이미지 데이터 기반의 숙소 추천시스템

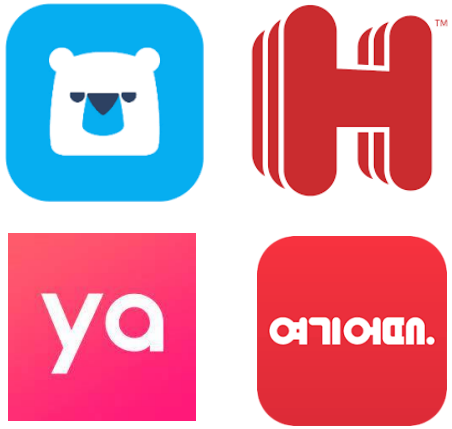
비타민 8기 추천시스템 3조
이선호 서진슬 조한준 최예은

INDEX

1. 주제 선정 이유

1 주제 선정 이유

1) 에어비앤비 숙소 추천시스템



다양한 숙소 추천 플랫폼들과의 차이점

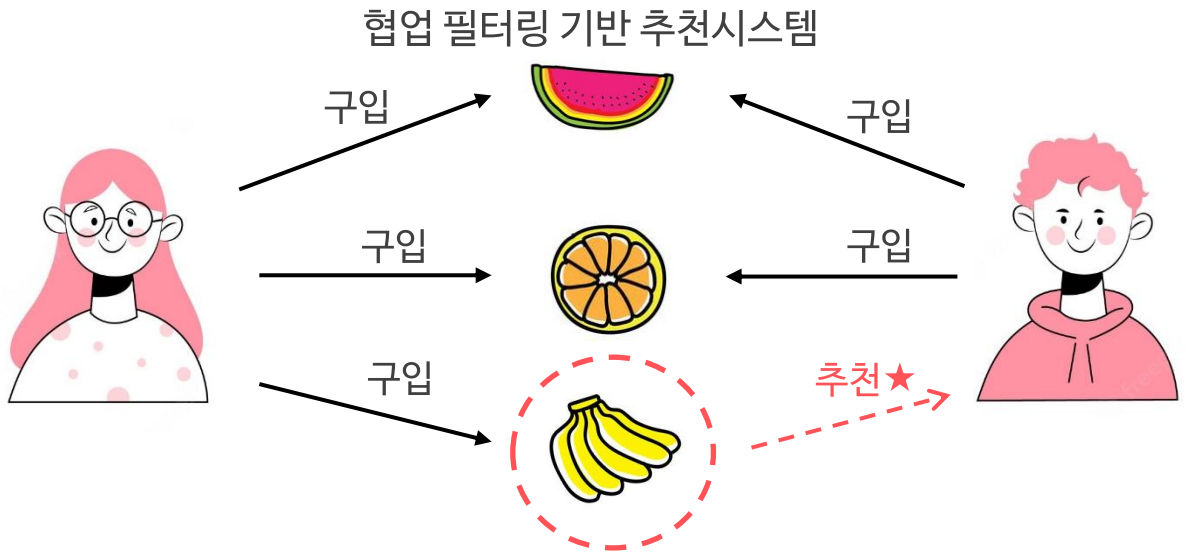
1. 감성 숙소
2. 다양한 숙소 이미지
3. 호스트가 직접 촬영한 사진 사용

비슷한 분위기를 가진 다른 숙소를 추천 → 사용자의 선택 폭 확장

숙소가 실제 예약될 확률 ↑

1 주제 선정 이유

2) 협업필터링 기반 추천시스템 문제점



- 콘텐츠에 대한 이질적이고 주관적인 선호도
- 데이터 희박성(Sparsity)문제
- 데이터 처리 속도 및 확장성 문제
- 콜드 스타트(Cold-Start)문제

