SW설명서

1. Code

* 노트북 실행 파일(.ipynb)이 담겨 있습니다. 전처리 코드는 아래 순서대로 실행해야함.

1. 전처리 코드
   * Preprocessing\_data.ipynb: 기존에 있는 4차년도.csv, 5차년도.csv, 5차년도\_2차.csv 세개의 데이터를 concat하고 전처리하는 코드입니다. processed\_sentiment.csv 파일 생성. 이 파일은 CNN.ipynb, LSTM.ipynb에서 사용됩니다.
   * Preprocessing\_new\_data.ipynb: aihub에서 가져온 새로운 데이터 senti\_training.xlsx, senti\_validation.xlsx 파일을 전처리하는 코드입니다. new\_processed\_sentiment.csv 파일을 생성.
   * Preprocessing\_total\_data.ipynb: processed\_sentiment.csv, new\_processed\_sentiment.csv 두가지 파일의 데이터를 concat하고 전처리합니다. 또 anxious, embarrassed, hurt 클래스에 대한 데이터를 따로 분리합니다. 세가지 클래스에 대한 샘플은 추후에 모델을 통해 다시 레이블링 됩니다. Total\_sentiment.csv, label\_anxious.csv, label\_embarrassed.csv, label\_hurt.csv 파일 생성.
   * Label\_anxious.ipynb, Label\_embarrassed.ipynb, label\_hurt.ipynb: CNN 모델을 통해 다시 레이블링 합니다. 그리고 각 감정 레이블에 맞게 적절히 수정하는 작업이 있습니다.(ex) ‘행복’으로 레이블링 된 샘플을 ‘두려움’으로 재레이블링) 각 relabel\_anxious.csv, relabel\_embarrassed.csv, relabel\_hurt.csv 파일 생성.
   * Processing\_end\_data.ipynb: total\_sentiment.csv, relabel\_anxious.csv, relabel\_embarrassed.csv, relabel\_hurt.csv 4가지 데이터를 concat하고 전처리한다. End\_sentiment.csv 파일 생성.
2. 모델 코드
   * CNN.ipynb: 기존에 있는 4만개 샘플(processed\_sentiment.csv)을 사용하는 CNN 모델. Confusion matrix를 시각화한 이미지, tokenizer, 학습된 최종 model 세가지 파일 생성.
   * LSTM.ipynb: 기존에 있는 4만개 샘플(processed\_sentiment.csv)을 사용하는 CNN 모델. Confusion matrix를 시각화한 이미지, tokenizer, 학습된 최종 model 세가지 파일 생성.
   * CNN\_aug.ipynb: 최종적으로 전처리된 12만개 샘플(end\_sentiment.csv)을 사용하는 CNN 모델. Confusion matrix를 시각화한 이미지, tokenizer, 학습된 최종 model 세가지 파일 생성.
   * LSTM\_aug.ipynb: 최종적으로 전처리된 12만개 샘플(end\_sentiment.csv)을 사용하는 CNN 모델. Confusion matrix를 시각화한 이미지, tokenizer, 학습된 최종 model 세가지 파일 생성.
3. Data

* Aihub\_sentiment 폴더 이하: 새로 가져온 데이터. 27만개 정도의 말뭉치가 있다.

1. model