음							

PyMuPDF  
:숫자 및 문자의 인식률이 낮다.

육종범  
2000  
. 09  
. 29 e  
대전  
유성구  
가나다  
311  
이-  
1 234  
- 5678  
홍길동  
국민은행  
1 234  
- 571234  
2011  
a  
22  
육종범  
m  
육종범  
DM

PyPDF2  
: 문자의 인식률이 매우 낮다.

₩2000.09.29₩e\_080xow31100-1234-5678₩jzj00₩|234-5712342011a22₩S₩m  
₩S₩DM

# 개발 방법 (시행착오)

보훈예우수당 지급신청서				처리기간 30일
신청인	성명	육종범		지급대상유형
	생년월일	2000.09.29 (남) 여		
	주소	대전 유성구 가나다 311		
	연락처	010-1234-5678		
예금계좌	예금주	금융기관명	계좌번호	
	홍길동	국민은행	1234-571234	
<p>「서울특별시 성동구 국가보훈대상자 예우 및 지원에 관한 조례」 제11조에 따라 보훈예우수당의 지급을 신청합니다.</p> <p>2011년 9월 22일</p> <p>신청인 육종범 (서명인)</p>				
성동구청장 귀하				
구비서류				수수료
1. 국가유공자증 또는 국가유공자 유족증. 고엽제 후유의증 환자는 지방보훈청장의 확인서 사본 1부. 2. 본인 통장사본 1부(계좌번호가 표시되어 있는 면을 말함).				없음
본인은 이 건 업무처리와 관련하여 「개인정보 보호법」 제15조제1항 및 「전자정부법」 제38조제1항에 따라 본인의 주민등록자료 등을 담당 공무원이 열람하는 것에 동의합니다.				
신청인: 육종범 (서명인)				

## pdfminer

: 숫자 및 문자의 인식률이 낮다.

육종범  
2000  
.  
09 . 29 e  
대전 유성구 가나다 311  
이 - 1 234 - 5678  
홍길동  
국민은행  
1 234 - 571234  
2011  
a  
22  
육종범 m  
육종범  
DM

## pdfplumber

: 숫자 및 문자의 인식률이 낮다.

육종범  
2000 09 29 e  
.  
대전 유성구 가나다 311  
이 - 1 234 - 5678  
홍길동 국민은행 1 234 571234  
-  
2011 a 22  
육종범 m  
DM  
육종범

# 개발 방법 (시행착오)

보훈예우수당 지급신청서				처리기간 30일
신청인	성명	육종범		지급대상유형
	생년월일	2000.09.29 (남) 여		
	주소	대전 유성구 가나다 311		
	연락처	010-1234-5678		
예금계좌	예금주	금융기관명	계좌번호	
	홍길동	국민은행	1234-571234	
<p>「서울특별시 성동구 국가보훈대상자 예우 및 지원에 관한 조례」 제11조에 따라 보훈예우수당의 지급을 신청합니다.</p> <p>2011년 9월 22일</p> <p>신청인 육종범 (서명인)</p> <p>성동구청장 귀하</p> <p>구비서류</p> <p>1. 국가유공자증 또는 국가유공자 유족증. 고엽제 후유의증 환자는 지방보훈청장의 확인서 사본 1부.</p> <p>2. 본인 통장사본 1부(계좌번호가 표시되어 있는 면을 말함).</p> <p>본인은 이 건 업무처리와 관련하여 「개인정보 보호법」 제15조제1항 및 「전자정부법」 제38조제1항에 따라 본인의 주민등록자료 등을 담당 공무원이 열람하는 것에 동의합니다.</p> <p>신청인: 육종범 (서명인)</p>				
구비서류				수수료
				없음

## tabula

: 테이블로 반환하는 장점은 있지만, 인식률이 떨어짐

	육종범	Unnamed: 0	Unnamed: 1
0	2000 09	NaN	NaN
1	.	NaN	NaN
2	.	29 e	NaN
3	대전 유성구 가나다 311	NaN	NaN
4	NaN	-	NaN
5	이- 1 234	5678	NaN
6	홍길동 국민은행	NaN	234 - 571234
7	2011 a 22	NaN	NaN
8	NaN	육종범 m	NaN
9	NaN	육종범	DM

## tika

: 숫자 및 문자의 인식률이 낮다.