기존 협업필터링 기반 추천시스템 : 유저, 아이템 데이터 사용

But 에어비앤비의 유저 데이터를 구할 수 없음 ❌

1 주제 선정 이유

airbnb

데이터 수집이 가능한 이미지 데이터

숙소 이미지를 사용하는 추천 시스템 개발

2. 데이터 소개

2 데이터 소개

1) 에어비앤비 크롤링 과정

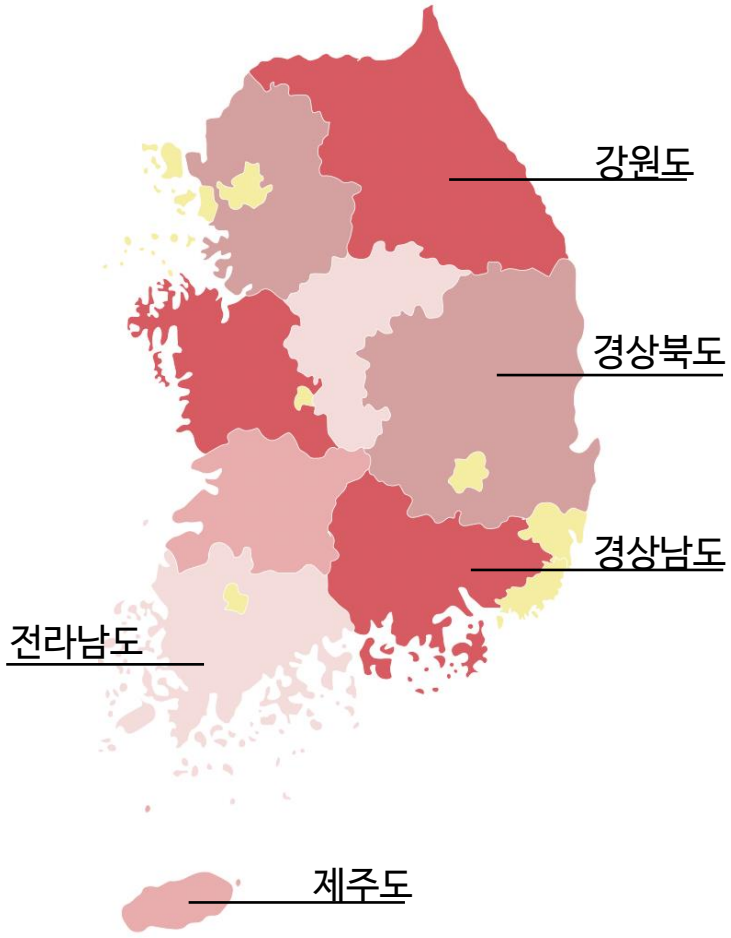


표 2-1-4 여행지별 여행 횟수 총량

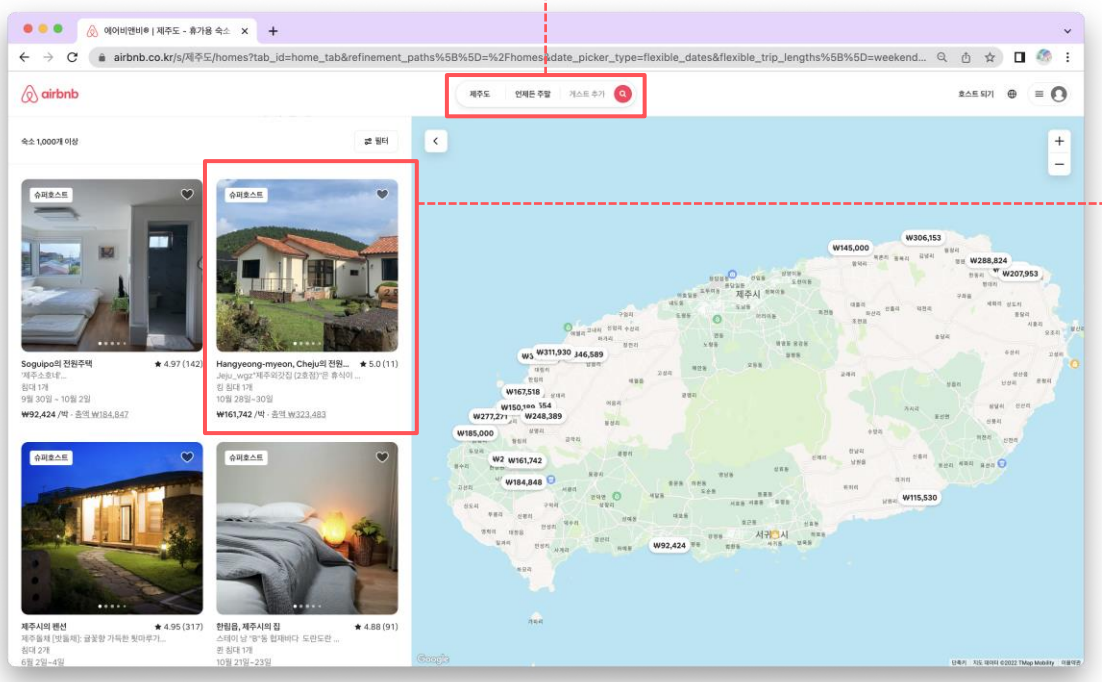
(단위: 천회)

