8주차 수행내역

2017103758 조문기

지난 2주간 수행했던 내역을 정리합니다.

10/16

새롭게 모은 데이터들을 바탕으로 FastText 모델을 새로 만들었습니다. 모델 만들 시 이전에 만든 모델에 오류가 있어 수정을 했습니다.

10/17

새로운 FT 모델을 바탕으로 BiLSTM+1DCNN 모델을 학습시켜서 정확도를 측정했습니다.

일련의 과정을 함수화 시켜 코드를 깔끔하게 정리했습니다.

10/18

데이터 불균형 문제를 해소하고자 oversampling 기법들에 대해 탐구해 봄.

SMOTE, K-neighbourshood 방법?

10/20

FastText 상의 단어간 유사도 순위를 이용해 비슷한 단어를 파악, 이를 활용해 단어를 바꾸는 방식 채용 -> 욕설 레이블의 데이터를 일반 댓글 레이블의 수와 비슷하게 맞춤.

BiLSTM + Attention 채용

새로운 FT 모델 적용 이후 계속 '안녕하세요'를 욕설로 구분하는 이슈 발생. 왜일까? FT모델이 잘못된 걸까? 그러기엔 FT모델 상에서 most_similar 메서드를 돌려보면 안녕의 비슷한 단어는 잘만 나온다. train data 셋에 문제가 있는 것은 아닐지 확인해볼 필요가 있을 듯.

이제는 데이터 상태와 모델링 기법을 어떤 기법을 써야할 지에 대해 생각해 볼 때인 것 같다.

10/21

FT모델 만들 때 불용어를 따로 지정하지 않고, Mecab의 pos-tagging중 조사, 용언, 접미사 등등을 제외한 욕설과 가까운 품사(명사, 동사 등)만 남기고 나머지 제외하는 방식을 사용. 이를 통해 욕설에 영향을 줄 수 있는 품사들만 남길 수 있도록 다시 조절했음.

학습했던 자료들은 개인 GitHub에 꾸준히 업로드 중입니다. 단순히 적용중인 모델들은 GitHub Repository의 Practice 폴더에 업로드 중이고, 실제로 적용될 것 같은 학습 내용들은 Works 폴더에 업로드 중입니다.

GitHub Link: https://github.com/siryuon/DataAnalysis_CapstoneDesign