산학협력프로젝트 – 룰루랩 x 성균관대학교

OCR을 이용한 화장품 성분정보제공 서비스 개발

성균관대학교
소프트웨어융합대학

lululab

Contents

2022 산학협력프로젝트 최종발표회

발표자: 소프트웨어학과 ***안윤지***

01 | 팀 & 회사 멘토 소개

성균관대학교

소프트웨어학과 3학년 민예은
-
소프트웨어학과 2학년 안윤지
-
소프트웨어학과 2학년 이승주

이주식 교수

×

룰루랩

이종하 연구 소장
-
이진희 연구원

과제명	OCR을 이용한 화장품 성분정보제공 서비스 개발
기대하는 결과물	화장품 성분이 찍힌 영상에서 화장품 이름과 성분에 해당하는 글자만을 추출하여 그것을 성분사전과 비교, 매칭할 수 있는 시스템

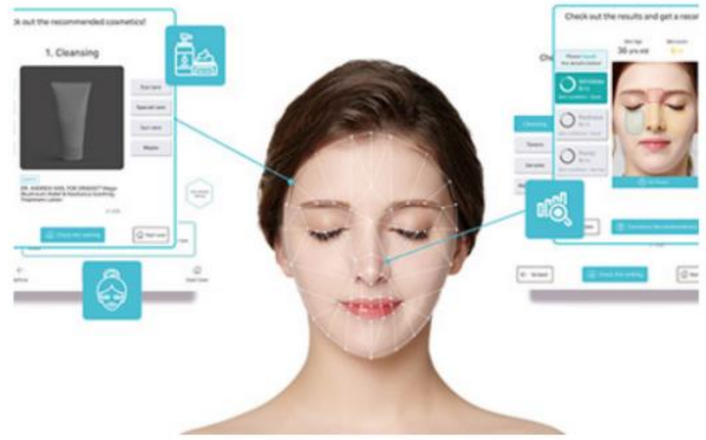
02 | 과제 배경 및 목적

lelelululab

- 자체 개발 인공지능 기반 피부진단기술을 보유하고 있으며 뷰티 브랜드사/유통사를 대상으로 피부진단 기술의 사업화 진행 중
- 17년 5월 삼성전자 C-Lab 에서부터 시작하여 혁신적인 아이디어와 기술을 바탕으로 피부 데이터를 기반으로 한 뷰티-헬스케어 AI 솔루션을 개발하고 있다.
- 현재 전 세계 80만개 이상의 피부 데이터와 전문평가기관에서 검증된 피부분석 정확도 92% 이상의 AI 기술력을 보유하고 있다.