구분	국내여행			관광여행			기타여행		
	전체	숙박	당일	전체	숙박	당일	전체	숙박	당일
전체	245,127	87,785	157,342	198,800	74,894	123,906	46,327	12,891	33,435
서울	13,259	2,569	10,690	7,608	1,730	5,878	5,651	839	4,812
부산	9,953	4,503	5,449	8,544	3,940	4,604	1,408	563	846
대구	4,007	1,354	2,653	2,970	897	2,073	1,037	457	580
인천	9,142	2,182	6,961	7,373	1,695	5,678	1,770	487	1,283
광주	1,808	837	971	1,187	571	615	621	266	356
대전	4,603	1,784	2,819	2,327	905	1,421	2,276	878	1,398
울산	4,194	1,265	2,929	3,304	1,021	2,283	890	244	646
세종	1,883	320	1,562	892	165	727	990	155	835
경기	53,400	7,142	46,259	41,183	5,427	35,756	12,218	1,715	10,503
강원	25,422	15,875	9,547	23,446	15,080	8,366	1,976	795	1,181
충북	10,146	4,102	6,044	7,643	3,151	4,493	2,502	951	1,551
충남	19,146	7,170	11,976	15,433	5,775	9,658	3,713	1,395	2,318
전북	13,838	5,618	8,220	11,984	4,612	7,371	1,854	1,006	849
전남	20,129	8,725	11,404	17,731	7,757	9,974	2,398	968	1,430
경북	22,634	7,556	15,078	19,393	6,454	12,939	3,241	1,102	2,139
경남	23,508	9,484	14,023	19,902	8,316	11,586	3,606	1,169	2,437
제주	11,097	10,052	1,045	10,734	9,979	755	362	72	290

주 : 1회 여행 시 여러 시도를 방문할 수 있으므로 각 시도별 합과 전체 결과는 상이함

2021 국민 여행 조사 결과 참고하여 5개의 국내 여행지 선택

2 데이터 소개



제주도 | 언제든 주말 | 게스트 추가

5개의 지역, 언제든 주말(유연한 일정)

Hangyeong-myeon, Cheju의 전원... ★5.0 (11)
Jeju_wgz"제주외갓집 (2호점)"은 휴식이 ...
킹 침대 1개
10월 28일~30일
₩161,742 /박 · 총액 ₩323,483

- 숙소 링크
- 숙소 대표 이미지
- 평점 및 후기 개수
- 숙소의 이름

각 지역 당 300개의 에어비앤비 숙소 크롤링

2 데이터 소개

2) 데이터 전처리 과정

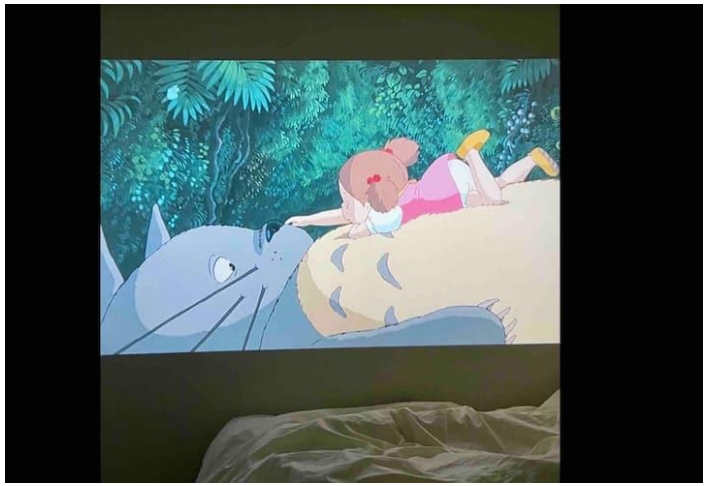
① 중복 숙소 및 대표이미지가 숙소와 관련 없는 경우 삭제

1500개 → 1118개

▼ 삭제된 데이터 예시

강원_283.jpg

경북_148.jpg



제주_296.jpg

2 데이터 소개

② 숙소 유형, 평점, 리뷰 개수 등 전처리

숙소이름
애월읍, 제주시의 전원주택
애월읍, 제주시의 전원주택
Hallim-eub, Cheju의 게스트용 별채
애월읍, 제주시의 캠핑카
Hangyeong-myeon, Cheju의 개인실
Namwon-eup의 집
한림읍, 제주시의 게스트용 별채
Hangyeong-myeon, Cheju의 집
애월읍, 제주시의 전원주택
애월읍, 제주시의 게스트용 별채
제주시의 펜션
Jochon-eup의 펜션
Hangyeong-myeon, Cheju의 게스트용 별채
성산읍, 서귀포시의 집
구좌읍, 제주시의 전원주택

펜션-빌라

호텔-리조트

캠핑

별채

다인실

아파트

개인실

총 7개 숙소유형

2 데이터 소개

최종 데이터 셋

숙소 이미지	숙소 링크	숙소 유형	평점	후기 개수
제주_123.jpg	'http://www.airbnb.…'	별채	4.95	106

[숙소 1118 개]