육종범

2000

.

09

. 29 e

대전 유성구 가나다 311

이- 1 234 - 5678

홍길동 국민은행 | 234 - 571234

2011 a 22

육종범 m

육종범 DM

# 개발 방법 (시행착오)

보훈예우수당 지급신청서			처리기간	
			30일	
신청인	성명	육종범	지급대상유형	
	생년월일	2000.09.29 (남 여)		
	주소	대전 유성구 가나다 311		
	연락처	010-1234-5678		
예금계좌		예금주	금융기관명	계좌번호
		홍길동	국민은행	1234-571234
<p>「서울특별시 성동구 국가보훈대상자 예우 및 지원에 관한 조례」 제11조에 따라 보훈예우수당의 지급을 신청합니다.</p> <p>2011년 4월 22일</p> <p>신청인 육종범 (서명인)</p> <p>성동구청장 귀하</p>				

# 개발 방법 (시행착오)

보훈예우수당 지급신청서				처리기간 30일
신청인	성명	육종범		지급대상유형
	생년월일	2000.09.29 (남 여)		
	주소	대전 유성구 가나다 311		
	연락처	010-1234-5678		
예금계좌	예금주	금융기관명	계좌번호	
	홍길동	국민은행	1234-5671234	
<p>「서울특별시 성동구 국가보훈대상자 예우 및 지원에 관한 조례」 제11조에 따라 보훈예우수당의 지급을 신청합니다.</p> <p>2011년 9월 22일</p> <p>신청인 육종범 (서인)</p>				
성동구청장 귀하				
<p>구비서류</p> <p>1. 국가유공자증 또는 국가유공자 유족증. 고엽제 후유의증 환자는 지방보훈청장의 확인서 사본 1부.</p> <p>2. 본인 통장사본 1부(계좌번호가 표시되어 있는 면을 말함).</p>				<p>수수료</p> <p>없음</p>
<p>본인은 이 건 업무처리와 관련하여 「개인정보 보호법」 제15조제1항 및 「전자정부법」 제38조제1항에 따라 본인의 주민등록자료 등을 담당 공무원이 열람하는 것에 동의합니다.</p> <p>신청인: 육종범 (서인)</p>				

## Google Vision ocr

: 체크박스를 잘 못읽는다.

[별지 제1호서식] <개정 2014.7.10.>  
보훈예우수당 지급신청서  
처리기간  
30일  
신청인  
성명  
육종범  
지급대상유형  
생년월일  
2000.09.29(남 여)  
신청인 주소  
대전 유성구 가나다 311  
연락처  
010-12345678  
예금주  
금융기관명  
계좌번호  
예금계좌  
홍길동  
국민은행  
1234-5671234  
「서울특별시 성동구 국가보훈대상자 예우 및 지원에 관한 조례」 제11조에 따라 보훈예우수당의 지급을 신청합니다.  
성동구청장 귀하  
2011년 9월 22일  
신청인 육종범 (서인)  
구비서류  
1. 국가유공자증 또는 국가유공자 유족증.  
고엽제 후유의증 환자는 지방보훈청장의 확인서 사본 1부.  
2. 본인 통장사본 1부(계좌번호가 표시되어 있는 면을 말함).  
없음  
본인은 이 건 업무처리와 관련하여 「개인정보 보호법」 제15조제1항 및 「전자정부법」 제38조제1항에 따라 본인의 주민등록자료 등을 담당 공무원이 열람하는 것에 동의합니다.  
신청인: 육종범 (서인)