인공지능 나이예측 & 얼굴촬영

딥러닝 기반 피부점수 제공



인공지능 제품추천

02 | 과제 배경 및 목적

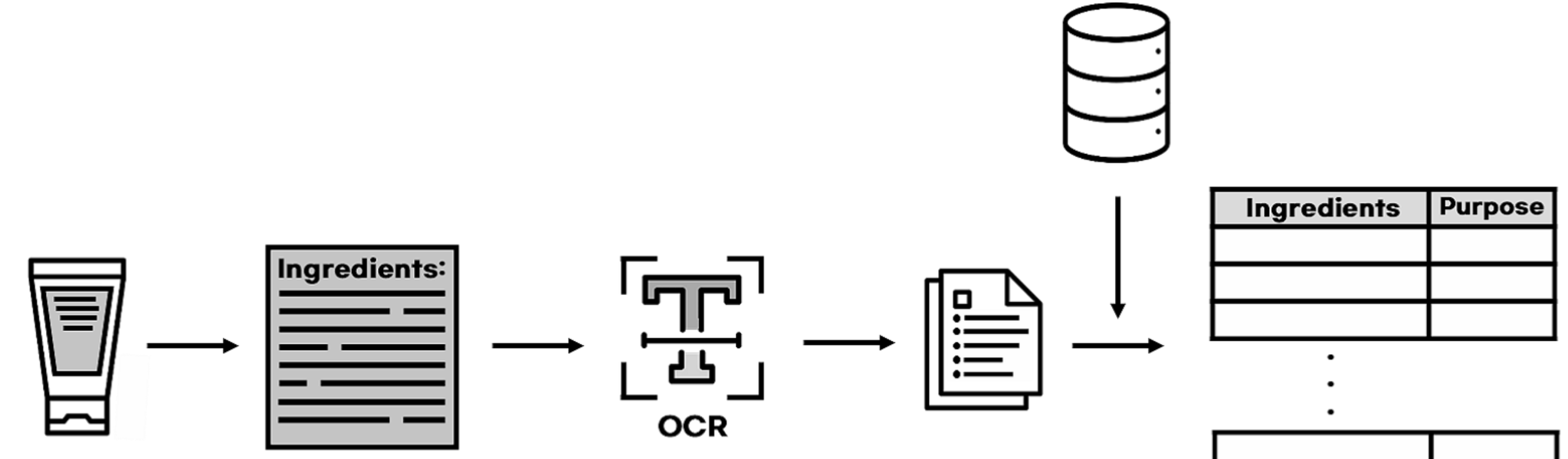
선정 배경

자사는 AI 피부분석서비스를 이용한 피부분석결과에 따라 피부에 맞는 화장품을 추천해주는 서비스를 계획 중에 있다. 또한 화장품 성분표를 찍어 올리면 그 화장품이 포함하고 있는 성분에 대한 정보를 제공해주는 서비스도 계획 중에 있다.
이를 위해서는 화장품 성분 데이터베이스 확보가 필요하다. 화장품 성분 정보를 수기로 작성할 수도 있지만, 매일같이 수십, 수백 개의 신상 화장품이 출시되고 있는 상황에서는 시간 면, 비용 면에서 매우 비효율적이다. 따라서 OCR 기술을 이용하여 화장품 용기에 적혀 있는 화장품 성분 정보를 수집할 수 있는 시스템을 개발하게 되었다.

목표

화장품 포장 상자 또는 용기의 이미지로부터 성분 텍스트를 인식해
이를 성분사전과 매칭함으로써 화장품의 성분정보를 수집하는 시스템 개발

02 | 과제 배경 및 목적



화장품 용기 이미지

성분표 부분만 크롭된 이미지

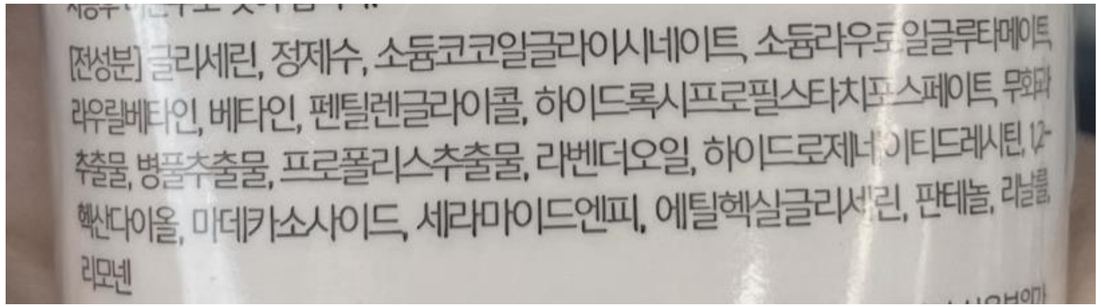
OCR을 통한 텍스트 추출

추출된 텍스트와 성분사전을 비교 매칭

성분사전에서 그 성분의 이름과 효능정보를 가져와서 정리

02 | 과제 배경 및 목적

OCR을 이용하여 화장품 용기 이미지로부터 성분 텍스트를 추출한다.



[전성분] 글리세린, 정제수, 소듐코코일글라이시네이트, 소듐라우로일글루타메이트 라우릴베타인, 베타인, 펜틸렌글라이콜, 하이드록시프로필스타치포스페이트 무화과 추출물,병풀추출물, 프로폴리스추출물, 라벤더오일, 하이드로제네이티드레시틴, 12-헥산디올, 마데카소사이드, 세라마이드엔피, 에틸헥실글리세린,판테놀, 리날룰, 리모넨

02 | 과제 배경 및 목적

OCR 인식 결과에서 성분을 하나씩 읽어서 성분사전과 매칭한다.

