빅데이터 분석 정의서

주제 : 비문 데이터를 활용한
유실 반려견 서칭 서비스

2021.00.00

Ⅰ. 개요

1. 아이디어 주제
   : 비문 데이터를 활용한 유실 반려견 서칭 서비스(찾아줄개)
2. 개발 목표
   : 견주에게 잃어버린 반려견을 찾아주는 앱 서비스
3. 개발 내용
   : 유실 반려견 얼굴 매칭 서비스(견종, 털색상)
   : 유실 반려견 비문 매칭 서비스

Ⅱ. 기능별 빅데이터 분석 명세서

기능명	유실 반려견 얼굴 매칭 서비스(견종, 털색상)
1. 데이터 준비	
데이터 정의	총 15가지 견종의 강아지 사진 데이터 (비숑, 보더콜리, 치와와, 차우차우, 골든래트리버, 허스키, 말티즈, 포메라니안, 푸들, 시바, 시츄, 웰시코기, 요크셔테리어, 닥스훈트, 진돗개, 믹스견)
데이터 획득 방법	1. 강아지 사진 구글 이미지에서 각 견종별 약 1000장씩 크롤링 2. Stanford Dogs Dataset 2만개 (http://vision.stanford.edu/aditya86/ImageNetDogs/) ** 데이터 수집 한 링크 꼭 넣을 것
2. 전처리	
전처리 과정	Keras의 Imgagegenerator를 사용해서 이미지 회전 및 이동을 통해 이미지 데이터 6배 증강(각 1000개 -> 6000개, 총 9만개)
3. 모델 생성 및 학습	
모델링 목표	견종 예측, 털 색상 예측
모델링 가능 알고리즘	견종 예측 : CNN 모델, VGG16 다중 분류 모델
	털 색상 예측 : Dlib, CNN, K-means 알고리즘, 수학적 거리 계산 공식
학습	견종예측 : Convolution층을 전이학습
	털 색상 예측 : 1. Dlib과 CNN을 사용해서 강아지의 얼굴 추출 2. K-means Color Clustering 알고리즘을 사용해서 사진의 평균 색상을 추출 후 가장 높은 비율의 색상 선택 3. 수학적 거리 계산 공식을 통해 검정색, 흰색, 갈색 중 대표색의

	RGB값과 가까운 RGB값 계산
4. 검증	
모델링 검증	15가지 견종의 강아지 이미지 데이터를 추가적으로 약 200장 정도 확보 후 견종 예측 모델 검증
모델링 평과 결과	견종 예측 - CNN 모델 : 0.08% 정확도 - VGG16 다중 분류 모델 : 87% 정확도

기능명	유실 반려견 비문 매칭 서비스
2. 데이터 준비	
데이터 정의	총 15가지 견종의 강아지 사진 데이터 (비숑, 보더콜리, 치와와, 차우차우, 골든래트리버, 허스키, 말티즈, 포메라니안, 푸들, 시바, 시츄, 웰시코기, 요크셔테리어, 닥스훈트, 진돗개, 믹스견)
데이터 획득 방법	1. 강아지 사진 구글 이미지에서 각 견종별 약 1000장씩 크롤링 2. Stanford Dogs Dataset 2만개 (http://vision.stanford.edu/aditya86/ImageNetDogs/)
2. 전처리	
전처리 과정	1. Stanford Dogs Dataset로 강아지 코를 학습한 YOLOv5를 사용해서 코 부분만 잘라내서 저장 2. 코의 주름을 잘 보이게 하기 위해 어두운 부분을 밝게 펴주는 이미지 전처리 : CLAHE(대비 제한 적응 히스토그램 평활화) 사용
3. 모델 생성 및 학습	
모델링 목표	비문 특징 추출하여 일치하는 비문 데이터 찾기
모델링 가능 알고리즘	비문 특징 추출 : SIFT(Scale-Invariant Feature Transform)
	추출한 특징 벡터화 : K-means 클러스터링
	일치하는 비문데이터 찾기 : SVM모델
학습	비문 특징 추출 : - SIFT를 사용하여 이미지의 크기를 확대하거나 축소 - 검출되는 특징 수(nfeatures) 200개 - 이미지 피라미드 사용할 계층 수(nOctaveLayers) : 3 - 빈약한 특징은 거르기 위해 문턱 값(contrastThreshold) : 0.0005

	추출한 특징 벡터화 : - 이미지크기에 따라서 특징이 존재하는데 사용자들에게 많은 사진을 요구할 수 없어서 입력 차원을 줄여주는 전처리 진행 - K-means 클러스터링을 사용하여 100차원 벡터로 변환하여 분류
	일치하는 비문 데이터 찾기 : - SVM모델을 활용 - 기존의 비문데이터와 사용자가 등록한 비문 데이터의 확률값을 비교하여 svm.predict_proba 값이 0.7 이상일 경우에 동일한 비문이라고 판단
4. 검증	
모델링 검증	- 검증용으로 강아지 이미지 데이터를 추가로 확보(약 200장) - K-fold 교차검증 법을 사용하여(k=5) 모델 검증 진행
모델링 평과 결과	- 오차 행렬(Confusion matrix)를 사용하여 평가 진행 - 견종별로 0.7 이상으로 동일한 비문으로 일치하는 정확도는 0.735