# 개발 방법 (시행착오)

추가정보	스마트폰 사용 여부	<input checked="" type="checkbox"/> 사용 <input type="checkbox"/> 미사용 ※ 스마트폰 사용할 경우 인터넷 이용하여 사용하는 지? <input checked="" type="checkbox"/> 인터넷 사용 <input type="checkbox"/> 인터넷 미사용
	주로 정보를 얻는 방법 (1가지 체크)	<input type="checkbox"/> TV <input type="checkbox"/> 신문 <input type="checkbox"/> 이웃 <input checked="" type="checkbox"/> 인터넷 <input type="checkbox"/> 가족
	주소록원 (1가지 체크)	<input type="checkbox"/> 근로활동 <input type="checkbox"/> 개인연금 <input type="checkbox"/> 공적연금 <input type="checkbox"/> 재산소득(임대료) <input type="checkbox"/> 가족지원 <input checked="" type="checkbox"/> 그 외(백수)
	근로활동 여부	<input type="checkbox"/> 예 <input checked="" type="checkbox"/> 부 ※ 근로활동 할 경우 그 종류는? <input type="checkbox"/> 상용직 <input type="checkbox"/> 기간제 <input type="checkbox"/> 일용직 <input type="checkbox"/> 자영업 <input type="checkbox"/> 농업 <input type="checkbox"/> 노인일자리
	주로 시간을 보내는 장소	<input checked="" type="checkbox"/> 감수면시간 포함 ( 24 시간) <input type="checkbox"/> 거주 동 주변 ( 시간) <input type="checkbox"/> 거주 동 외부 ( 시간)
	받고 싶은 교육은? (1가지 체크)	<input type="checkbox"/> 건강관리 <input type="checkbox"/> 문화활동 <input checked="" type="checkbox"/> 경제 <input type="checkbox"/> 운동 <input type="checkbox"/> 그 외
하고 싶은 여가활동은?	(주관사) 그림 그리기	

## 2. 노쇠평가 및 건강상태

■ 노쇠평가 및 건강상태	
1. 지난 한 달 동안 피곤하다고 느낀 적이 있습니까? <input checked="" type="checkbox"/> 항상 그렇다 <input type="checkbox"/> 거의 대부분 그렇다 <input type="checkbox"/> 종종 그렇다 <input type="checkbox"/> 가끔씩 그렇다 <input type="checkbox"/> 전혀 그렇지 않다	
2. 도움이 없어 혼자서 쉬지 않고 10개의 계단을 오르는데 힘이 듭니까?	<input checked="" type="checkbox"/> 예 <input type="checkbox"/> 아니요
3. 도움이 없어 300미터를 혼자서 이동하는데 힘이 듭니까?	<input checked="" type="checkbox"/> 예 <input type="checkbox"/> 아니요
4. 의사에게 다음 질병이 있다고 들은 적이 있습니까? <input type="checkbox"/> 고혈압 <input type="checkbox"/> 당뇨병 <input type="checkbox"/> 암 <input type="checkbox"/> 만성폐질환 <input type="checkbox"/> 심근경색 <input type="checkbox"/> 협심증 <input type="checkbox"/> 천식 <input type="checkbox"/> 관상염 <input type="checkbox"/> 뇌경색 <input type="checkbox"/> 신장질환 <input type="checkbox"/> 치매 <input type="checkbox"/> 기타( ) <input checked="" type="checkbox"/> 없음	
5. 현재와 1년 전의 체중은 몇 kg이었습니까? 현재 ( 60 ) kg 1년전 ( 75 ) kg ※ (체중을 모를시) 최근 1년 사이 벨트나 옷이 헐렁할 정도로 체중이 줄었습니까? <input type="checkbox"/> 예 <input checked="" type="checkbox"/> 아니요	
6. 복용약 수 (하루 기준) <input checked="" type="checkbox"/> 1~2개 <input type="checkbox"/> 3~4개 <input type="checkbox"/> 5~10개 <input type="checkbox"/> 10개 이상	
7. 병원진료 <input type="checkbox"/> 월 1회 미만 <input type="checkbox"/> 월 2회 <input type="checkbox"/> 주 3~4회 <input checked="" type="checkbox"/> 주 4회 이상	