▼ 숙소 이미지 예시

강원_28.jpg

전남_296.jpg

숙소 이미지와 숙소 유형 → **Feature Engineering**

3. 모델링

3 모델링

1) 모델 개요

☑ 유저가 원하는 분위기의 숙소와 비슷한 숙소를 추천해주는 모델

3가지의 모델 비교 후 선택

Input

모델 1 코사인 유사도

모델 2 KNN

모델 3 K-means

Output

3 모델링

2) 이미지 feature 추출

Resnet18

- 2015 ILSVRC(이미지 인식 경진대회) 에서 우승, 처음으로 사람보다 높은 정확도를 기록함
- 기존 딥러닝의 문제점인 Vanishing Gradient 문제를 해결하여 아주 깊은 네트워크의 학습이 가능함을 증명한 모델
- 많은 논문과 모델에서 응용되고 있는 모델
- 이미지넷(1000개의 클래스)으로 학습된 pre-train 모델 사용 가능

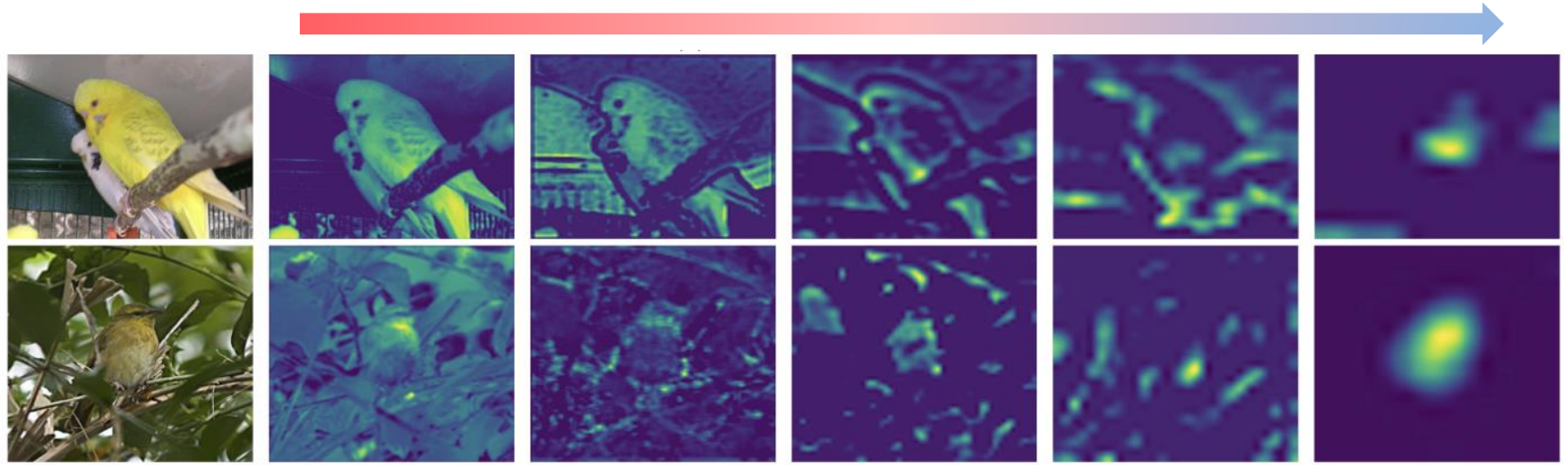
Resnet18을 선택한 이유

- Resnet 중에 가장 가볍고 빠른 모델 → 실시간 추천 시스템 적용할 때 장점
- 각각의 레이어에서 상대적으로 적은 수의 피처를 얻을 수 있음

ex) resnet18(최종 레이어) : 512개, resnet50(최종 레이어) : 2048개

3 모델링

LayerCAM: Exploring Hierarchical Class Activation Maps for Localization



(c) Feature maps

Fig. 4. (a-b) show the variances of the gradients corresponding to each feature map at different stages of VGG16. (c) illustrates the feature maps randomly selected from different stages.

"초반 레이어는 이미지의 타겟, 배경 등 모두에게서 세밀한 특징 정보를 뽑고,
후반 레이어는 타겟 레이어에 대한 정보를 뽑는다."