[전성분] 글리세린, 정제수, 소듐코코일글라이시네이트, 소듐라우로일글루타메이트 라우릴베타인, 베타인, 펜틸렌글라이콜, 하이드록시프로필스타치포스페이트 무화과 추출물,병풀추출물, 프로폴리스추출물, 라벤더오일, 하이드로제네이티드레시틴, 12-헥산디올, 마데카소사이드, 세라마이드엔피, 에틸헥실글리세린,판테놀, 리날룰, 리모넨

〈성분 사전〉

Ingre_index	Kor_name	Eng_name	Other_name	Purpose
7134	일로마스탯	Ilomastat	Butanediamide;...	금속이온봉쇄제, 피부컨디셔닝제(기타)
7135	세라마이드엔피	Ceramide NP	세라마이드 3;...	컨디셔닝제, 피부컨디셔닝제(기타)

03 | 적용 기술
(1) OCR

OCR

OCR: 사람이 쓰거나 인쇄된 문자를 이미지 스캐너로 획득하여 기계가 읽을 수 있는 문자로 변환하는 기술

-> OCR을 이용하여 화장품 용기에 적혀 있는 화장품 성분표가 찍힌 이미지를 받아서 이미지 속 텍스트를 인식하고 추출한다

Google Cloud Vision API

성능	오픈소스 OCR 모델 중에서 가장 성능이 우수함.
선정 배경	- 프로젝트 초기 단계에서는 무료 오픈소스 OCR 모델을 자체 학습을 통해 성능을 향상시켜 사용하려고 시도 - 그러나 자체학습에 필요한 데이터셋이 없는 점, 자체 학습 외의 다른 과제에 투자할 시간도 고려해야 하는 점, 유의미한 성능 향상을 보장할 수 없는 점 등의 이유로 모델 학습은 진행하지 않기로 함. - 따라서 성능이 가장 우수하다고 판단되는 Google cloud vision api를 기준으로 전후처리 성능을 향상시키는 것을 중점으로 두고 프로젝트를 진행함.

03 적용 기술
(2) 유사도 검사 알고리즘

〈 유사도 검사 〉

유사도를 측정하여 가장 유사도가 높은 성분과 매칭함으로써 오타가 있는 성분텍스트도 사전과 정확하게 매칭할 수 있다.

편집거리 알고리즘(Levenshtein distance, edit distance)

- 두 문자열의 유사도를 측정하는 가장 대표적인 방법.
- 둘 중 하나의 문자열이 다른 하나의 문자열과 같아지기 위해서 더하거나(insert), 빼거나(delete), 바꿔보거나(replace), 위치를 변경(transposes) 하는 연산이 최소 몇 번 이루어져야 하는지 계산함으로써 두 문자열의 유사도를 측정한다.
- 대표적으로 Peter Norvig, Symspell이 있다.

파이썬 표준 라이브러리 difflib

- 별도의 설치가 필요 없는 파이썬 표준 라이브러리
- 두 문자열의 유사도를 계산해준다.