## Naver Clova ocr

: 체크박스가 가끔 누락됨

☒ 사용 ☐ 미사용  
스마트폰 사용 여부  
스마트폰 사용할 경우 인터넷 이용하여 사용하는 지?  
☒ 인터넷 사용 ☐ 인터넷 미사용  
주로 정보를 얻는 방법  
☐ TV ☐ 신문 ☒ 이웃 ☒ 인터넷  
주소록원  
☐ 근로활동 ☐ 개인연금 ☐ 공적연금 ☐ 재산소득(임대료)  
(1가지 체크)  
☐ 가족지원 ☒ 그 외(백수)  
근로활동 여부  
☐ 예 ☒ 부  
※ 근로활동 할 경우 그 종류는?  
☐ 상용직 ☐ 기간제 ☐ 일용직 ☐ 자영업 ☐ 농업 ☐ 노인일자리  
주로 시간을 보내는 장소  
☒ 감수면시간 포함 ( 24 시간)  
☐ 거주 동 주변 ( 시간)  
☐ 거주 동 외부 ( 시간)  
받고 싶은 교육은?  
(1가지 체크)  
☐ 건강관리 ☐ 문화활동 ☒ 경제 ☐ 운동 ☐ 그 외  
하고 싶은 여가활동은?  
(주관사) 그림 그리기

체크박스가 나와야 하는데 누락

	<p>스마트폰 사용 여부</p> <p><input checked="" type="checkbox"/> 사용 <input type="checkbox"/> 미사용</p> <p># 인터넷 사용할 경우 인터넷 이용에 사용하는 기기</p> <p><input checked="" type="checkbox"/> 인터넷 사용 <input type="checkbox"/> 인터넷 미사용</p>
	<p>주요 정보를 얻는 방법</p> <p><input type="checkbox"/> TV <input type="checkbox"/> 신문 <input type="checkbox"/> 이웃 <input checked="" type="checkbox"/> 인터넷 <input type="checkbox"/> 가족</p>
	<p>주요소통</p> <p><input type="checkbox"/> 전화 <input type="checkbox"/> 카톡지침 <input type="checkbox"/> 이메일 <input type="checkbox"/> 음성메시지</p> <p><input type="checkbox"/> 문자 <input type="checkbox"/> SNS <input type="checkbox"/> 외 <input type="checkbox"/> 기타</p>
유기정보	<p>근로활동 여부</p> <p><input type="checkbox"/> 여 <input type="checkbox"/> 부</p> <p># 근무할 경우 어떤 곳 종사하는가?</p> <p><input type="checkbox"/> 상용직 <input type="checkbox"/> 기간제 <input type="checkbox"/> 일용직 <input type="checkbox"/> 계약직 <input type="checkbox"/> 농업 <input type="checkbox"/> 노인요양직리</p>
	<p>주요 시간강을 받든 강소</p> <p><input checked="" type="checkbox"/> 강소(연간강 포함) <input checked="" type="checkbox"/> 시간</p> <p><input type="checkbox"/> 거주주 주변 <input type="checkbox"/> 시간</p> <p><input type="checkbox"/> 거주주 외부 <input type="checkbox"/> 시간</p>
	<p>평균 일련교은?</p> <p><input type="checkbox"/> 12시간 이하 <input type="checkbox"/> 12시간 이상</p>
	<p>하루 평균 야간활동은?</p> <p><input type="checkbox"/> 주간 <input checked="" type="checkbox"/> 야간</p>

