빅데이터 2024 2학기 프로젝트 계획서

한국외국어대학교 경영학전공 202102858 임산별

1. 프로젝트 주제 : 메신저 채팅 짤 이미지 추천 기능 기획

   1.1 선정 동기

   1.2 최근 MG세대에서는 짤(meme) 이미지를 활용한 소통이 증가하고 있음. 실제 기업도 이들의 관심을 끌기 위해 짤을 활용해 적극적인 마케팅과 콜라보 상품 출시 등을 진행 중임. 이와 같은 트렌드를 반영하여 메신저 채팅에서 사용자가 보다 쉽게 짤을 활용할 수 있도록 짤 추천 서비스를 기획하게 됨.

   1.3 서비스 기획 내용

   - 이용자가 텍스트를 입력하면, 텍스트 내 단어에 적합한 짤 이미지들을 추천해줌(예: '웃겨'입력 -> 웃는 사진의 짤 추천)
   - 이용자는 이모티콘 리스트 중에 원하는 이미티콘을 선택하는 것 처럼 원하는 이미지를 선택해서 전송 가능함.

2. 데이터 수집

   데이터 수집 출처 : Pinterest(사진 추천 및 포스팅 사이트)

   데이터 수집 방법 : 자체 수집(특정 단어를 검색해 해당하는 이미지들 크롤링)

   - Selenium,ChromeWebDriver을 이용해 특정 키워드(예: "무한도전")로 이미지를 크롤링함(약 1000만건)
   - 특정 단어 : 짤이 많이 존재하는 밈과 유행어 100개 사전을 구축(자체 수집)
   - 각 단어별 10만개 사진 수집(빠른 크롤링을 위해 5개 사이트로 병렬식 진행/총 1000만개 사진 수집)

3. 전처리 계획

   3.1 이미지 전처리

   - 수집된 단어별 이미지에 해당 단어로 라벨링
   - 중복되는 이미지, 화질이 낮은 이미지 제거
   - 수집된 이미지가 500개 미만인 단어들 제거

4. 분석 계획

   4.1 이미지 분류 모델 구축

   - 머신러닝의 전이학습 CNN(Convolutional Neural Network)을 사용해 라벨별 특징을 파악함
     * 전이학습 사용이유 : 짤 이미지의 특성상 텍스트,곡선,색감 이외에 감정을 표현하는 이미지가 많음
   - 이미지 벡터를을 사용해 K-means클러스터 알고리즘을 활용해 이미지 그룹화 -> 이상치 제거

   4.2 이용자 입력 단어 유사한 형태로 변경해 유행어 사전에 대치

   - 이용자가 입력하는 단어 전처리(토큰화, 정규화 / 짤(meme)특성상 Konlpy같은 기존 사전에 등록되지않은 불용어가 존재할 가능성이 높기 때문에, 불용어,기호,외국어 제거하지 않음)

   - Word2Vec, FastText를 사용하여 단어들을 벡터 형태로 변환(임베딩)

   - 코사인 유사도, dit Distance, Jaccard 유사도 등을 사용해 유사도 계산(세개 중 성능 높은 것 선택)

   - Nearest Neighbor Search를 이용해 유사한 단어 탐색 -> 해당 단어의 이미지 추천

5. 시각화 계획

   - 집단별 이미지 특성 이해를 위해. T-SNE 차원 축소방법을 사용해 이미지 벡터 시각화
   - Confusion Matrix를 사용해 모델의 분류 성능 시각화
   - Triplet loss 를 사용해 이미지 유사도 측정-> 네트워크 그래프를 통해 시각화
   - 사용자 인터페이스 대시보드로 서비스 이용 예시 시각화

6. 레퍼런스 및 참고자료 출처

   - 토이프로젝트
   - 인텔 이미지 분류