03 적용 기술
(3) 후처리 알고리즘

〈 후처리 알고리즘 〉 – 유사도 검사를 통한 매칭

적용 대상 OCR 결과에서 성분 문자열 안에 오타가 있는 경우 ex) 헥산다이올 –〉 헥산디올

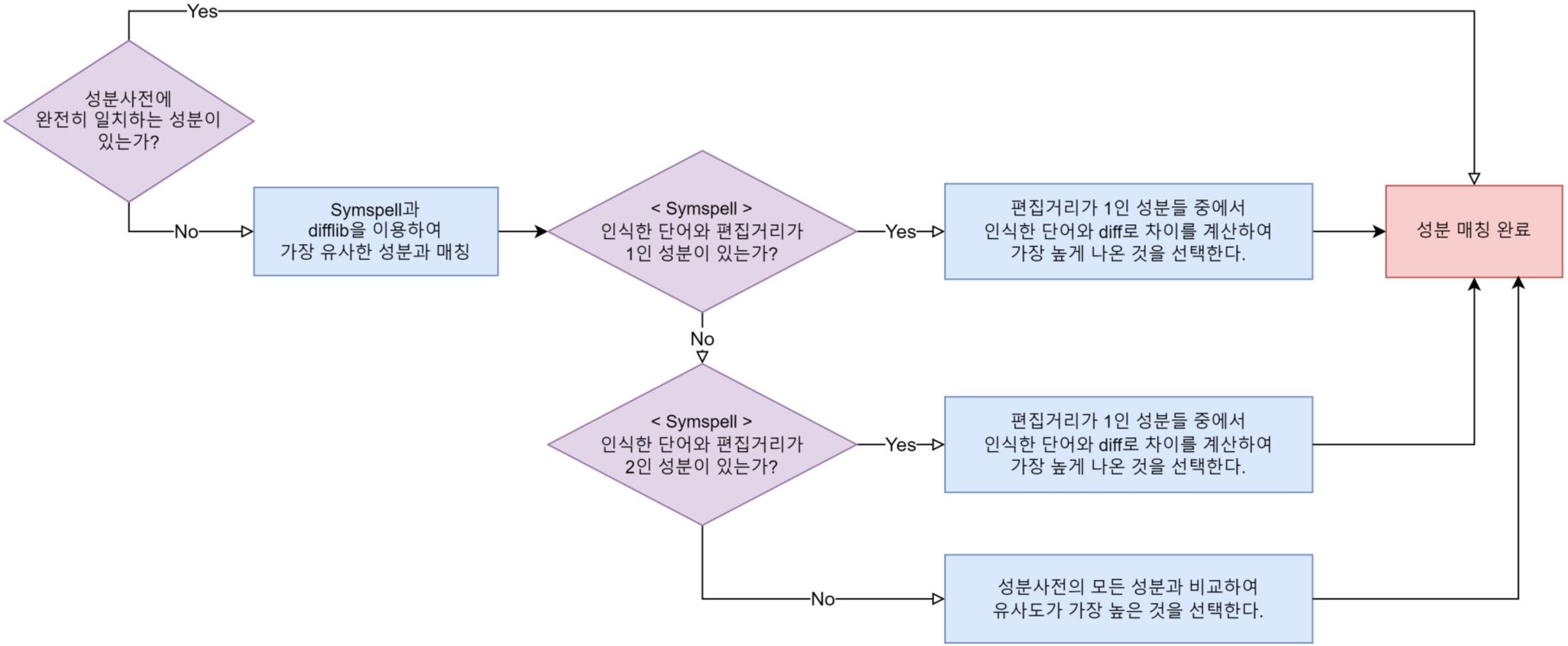
대응 방법 Symspell과 difflib library를 같이 사용하여 유사도 검사를 통해 성분 매칭

- diff만 사용하면 2만개의 성분을 다 비교하여 유사도가 제일 높은 것을 알아내는 것이므로 시간이 오래 걸림
- symspell만 사용하면 편집거리가 같은 단어에서 무엇이 더 높은 유사도를 가지는지 측정하기 어려움

>> symspell을 사용하여 2만개 성분 사전에서 유사한 단어를 추려 diff 사용범위를 좁혀주면, 위와 같은 단점을 보완해줄 수 있음.

03 적용 기술

(3) 후처리 알고리즘



03 적용 기술
(3) 후처리 알고리즘

〈 후처리 알고리즘 〉 - 파싱 문제

적용 대상 OCR 결과에서 콤마, 점 등의 구분자가 누락되어 두 성분 사이에 공백만 남아서 이를 구분해주기 위한 작업이 필요한 경우
ex) "헥산다이올, 페녹시에탄올" -〉 "헥산다이올 페녹시에탄올"

대응 방법 Symspell과 difflib library를 같이 사용하여 유사도 검사를 통해 성분 매칭

① 어떤 문자열의 Symspell+diff를 거쳐 나온 결과를 검사한다. 만약 유사도 0.85 이하로 매칭이 된다면, 파싱의 문제가 있다고 판단.
② 공백을 기준으로 단어들을 분리해주고, 각각 Symspell+diff의 과정을 거친다.
③ 처음 문자열로 매칭한 결과와, 공백으로 분리하여 각각 유사도를 검사한 결과를 비교하여 더 높은 유사도로 매칭한 것을 결과로 낸다.
- 전자와 매칭된 성분과의 유사도 〉 후자와 매칭된 성분과의 유사도 = 파싱문제가 아니었다.
- 전자와 매칭된 성분과의 유사도 〉 후자와 매칭된 성분과의 유사도 = 파싱문제였고, 공백을 기준으로 나눠준 각 단어들이 독립적인 성분이다.