3 모델링

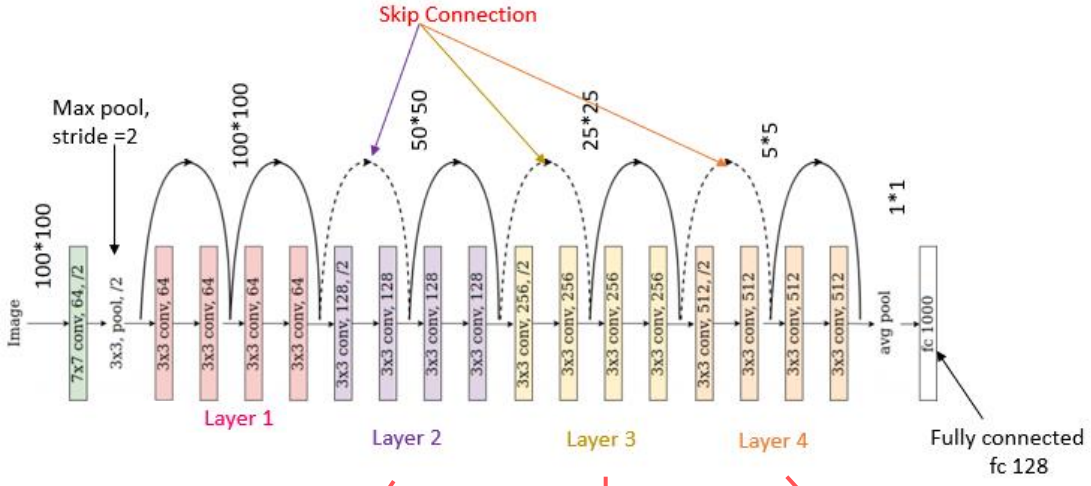
Layer2를 선택한 이유

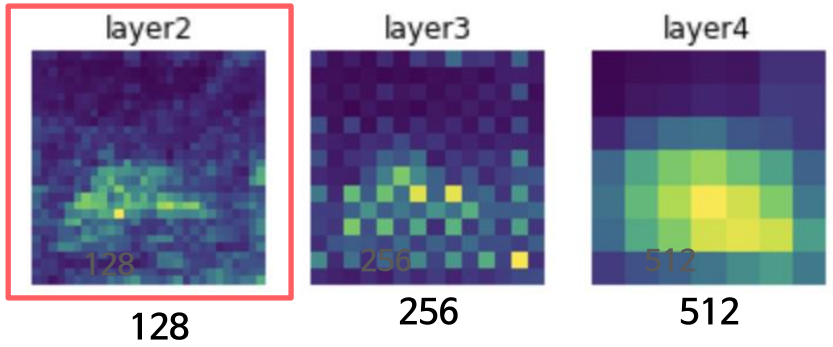
강원_155.jpg

- 가장 전체적인 이미지의 피처를 뽑을 수 있다.
- 아웃풋 피처 차원이 모든 레이어중에서 작아서 차원의 저주 위험성이 다른 레이어에 비해 비교적 낮다.

→ Layer 2 선택

Resnet18로 layer별 feature 추출한 결과





3 모델링

Resnet 18 Layer2로 각 이미지 feature 추출한 결과 ✚ 숙소 타입 원핫 인코딩 결과

128개 ↓ ↘ 7개

image_index	L2_0	L2_1	L2_2	L2_3	L2_4	L2_5	L2_6	L2_7	L2_8	L2_9	...	L2_125	L2_126	L2_127	single	multi	outhouse	apartment	camping	pension	hotel
gyeongnam_203	0.139648	0.169798	0.136722	0.108873	0.176598	0.123397	0.120343	0.245084	0.149954	0.288930	...	0.113126	0.305492	0.151653	1	0	0	0	0	0	0
gyeongbuk_194	0.140625	0.193402	0.150842	0.108315	0.186995	0.132722	0.113584	0.244671	0.152429	0.307828	...	0.118192	0.322847	0.149029	1	0	0	0	0	0	0
gyeongnam_217	0.141954	0.178325	0.142441	0.102194	0.190987	0.121189	0.114551	0.233060	0.150725	0.302136	...	0.106955	0.321539	0.152065	0	0	0	0	0	1	0
jeju_102	0.151979	0.173062	0.145052	0.106857	0.177310	0.133428	0.102689	0.248616	0.144603	0.297525	...	0.106826	0.306511	0.142500	0	0	0	1	0	0	0
jeju_116	0.132614	0.175584	0.139151	0.111856	0.175659	0.150412	0.101449	0.255162	0.157335	0.303659	...	0.096464	0.333100	0.142683	0	0	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
gyeongnam_232	0.135774	0.180139	0.146762	0.105828	0.186918	0.125276	0.125666	0.242583	0.143527	0.314263	...	0.118857	0.323966	0.153585	0	0	0	0	0	1	0
gyeongnam_226	0.130606	0.186938	0.146593	0.120030	0.192117	0.121673	0.114408	0.237203	0.151742	0.305917	...	0.105625	0.323678	0.161128	0	0	0	0	0	1	0
jeju_133	0.138077	0.169712	0.142581	0.103773	0.180805	0.136002	0.108735	0.218698	0.163893	0.294224	...	0.101078	0.305961	0.161473	0	0	0	0	0	1	0
jeju_127	0.142617	0.171123	0.140616	0.106747	0.173767	0.127077	0.097530	0.251893	0.140797	0.298628	...	0.101218	0.304679	0.142151	0	0	0	0	0	1	0
gyeongbuk_199	0.130180	0.186989	0.155539	0.121664	0.179350	0.132715	0.115970	0.228854	0.162364	0.294402	...	0.110830	0.312408	0.170806	0	0	0	1	0	0	0