## 2 노쇠평가 및 건강상태

■ 보시방 및 길찾기	
1. 가는 한도 동안 기관차와 노면 접촉이 있었나?	
✓ <input type="checkbox"/> 없었음 <input type="checkbox"/> 제1면 접촉 <input type="checkbox"/> 종동 접촉 <input type="checkbox"/> 2면 접촉 <input type="checkbox"/> 전면 접촉	
2. 노면이 얼어 있거나 휘어진지 10cm² 이상 얼음으로 덮여 있는지 묻나?	✓ 예 <input type="checkbox"/> 아니오
3. 도랑이 없어 300mm를 초과치 아물때에 해당 묻나?	✓ 예 <input type="checkbox"/> 아니오
4. 마개재가 눈에 걸렸어 있고 그로 인해서 아물때가?	
<input type="checkbox"/> 그렇고 <input type="checkbox"/> 없음 <input type="checkbox"/> 눈이 아물때에 해당 <input type="checkbox"/> 심하게 아물때 <input type="checkbox"/> 아물때 <input type="checkbox"/> 천천히 <input type="checkbox"/> 결빙 <input type="checkbox"/> 비가 <input type="checkbox"/> 눈이 아물때에 해당 <input type="checkbox"/> <input type="checkbox"/> 없음	
5. 한해 1년 전의 제빙은 몇 kg이었나? 한해 ( <u>60</u> ) kg 미만 ( <u>10</u> ) kg	
※ 아물때를 다루는 한해 1년 전의 제빙나 호의 아물때 정도를 제빙이 있었나?	
	✓ 예 <input type="checkbox"/> 아니오
6. 폭우가 수위(하)만 <input type="checkbox"/> 1~2기 <input type="checkbox"/> 3~4기 <input type="checkbox"/> 5~10기 <input type="checkbox"/> 10기 이상	
7. 병력(보) 일회 미관 <input type="checkbox"/> 일회 <input type="checkbox"/> 2회 3회 <input type="checkbox"/> 5회 이상	

사를 ☐ 미사용  
 스마트폰  
 ✕ 스마트폰 사용할 경우 인터넷 이용하여 사용하는 지?  
 사용 여부  
 √ 인터넷 사용 ☐ 인터넷 미사용  
 주로 정보를 얻는 방법 ☐ TV ☐ 신문 ☐ 이웃 √인터넷 ☐ 가족  
 (가치 체크)  
 주소록만 ☐ 근로활동 ☐ 개인연금 ☐ 공적연금 ☐ 재산소득(임대료)  
☐ 주식 ☐ 가족지원 √ 외 (백 수)  
☐ 며 √부  
 근로활동  
 거주지역  
 여부 ✕ 근로활동 할 경우 그 종류는?  
☐ 실종직 ☐ 기간제 ☐ 일용직 ☐ 자영업 ☐ 농업 ☐ 노인일자리  
☐ 전(주말시간 포함) ( 24 시간)  
 주를 시간을  
☐ 거주 중 주변 ( 시간)  
 보내는 장소  
☐ 거주 중 외부 ( 시간)  
 받고 싶은  
 교육은? ☐ 건강관리 ☐ 문화활동 √경제 ☐ 운동 등 ☐ 외  
 (가치 체크)  
 하고 싶은  
 (주관식) 그림 그리기  
 여가활동은?  
☐ 노숙생활 및 건달살대  
☒ 노숙생활 외 건달살대  
 1. 지난 한 달 동안 교전했다고 느낀 적이 있습니까?  
 √ 항상 그렇다 ☐ 거의 대부분 그렇다 ☐ 종종 그렇다 ☐ 가끔씩 그렇  
 2. 도둑이 없애 혼자서 처지 말고 10개의 계단을 오를데 힘이 든다  
 3. 도둑이 없애 900미터를 혼자서 이동하는데 힘이 든다? √ 예 ☐  
 4. (가치 체크) 건강에 약하고 혹은 술이 많습니까?  
☐ 건강에 ☐ 신장질환 ☐ 치매 ☐ 머릿기( ) √ 위염  
 5. 현재와 1년 전의 체중은 몇 kg이었습니까? 현재 ( 80 ) kg 1년전  
 ✕ (체중을 모르신다) 최근 1년 사이 발트나 듯이 혈당할 정도로 체중이  
 늘었어 아니요  
 6. 복통약 수 (하루 기준) √ 1-2개 ☐ 3-4개 ☐ 5-10개 ☐ 10개 이상  
 7. 병원에서 ☐ 소화 1회 ☐ 미만 ☐ 대변 2회 ☐ 주 3-4회 ☐ 주 4회 이상