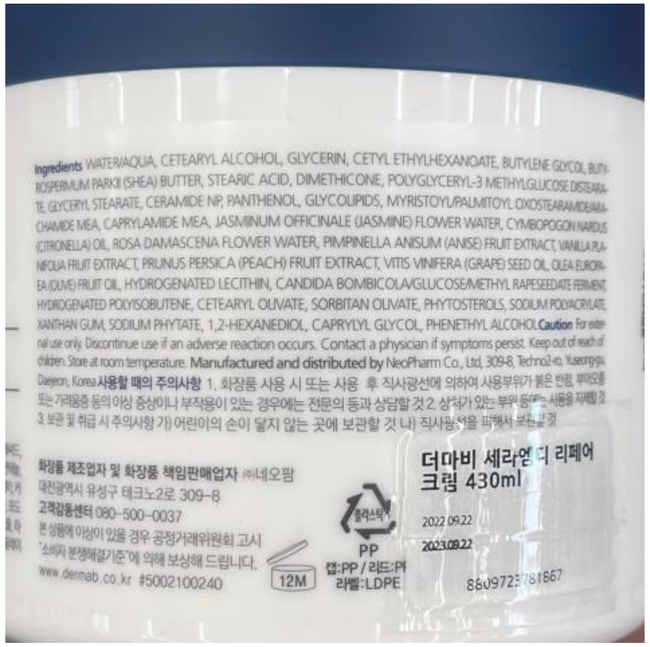
04 | 기술의 차별성 및 독창성

문자를 정확하게 읽어내기 위한 OCR 알고리즘에 대한 연구는 인터넷에서 쉽게 찾아볼 수 있으며, 오픈소스의 형태로도 존재하여 변형하여 쓸 수 있고, 시중에서 구매도 할 수 있음.

그러나 본 프로젝트에서 개발하고자 하는 화장품 성분에 대한 OCR 알고리즘은 화장품 성분표의 전문용어를 인식하는 것에 중점을 두고,
이를 데이터베이스화 시키거나 일치하는 성분을 찾아내야 한다는 점에서 단순히 글자를 인식하는 단계의 OCR 연구와는 차별화됨

