딥러닝 기반 비속어 필터링

- 9주차 -

2021202057 황세영 2021202058 송채영



Doing 진행사항

[9주차]

- 8주차에 중간 계획서 발표를 진행하였고, 발표를 통해 피드백을 받음
- 9주차에는 피드백을 받고 앞으로의 개선 방향과 추가해야 할 점을 회의함
 - + 욕을 변형하여 사용하는 경우에 비속어 처리를 어떻게 할지
 - + 맞춤법을 틀리게 작성하거나 중간에 다른 글씨를 섞어서 쓰는 등의 방법을 사용할 때는 어느 정도까지 필터링 할지
 - + 비속어가 계속해서 만들어지고 있는데, 신종 비속어들을 어떻게 처리할 지에 대한 대책이 필요
 - + 비속어 필터링 같은 경우에는 CNN보다 더 좋은 Bert나 LSTM 같은 모델들이 존재하는데 CNN을 사용하는 이유
- 임베딩 기법으로 분류하는 과정에서 Type error가 났지만, 해결하지 못 함

To Do 앞으로 할 것

딥러닝 기반 비속어 필터링 에브리타임, dcinside, Instagram 게시물, youtube, 뉴스 등 데이터 크롤링으로 예외처리로 비속어로 분류될 수 있 자주 방문하는 사이트 내 댓글을 통해 단어(data) 획득 으나 비속어가 아닌 데이터 찾기 욕설의 표준, 기준이 되는 것들을 얻을 예정 sst 변환을 통해 유사 텍스트 변환을 통해 유사성 비교와 단어 학습(라벨링) 비속어인지 아닌지를 비교함 단어를 임베딩 기법으로 분류해 Word2Vec or fastttext 딥러닝 모델의 입력값으로 사용 비속어로 추출된 어절을 중심으로 앞 뒤 단어를 trigram으로 반환해 벡터화함 2 class로 나누어 벡터화한 데이터를 비속어인 경우 0. CNN 모델에 넣음 비속어가 아닌 경우 1 그래드 캠과 shap을 데이터를 시각화함

[10주차]

- 획득한 데이터를 임베딩 기법으로 분류 [10주차 ~ 12주차]

- 벡터화, CNN모델에 넣기, 출력, 데이터 시각화 진행 [13주차]

- 프로젝트 최종 보고

Done 한 것

- 주제선정
- 데이터 획득

(에브리타임, dcinside, Instagram 게시물, youtube, 뉴스 등 자주 방문하는 사이트 내 댓글을 통해 욕설의 표준, 기준이 되는 것들을 얻음)

- + 데이터 획득(5000개 이상)
- -> 띄어쓰기가 포함된 비속어 데이터를 수집함 (ex, 아저씨 발냄새 나요, 해바라기씨 발아, 무지개 같은 사장님, 곧 휴가 철이네 등)
- -> 추가 데이터 네이버 금칙어 데이터, 인스티즈 금칙어 데이터를 수집함
- + 딥러닝 모델의 입력값으로 사용하기 위해 임베딩 기법으로 분류하기를 진행 중
- -> 완성 x(오류 수정 필요)