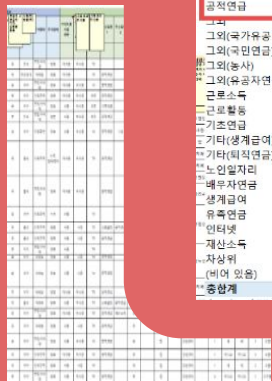
## 추출된 데이터 후보정

## ChatGPT



2024  
AAiCON

# 기대 효과



행 레이블 - 개수 : 주소득원1

개인연금	56
개인연금	1
공적연금	984
공적연금	1
오타	
그외(국가유공자)	21
그외(국민연금)	1
그외(농사)	1
그외(유공자연금)	1
근로소득	4
근로활동	109
기초연금	1
기타(생계급여)	3
기타(퇴직연금)	1
노인일자리	1
배우자연금	1
생계급여	5
유족연금	1
인터넷	1
재산소득	33
자상위	1
(비어 있음)	
중산계	1387

- 많은 데이터의 수기 작성
- 인적 오류 (오타 등)

## 프로젝트 추가시 장점

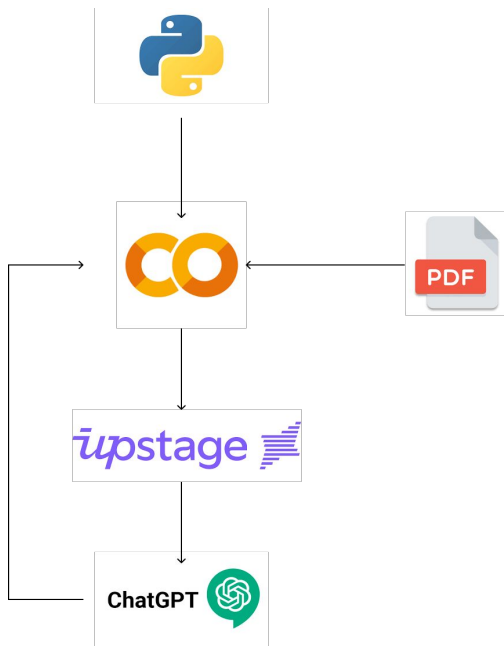
- 많은 작업량을 효과적으로 줄일 수 있다
- 인적오류(오타 등)를 모두 예방가능
- 데이터의 관리가 용이해짐(데이터 활용율 (증가))

1. 행정 문서처리
2. 고객 피드백 분석
3. 교육기관 서류 정리
4. 연구 데이터 처리

설문조사 뿐 아니라 PDF에  
관련한 데이터를 다루고 있기  
때문

# 향후 계획

## 현재 진행 상황



## 현재 만들어 놓은 프로그램에 추가해야할 사항

1. 예외처리에 관련된 대응 방안 생각  
- 악필 (글씨체)에 관하여 텍스트 불러오기
2. 특이한 설문방식 데이터 추출 방안 생각

## 추후 진행 예정

1. 정렬된 파일 Excel에 양식처럼 옮기기

# 참고 문헌

---

## 정보 출처

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

## 이미지 출처

<https://velog.io/@gmlstjq123/%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=S2RCTlh2UkpuMINVYWZXWU9hRmltZz09>

<https://chatgpt.com/>

# 감사합니다