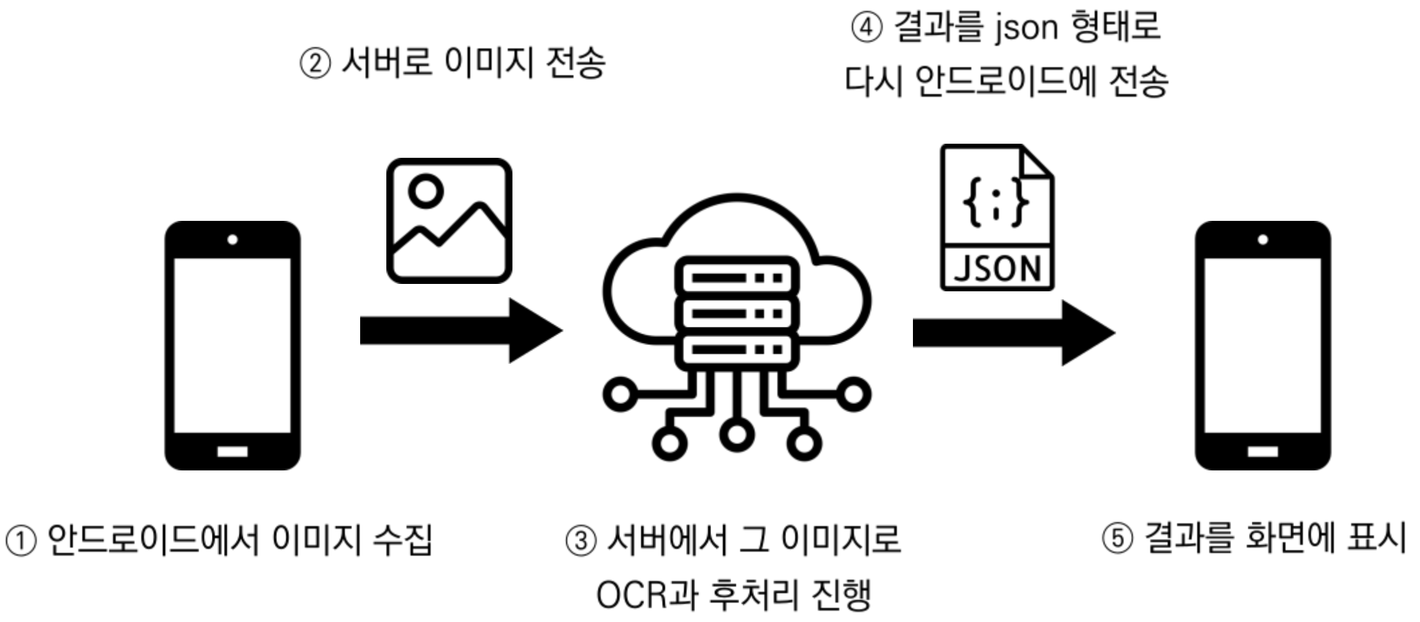
화장품 성분표는 주로 용기나 박스에 붙어있어
용기나 환경의 특성에 따라 많은 제약을 받게 됨

- 대체로 한정된 공간 안에 몇 십 개에 달하는 성분정보를 적어야 하기 때문에 글씨가 매우 작으며, 글씨 사이의 간격이 매우 좁다.
- 화장품 용기의 대다수는 튜브 형태이거나 둥근 굴곡을 가진 경우가 많아 왜곡과 빛 반사가 심하다.



05 | 작품 소개

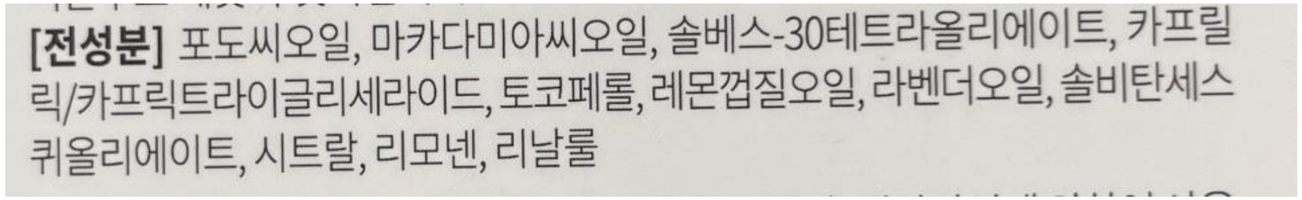
플라스크 서버를 구축하여 안드로이드에서 서버로 이미지를 전송하고,
서버에서 그 이미지로 OCR과 후처리를 진행한 뒤,
그 결과를 안드로이드로 다시 반환하도록 설계함



05 | 작품 소개

화장품의 성분정보 수집 및 제공

① 화장품의 성분정보가 담긴 이미지 수집
② OCR을 이용하여 텍스트 추출
③ 성분정보 수집


[전성분] 포도씨오일, 마카다미아씨오일, 솔베스-30테트라올리에이트, 카프릴릭/카프릭트라이글리세라이드, 토코페롤, 레몬껍질오일, 라벤더오일, 솔비탄세스퀴올리에이트, 시트랄, 리모넨, 리날룰

성분명
성분의 효능, 배합 목적

위험도 (낮은/중간/높은)

포도씨오일 피부컨디셔닝제	낮은 위험도
마카다미아씨오일 피부컨디셔닝제(유연제)	낮은 위험도
솔베스-30테트라올리에이트 계면활성제(유화제)	낮은 위험도
카프릴릭/카프릭트라이글리세라이드 착향제, 피부컨디셔닝제(수분차단제), 용제	낮은 위험도
레몬껍질오일 착향제, 피부컨디셔닝제(기타)	중간 위험도
라벤더오일 착향제, 피부컨디셔닝제(기타)	낮은 위험도
솔비탄세스퀴올리에이트 계면활성제(유화제)	낮은 위험도
시트랄 감미제, 착향제	중간 위험도
리모넨 착향제, 용제	중간 위험도
리날룰 착향제	중간 위험도