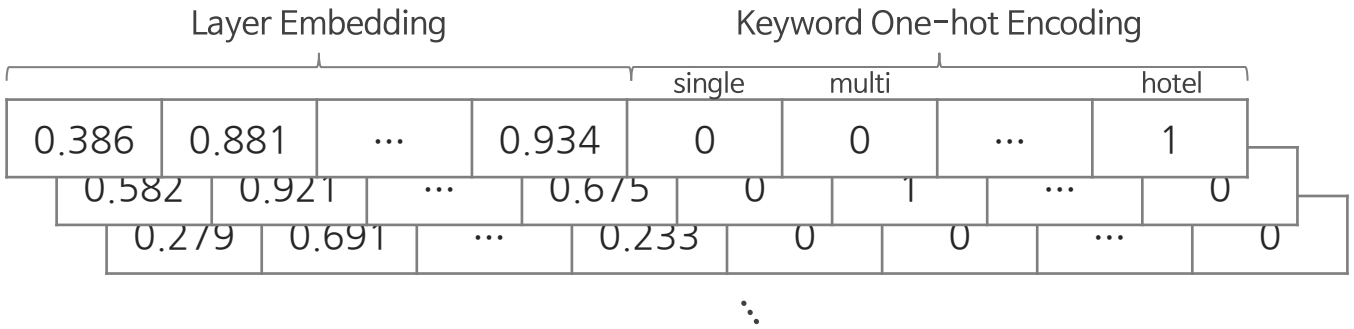
1117 rows × 135 columns

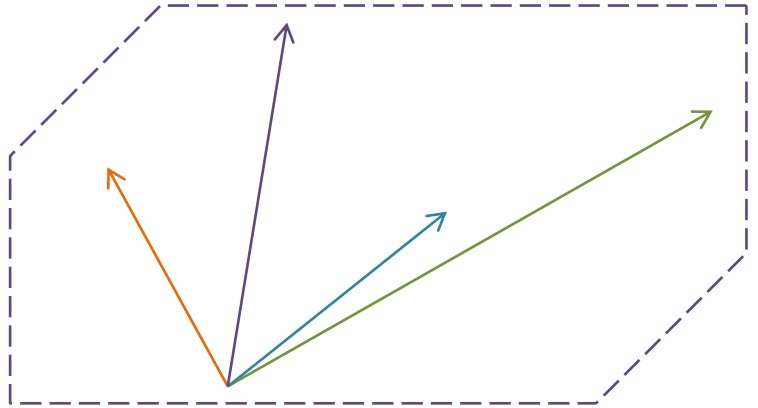
→ 모델링에 사용할 최종 데이터 셋

3 모델링

3) 모델링

① 코사인 유사도





공간상의 좌표로 나타냈을 때 서로 비슷한 방향을 바라볼수록 유사

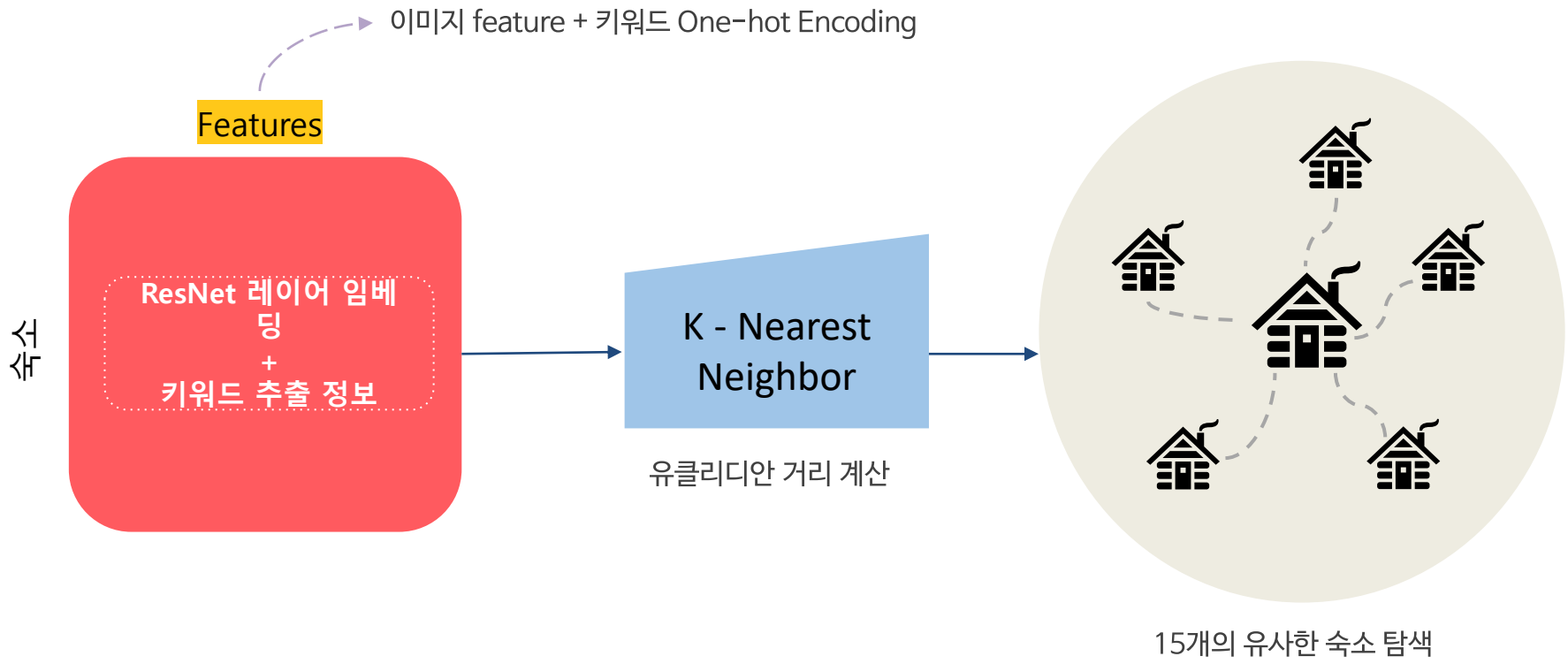
$$\cos(a,b) = \frac{a \cdot b}{|a||b|}$$

