추천 시스템 3조 - 이미지 데이터 기반의 에어비앤비 숙소 추천 시스템

0. 기존 추천시스템의 문제점

- 기존의 협업필터링기반 추천시스템은 다음과 같은 문제점이 있었다.
    - 콘텐츠에 대한 이질적인 선호도와 사용자들의 객관적이지 못 하고 의사와 다 르게 평가하여 측정된 선호도에 대한 문제
    - 새로운 콘텐츠에 대한 평가가 이루어지지 않은 경우에 생기는 콜드 스타트(Cold-Start)문제
    - 방대한 양의 콘텐츠에 비해 사용자들에 의해 평가된 콘텐츠의 양이 적어 생기는 데이터 희박성(Sparsity)문제
    - 사용자가 많아지고 평가 데이터가 많아질수록 처리 속도가 급격히 느 려지는데 이를 해결해야하는 확장성 문제
- 기존 추천시스템은 많은 양의 유저 데이터(구매내역, 리뷰, 평점 등등)가 꼭 필요하다.
- 제품이나 유저의 특성에 대한 전반적인 설명을 파악하기 어려운 상황에서는 컨텐츠 기반 필터링의 추천 퀄리티가 매우 낮아진다는 문제가 있다.
- 신규 유저, 신규 숙소와 같이 데이터가 없거나 매우 적은 경우 좋은 추천을 할 수 없다. (**Cold Start)**
- **Cold Start 문제를 해결하기 위해 새로운 접근 방법**
  → **다른 데이터 필요 없이 숙소 이미지**만을 사용해 추천을 할 수 있는 방법 제시
- 에어비앤비를 하게된 이유
    - 숙박 사이트와 에어비앤비의 차이점 = 감성 숙소를 찾을 수 있다, 에어비앤비는 사람이 찍은 사진이다
    - 숙소는 이미지 기반 추천시스템이 흔치 않음으로 시범모델을 제안해보겠다.
    - 추천의 이유를 제시할 수 있다. 고객들이 추천 원인을 알아볼 수 있다.

1. 데이터

- 2021 국민 여행 조사 결과 참고하여 5개의 국내 여행지 선택

- 에어비앤비에서 데이터 스크래핑(각 지역 당 300개의 숙소)
- 전처리
  - 중복 숙소 삭제
  - 숙소의 썸네일 이미지가 숙소와 관련 없는 경우 삭제
  - 숙소타입, 평점, 리뷰 개수 등 전처리
- 최종 데이터

2. 이미지 피처 추출

- 모델 : resnet18
  - 선택 이유
    - ResNet은 Identify Mapping을 활용하여 기존 딥러닝의 문제점인 Vanish- ing Gradient 문제를 대응하여 아주 깊은 네트워크의 학습이 가능함을 확인시켜 주었다. 그림에서 보는 것처 럼 기본 구조의 출력에 다시 입력을 더해서 다음 레이어 로 넘어가도록 되어 있고, 이를 통해서 에러를 역전파하 는 과정에서 미분값이 상수로 남아서 계속 더해져서 값 이 작아지지 현상을 방지해 준다. 이러한 구조를 통해서 설계된 네트워크로 새로운 특징(Feature)을 추출하여 기존의 객체 분류나 검출의 구조의 특별한 변 화 없이 높은 성능 개선을 할 수 있었다.
    - 2015 **ILSVRC 에서 우승, 처음으로 사람보다 높은 정확도를 기록**
    - 많은 논문과 모델에서 응용되고 있는 모델
    - 이미지넷으로 pre-train된 모델 사용 가능
    - resent 중에서 가장 사이즈 작고 빠른 모델
    - 추가 근거
- 피처 추출할 레이어 선택
  - 이유
    - layer2 : 이미지 전체적인 정보를 학습
    - layer4 : 이미지에서 특정 물체에 대한 구체적인 정보를 학습
    - 결론 : 레이어2 위주로 진행했다.
- 최종 피처 추출 데이터