어플리케이션을 통해 화장품 성분 정보 제공

06 | 과제 성과 및 기대효과

〈 과제 성과 – 후처리 성능〉

시간	한 이미지를 돌리는데 5~30초 소요. 파싱 문제가 없고, 오타가 적은 경우 굉장히 빠른 속도로 매칭할 수 있으며, 오타가 많은 경우 최대 30초가 걸림. (대체로 한 사진에 오타는 평균 2~4개 수준)
매칭 정확도	[영어] - 후처리 적용 전 : 정확하게 인식된 성분 개수 / 전체 성분 개수 = 4532/4778 = 94.85% - 후처리 적용 후 : 정확하게 인식된 성분 개수 / 전체 성분 개수 = 4391/4715 = 93.13% [한글] - 후처리 적용 전 : 정확하게 매칭된 성분 개수 / 전체 성분 개수 = 3134/3481 = 90.03% - 후처리 적용 후 : 정확하게 매칭된 성분 개수 / 전체 성분 개수 = 3277/3481 = 94.14%

06 | 과제 성과 및 기대효과

〈 기대 효과 〉

화장품 성분 DB 구축

성분 DB를 수집하기 위한 전략 중 하나로 사용할 수 있다.
어떤 화장품의 성분 정보가 아직 등록되지 않은 경우, 불특정다수의 사용자가 자사의 어플리케이션을 이용하여
자신의 화장품 용기를 사진으로 찍으면, 그것을 OCR을 이용하여 텍스트를 추출함으로써 성분정보를 수집할 수 있다.