숙소의 feature vector끼리 cosine 유사도를 구하여 가장 값이 높은 상위 N개 추출

3 모델링

3) 모델링

② KNN

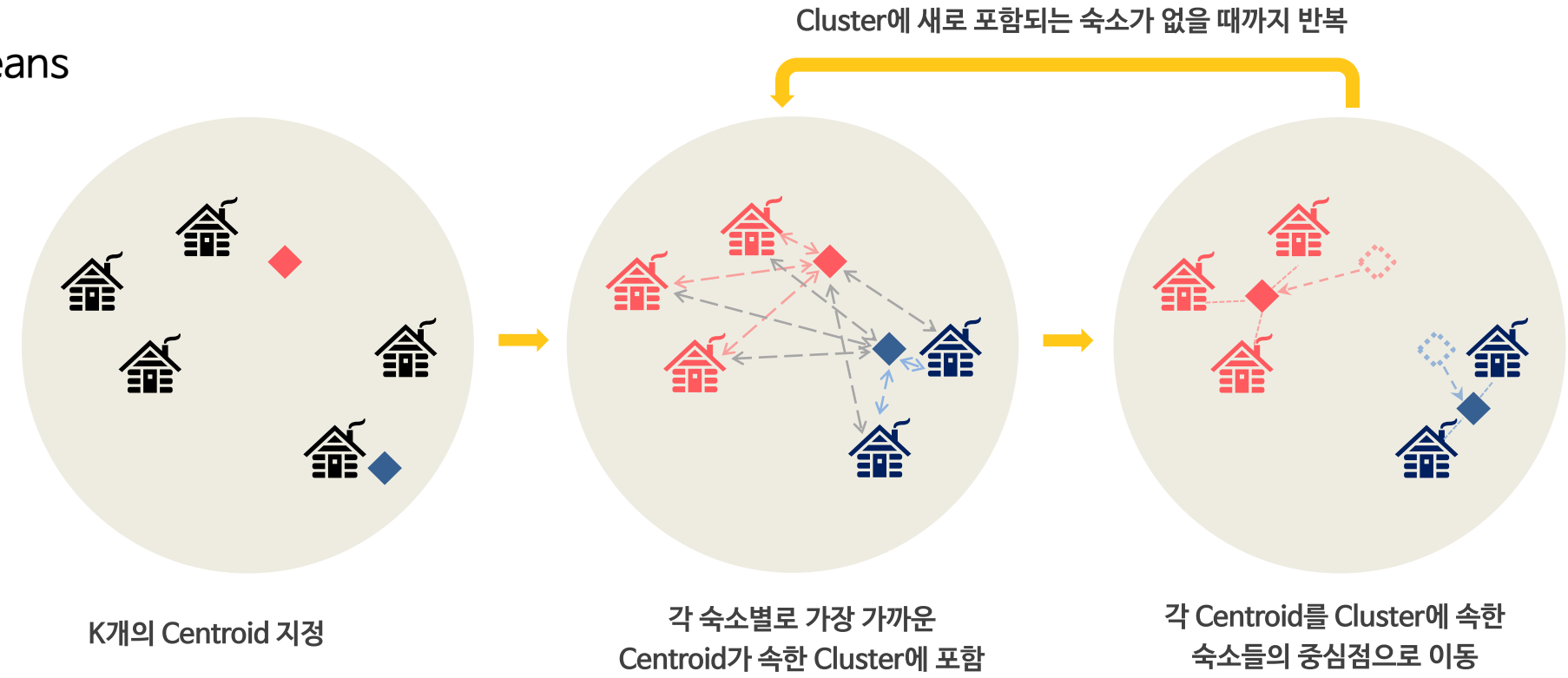


Features에 관한 유클리디안 거리 기반으로 가장 가까운 15개의 숙소 아이템 찾기 → 선정한 숙소 후보를 cosine 유사도로 정렬

3 모델링

3) 모델링

③ K-means



선정한 숙소 후보를 cosine 유사도로 정렬

3 모델링

4) 모델 평가

① 내부/외부 score

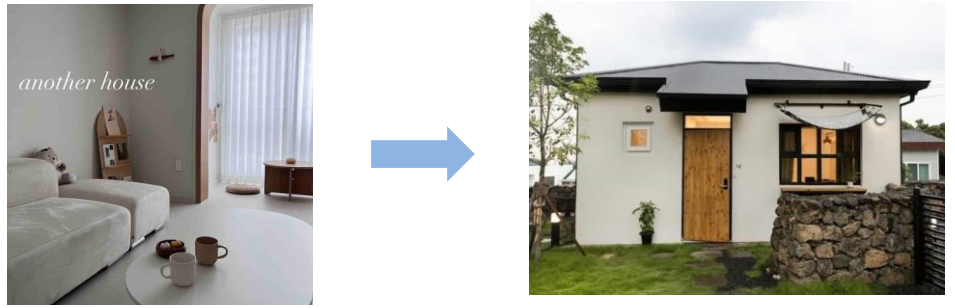
Input image	Output image (추천결과)	Score
숙소 내부	숙소 내부	+1
숙소 외부	숙소 외부	+1
숙소 내부	숙소 외부	+0
숙소 외부	숙소 내부	+0

▼ 내부(input) → 내부(output) 예시 : +1점

input

output

▼ 내부(input) → 외부(output) 예시 : +0 점



input

output

3 모델링

4) 모델 평가

② 컬러 비교 score

Input

Output

ColorThief를 사용하여 이미지의 대표색 3가지를 추출

- 3가지 중 2가지 색 이상 비슷하면 +1
- 3가지 중 2가지 색 이상 비슷하지 않으면 +0

3 모델링

4) 모델 평가 – 예시

Input

Inside

Output

Inside

👍

outside

👎

3 모델링

4) 모델 평가

지역별로 테스트한 결과

모델 \ Input	제주_35	전남_138	경북_260	강원_237	경남_103	SUM	MEAN
COS	10	7	7	10	8	42	8.4
KNN	10	10	9	9	9	47	9.4
K-means	10	9	7	10	9	45	9

추천 결과 이미지 5개에 대한 계산 결과(외부내부:5점 + 색 5점)