3. 추천 시스템 구성

- 선호님 군집화(15개)
  - 코사인 유사도 기반 추천
  - KNN으로 15개의 가장 유사한 숙소 추천
  - **KNN으로 15개 nearest neighbor 뽑은 후 그중에서 5개 cosine 유사도 높은 순으로 출력 (최종 모델)**
  - Kmeans사용하여 15개로 군집화 후 코사인유사도 정렬 = 내부 외부 결과 다른 경우가 있었다.
  - → 레이어2 or 레이어 2+3이 좋아 보임.
  - 숙소 정렬 시 하이브리드 사용 가능
- 진슬님 유사도 기반 클러스터링(15개)
  - 코사인 유사도 기반 클러스터링
  - 실루엣 계수 사용하여 군집화 결과 평가
  - 군집 별 숙소 개수 고르게 분포했는지(1개인 군집이 많았는지) → 레이어2에서 군집수 15의 실루엣계수가 가장 좋았다
  - 레이어 조합별 클러스터링 결과 확인 → 레이어2 or 레이어 2+3이 좋아 보임.
- 레이어 선택 이유
  - 기준
    - 15개를 뽑았을 때 내부 외부, 색감을 가장 잘 군집화 했던 분류 했던 레이어
    - 기준1 내부 외부
    - 기준2 색감 : 가장 많은 영역을 차지한 색 3개 추출해서 2개 이상 색이 동일하면 동일하다고 판단.
- 진슬님+선호님 분석 결과 = 레이어2 or 레이어 2+3이 좋았다.
- 후처리(평점, 리뷰 개수 활용)
  - 1개 이미지 → 5개 숙소(이미지, 숙소 링크, 지역, 숙소 유형, 평점, 리뷰 개수)
- 최종 추천

4. 평가 및 최종 모델 선정

- 평가 지표
- 모델 평가 및 비교
- 최종 모델 선택

5. 결론(의의와 한계)

- 결론
  - 유저 데이터 없이도 추천 시스템 구현 가능
  - 숙소의 이미지 데이터만 있다면 다른 숙소 추천 가능
  - 신규 유저, 신규 숙소에 대해서도 추천 할 수 있다.
- 한계점
  - 숙소의 썸네일 이미지 한 장만 사용했다.
  - 숙소와 상관 없는 이미지의 숙소는 제외되었다.
  - 사용한 데이터 양의 부족(?)

레이어 선택하게된 이유

- 외부 사진, 내부 사진
  - 외부 사진 넣었을 때 외부 사진의 숙소가 추천되는지 체크
    ex) 5개의 추천 숙소 중에서 5개 모두 외부 사진이면 만점
  - 외부 사진 넣었을 때 내부 사진의 숙소를 추천 → 패널티
    ex) 외부 사진 넣었을 때 외부4개 내부1개 숙소 추천되면 5점 만점에 4점
- 사진간의 컬러 비교
- 이미지내 동일한 사물 체크
  - image net으로 학습된 object detection으로 사물 개수 체크
    - 모델 후보1: yolov5(coco)

        - 모델 후보2: **ONNX-ImageNet-1K-Object-Detector(이미지넷)**
          https://github.com/ibaiGorordo/ONNX-ImageNet-1K-Object-Detector
        - 사람이 직접한다.
            - 사물 리스트
    - 몇개의 사물이 동일한지 점수화

모델 평가 방법

- 외부 사진, 내부 사진
    - 외부 사진 넣었을 때 외부 사진의 숙소가 추천되는지 체크
      ex) 5개의 추천 숙소 중에서 5개 모두 외부 사진이면 만점
    - 외부 사진 넣었을 때 내부 사진의 숙소를 추천 → 패널티
      ex) 외부 사진 넣었을 때 외부4개 내부1개 숙소 추천되면 5점 만점에 4점
- 사진간의 컬러 비교
- 이미지내 동일한 사물 체크
    - image net으로 학습된 object detection으로 사물 개수 체크
        - 모델 후보1: yolov5(coco)
        - 모델 후보2: **ONNX-ImageNet-1K-Object-Detector(이미지넷)**
          https://github.com/ibaiGorordo/ONNX-ImageNet-1K-Object-Detector
        - 사람이 직접한다.
            - 사물 리스트
    - 몇개의 사물이 동일한지 점수화