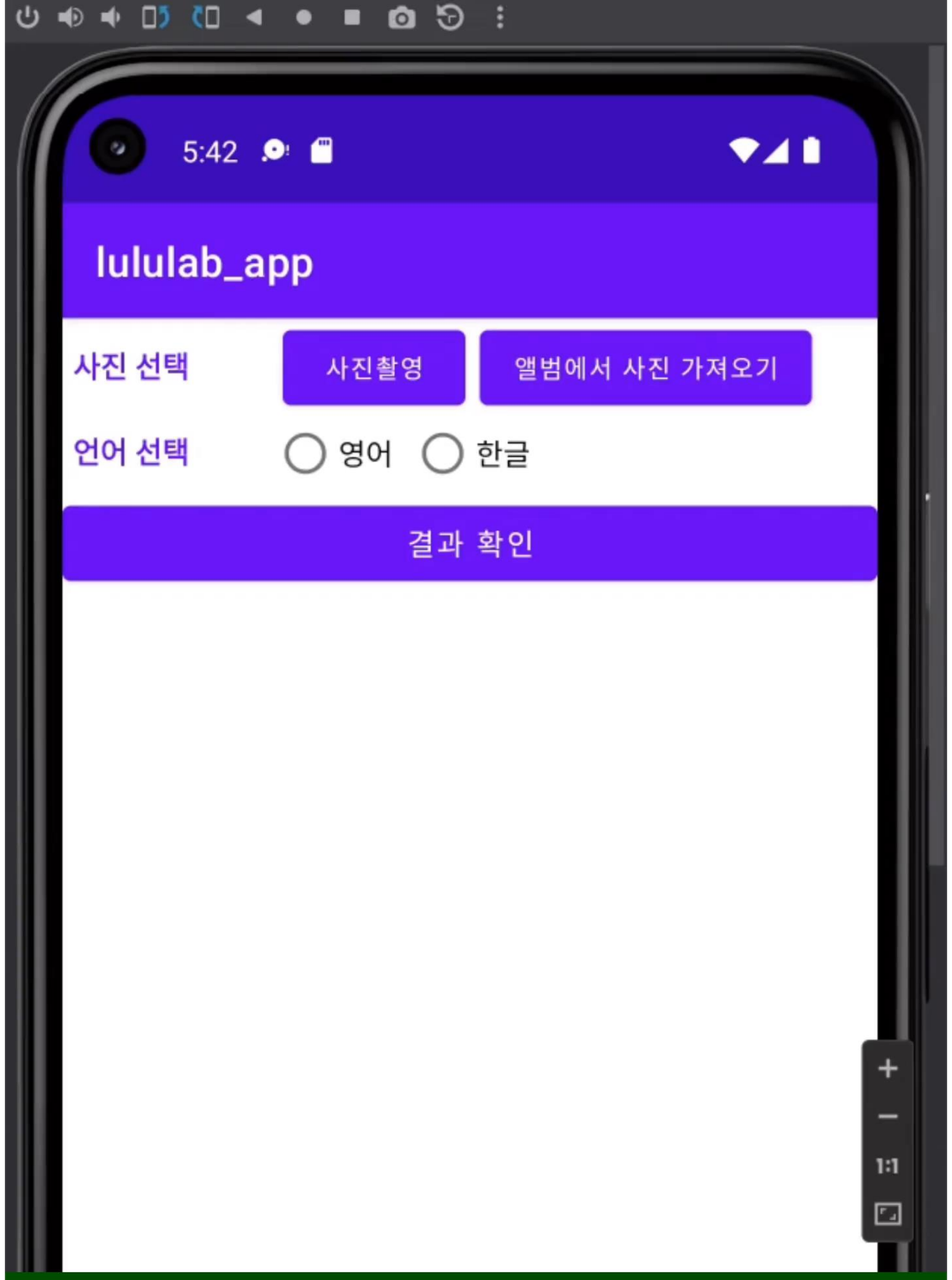
화장품 성분 DB 활용

- 각 피부타입에 맞는 성분이 있는 화장품을 추천해주기 위해 사용한다.
- 어떤 성분이 피부에 맞아서, 혹은 맞지 않아서 그 화장품을 추천해주었는지를 설명할 수 있다. >> 신뢰성
- 사용자가 자신의 화장품의 성분이 궁금할 때 그 정보를 제공할 수 있다. >> 지속적인 사용 유도

07 | 개선점 및 향후 계획

개선할 점	후처리 성능 향상 - 특수한 구분자에 대한 대응, 성분이 아닌 글자가 인식되었을 때 대응 - 영어도 한글만큼 성능을 향상시킬 필요가 있음.
향후 계획	12/31까지 프로젝트를 계속 진행할 예정. - 화장품 이름으로 성분 검색하는 기능 추가하기 - 사전에 등록되어 있지 않은 성분이 검출되었을 때 성분사전에 추가하기 - 후처리 속도 향상시키기

시연 동영상



09 Appendix

(1) 산학협력과제 요약

<table>
<tr><th>과제명</th><th>OCR을 이용한 화장품 성분정보제공 서비스 개발</th><th>기업명 / 담당자</th><th>룰루랩</th></tr>
<tr><td rowspan="3">과제개요</td><td rowspan="3">화장품의 성분 데이터베이스 확보를 위해
화장품 포장 상자 또는 용기의 이미지로부터
성분 텍스트를 인식해
데이터베이스에 저장하는 시스템 구축</td><td rowspan="3">참여 학생</td><td>안윤지(소프트웨어학과, 2학년), 팀장</td></tr>
<tr><td>민예은(소프트웨어학과, 3학년)</td></tr>
<tr><td>이승주(소프트웨어학과, 2학년)</td></tr>
<tr><td rowspan="3">과제 기간</td><td rowspan="3">2022년 4월 6일 ~ 2022년 12월 31일</td><td>기업 멘토</td><td>이종하 수석, 이진희 연구원</td></tr>
<tr><td>지도교수</td><td>이주식 교수</td></tr>
<tr><td>산학교수</td><td>이주식 교수</td></tr>
<tr><td>기대하는
결과물</td><td colspan="3">화장품 성분이 찍힌 영상에서 화장품 이름과 성분에 해당하는 글자만을 추출하여
그것을 성분사전과 비교, 매칭할 수 있는 시스템</td></tr>
</table>

Thank you