**실무 프로젝트 기획안**

**2020년 11월 13일**

**과정명: 서비스 산업 데이터를 활용한 빅데이터 분석 실무**

|  |  |
| --- | --- |
| 팀 명 | 악!플원정대 |
| 팀 원 | 최호진(팀장) , 김선림 , 김한준 , 윤선영 , 이재원 , 최가은 |
| 프로젝트 타이틀 | 악성댓글 탐지봇 |
| 프로젝트 주제  및 내용  해결하고자 하는 문제  최종 산출물의 청사진 | * 머신러닝 혹은 딥러닝을 활용한 “악성댓글 탐지봇” 개발 * NLP를 활용한 악성댓글 분류 서비스를 Web에 구현 |
| 팀원간 역할  분담 및 일정 | * 주제 선정 : 팀 전원 * 기획안 및 WBS 작성 : 팀 전원 * 데이터 수집 * 인플루언서 인스타 계정 댓글 크롤링 : 이재원, 최가은 * 일간베스트 제목 및 댓글 크롤링 : 이재원 * 디씨인사이드 댓글 크롤링 : 김한준 * 에브리타임 댓글 크롤링 : 윤선영 * 데이터 전처리 * 댓글 데이터 라벨링 : 팀 전원 * 댓글 텍스트 토큰화 : 이재원, 최가은 * 댓글 텍스트 고유 정수 라벨링 : 이재원, 최가은 * 댓글 텍스트 패딩 : 이재원, 최가은 * 모델링 * 딥러닝 : 최호진, 이재원, 김선림 * 머신러닝 : 김한준, 윤선영 * 웹 구축 : 이재원, 김선림 * 발표 준비 * PPT 제작 : 김선림 * 시연 및 수정 : 팀 전원 * 최종 발표 : 발표자 미정 |
| 프로젝트 수행 방향  수행 방법/도구 | * 데이터 수집 : Selenium , BeautifulSoup 을 활용한 소셜미디어 및 웹사이트 댓글 크롤링 * 댓글 데이터 전처리 :  1. 텍스트 데이터 토큰화  * 특수문자와 같은 불용어를 제거한다. (악플을 분류하는 것이 목표이기에 특수문자와 같은 문자를 불용어로 지정하여 제거한다.) * 특수문자가 제거된 텍스트 데이터를 음절 단위로 토큰화 한다. 음절 단위로 토큰화하기 때문에 꼬꼬마와 같은 형태소 분석기를 사용하기보다는 파이썬 내장 함수로 음절을 쪼갠다.  1. 토큰화된 텍스트 데이터 고유정수 라벨링  * Tensorflow.keras가 제공하는 Tokenizer함수를 사용하여 음절 단위로 토큰화된 텍스트에 고유정수를 부여한다.  1. 토큰화와 고유정수 라벨링이 완료된 데이터 패딩  * 토큰화와 고유정수 라벨링이 완료된 데이터를 tensorflow.keras가 제공하는 pad\_sequence함수를 사용하여 행렬화한다. 여기서 행렬화는 텍스트마다 가지고 있는 토큰의 개수(길이)가 다르기 때문에 의미 없는 고유 정수인 0을 이용하여 댓글 데이터의 길이를 동일하게 맞추고, 데이터프레임(행렬)화하는 것이다. * 모델링 * 딥러닝 : 전처리한 데이터를 활용해 앞뒤 문맥을 고려해주는 RNN계열(LSTM) 딥러닝을 주 모델로 학습, 악성 댓글 여부를 예측한다. * 머신러닝 : 비교적 성능 좋은 모델로 평가되는 최신 기법인 lightGBM 계열 머신러닝과, 감성분석, 텍스트 데이터 분석에 자주 사용되는 Naïve Beyes, 그리고 Support Vector Machine과 같은 기법을 사용할 예정이다. * 웹 구축 : 장고를 활용해 악성 댓글 탐지 기능을 구현한 웹을 구축한다. |