[6회차] NLP 과제

발제자 : 이준찬

기간

마감: 7월 31일 목요일 23시 59분

지각 제출: 8월 1일 금요일 23시 59분

Intro

안녕하세요! 이번 과제는 개별로 수행해주시면 됩니다.

이번과제는 2단계로 구성됩니다!

- 1. Corpus 에서 Word2Vec을 학습해 단어 embedding 완성하기
- 2. 1에서 구한 embedding을 활용해 GRU를 구현하여 감정분석하기

명세

제공되는 파일

- config.py
- gru.py
- load_corpus.py
- model.py
- pyproject.toml
- requirements.txt
- test.py
- train_model.py
- train_word2vec.py
- word2vec.py

[6회차] NLP 과제 1

작성할 파일

- config.py
- gru.py
- train_model.py
- load_corpus.py
- word2vec.py

1. 데이터셋 확인

- dataset 링크
- 2. word2vec 구현 및 학습
 - word2vec.py 의 Word2Vec Class를 구현하세요! 이때, 생성자의 method 에서 넣어 준 값에 따라 _train_cbow 나 _train_skipgram 을 호출해야합니다. 그래서 두가지 모두 구현해야합니다!

(주의: Word2Vec을 훈련할 때는 padding token이 들어가지 않는 게 좋습니다!)

- load_corpus.py 의 load_corpus 를 구현하세요! load_corpus는 말 그대로 word2vec을 학습시킬 corpus를 가져오는 함수입니다. corpus를 어디서 어떻게 가져오실지는 자유입니다! 단, 사용할 수 있는 라이브러리는 Python 기본 라이브러리, torch, transformers, datasets로 제한됩니다. 또, 제출된 코드를 실행했을 때접근 불가능한 파일로부터 가져오는 게 있으면 안 됩니다. 니다. (코드가 1. 인터넷에서 파일을 로컬로 저장하고 2. 그 파일을 로컬에서 불러오
 - 니다. (코드가 1. 인터넷에서 파일을 로컬로 저장하고 2. 그 파일을 로컬에서 불러오는 과정을 가진다면 괜찮겠지만 그냥 로컬 파일을 불러오기만 한다면 안되겠죠?)
- train_word2vec.py 를 실행하시면 word2vec.pt 체크포인트 파일이 생성됩니다. 이 파일은 우리의 GRU 모델을 학습시키는 데 사용됩니다.

3. GRU 구현하기

• gru.py에 GRUCell 과 GRU 를 구현하세요! 힌트를 위해 input size를 드리겠습니다!

GRUCell.forward:

- x: (batch_size, d_model)
- h: (batch_size, d_model)

GRU.forward:

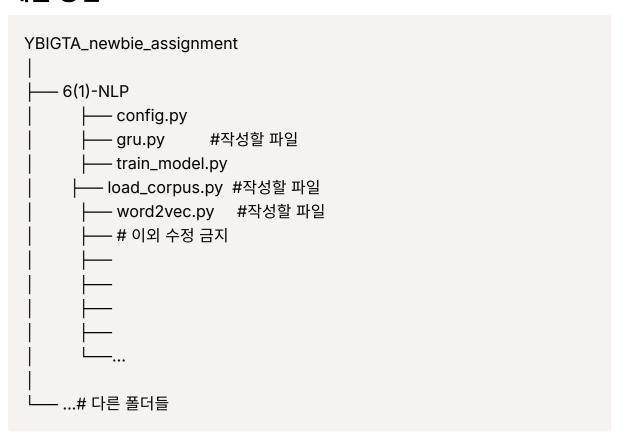
[6회차] NLP 과제

inputs : (batch_size, sequence_length, d_model)

구현하신 GRU 는 model.py 의 MyGRULanguageModel 이 사용합니다. 이 모델이 sentiment analysis를 위해 train_model.py 에서 학습됩니다.

train_model.py 를 실행하시면 checkpoint.pt 체크포인트 파일이 생성됩니다. 학습된 MyGRULanguageModel 의 체크포인트이며, test.py 에서 최종 test를 할 때 사용됩니다.

제출 방법



채점 기준

- ☐ test macro 0.3 이상
- ☐ mypy 통과



<u>과제 미흡:</u> 두 조건 중 하나만 충족 (1/2)

<u>과제 미제출:</u> 두 조건 모두 충족 못함 (0/2)

[6회차] NLP 과제