score가 가장 높은 KNN을 모델로 선택

4. 결과

4 결과

추천결과 1 (내부)

INPUT

전라남도 아파트

OUTPUT

* 위치 : 강원도
* 숙소유형 : 아파트
* 평점 : 4.99/5점
* 후기개수 : 102개

* 위치 : 경상북도
* 숙소유형 : 아파트
* 평점 : 신규숙소
* 후기개수 : 0개

* 위치 : 전라남도
* 숙소유형 : 아파트
* 평점 : 4.94/5점
* 후기개수 : 36개

* 위치 : 전라남도
* 숙소유형 : 아파트
* 평점 : 4.92/5점
* 후기개수 : 61개

* 위치 : 경상남도
* 숙소유형 : 아파트
* 평점 : 4.97/5점
* 후기개수 : 37개

4 결과

추천결과 2 (외부)

INPUT

경상북도 개인실

OUTPUT

* 위치 : 경상북도
* 숙소유형 : 개인실
* 평점 : 5.0/5점
* 후기개수 : 57개

* 위치 : 전라남도
* 숙소유형 : 개인실
* 평점 : 4.87/5점
* 후기개수 : 86개

* 위치 : 강원도
* 숙소유형 : 개인실
* 평점 : 4.88/5점
* 후기개수 : 8개

* 위치 : 경상남도
* 숙소유형 : 개인실
* 평점 : 4.95/5점
* 후기개수 : 222개

* 위치 : 경상북도
* 숙소유형 : 개인실
* 평점 : 4.75/5점
* 후기개수 : 55개

4 결과

의의

01

유저 데이터 없이
숙소 이미지만으로도
추천 시스템 구현 가능

02

피처 추출을 할 때
마지막 레이어를 사용하는
기존 방법들과 달리
초반 단계의 레이어 활용
가능성 제시

03

KNN과 코사인 유사도
두 가지 방법을 결합하여
새로운 추천시스템
가능성제시

04

가벼운 딥러닝 모델

+ 낮은 단계의
레이어 임베딩 사용

→ 실제 서비스 도입에
큰 이점

4 결과

한계점 ☹

01

숙소의 썸네일 이미지 한 장만 사용

02

사용한 데이터 양의 부족

03

정확한 평가지표의 부재

감사합니다