AIM5002 Homework 1: Korean Emotion Recognition

JinYeong Bak (jy.bak@skku.edu)

Deadline: Oct 23 2022, 23:59*

1 Introduction

과제 설명서를 잘 읽고 따라주세요. 이 과제는 팀 과제가 아닌 개인 과제로 academic integrity를 위반하셔서는 안 됩니다. 만약 그럴 경우에는 확인 후 문제가 있음으로 판정될 경우 F 성적이 나가게 됩니다.

2 Goal

이번 과제에서는 한국어 감성 분석을 진행하는 것입니다. 공개된 데이터를 살펴보고 가공한 후 한국어 감성 분석 모델을 학습하여 그 성능 확인을 진행할 것입니다. 추가적으로 실제 자연어 데이터 및 문제들을 살펴보고 그에 대한 프로젝트 계획 및 개선점 등에 대해 논의할 것입니다.

3 Task 1

첫 번째 task는 한국어 감성 데이터를 가공하는 것을 목표로 두고 있습니다.

먼저, AIHub[1]에 공개되어진 한국어 감성 데이터인 '감성 대화 말뭉치' 데이터를 다운로드 받습니다.

페이지 URL: `https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=86`

해당 페이지에서 파일을 다운로드 받으면 크게 'Training'과 'Validation' 이름의 두 개의 폴더가 있습니다. 그리고 각 폴더 안에는 '원천데이터'와 '최종데이터'로 나눠져있는 것을 확인하실 수 있습니다. 이번 과제에서는 '최종데이터'를 사용하도록 하겠습니다. 해당 파일들의 압축을 풀면 xlsx와 json 파일을 얻을 수 있습니다. 이 중 raw data인 json 파일을 통해 우리가 원하는 감성 분류 문제를 풀 데이터를 추출하도록 하겠습니다.

감성 대화 말뭉치는 사람과 시스템의 대화들의 뭉치입니다. 그리고 대화 처음에 사람의 말에서 오는 감정과 페르소나 등의 라벨이 담겨져있습니다. Table 1은 그 하나의 예시를 보여주고 있습니다. 사람과 컴퓨터의 대화가 이어지고 첫 번째 사람의 감정에 대해 'E68'이라는 감정 레이블이 있습니다. 우리는 이러한 데이터에서 사람의 말과 감성만을 추출하여 감성 분석을 진행하고자 합니다. 데이터에 대한 자세한 설명은 해당 페이지에서 확인하실 수 있습니다.

기존 데이터에서는 감성에 대해 60가지로 분류하였습니다. 이 중 우리는 여섯 가지의 감성 대분류만을 이용할 것입니다. 해당 감성은 분노, 슬픔, 불안, 상처, 당황, 기쁨이 있습니다. 기계 학습 모델은 이러한 감성 레이블을 텍스트로 바로 이해를 하지 못합니다. 따라서 각각의 감성을 숫자에 매칭시켜 최종 출력 파일에 표시해야합니다. 이번 과제에서 분노는 0번, 슬픔은 1번, 불안은 2번, 상처는 3번, 당황은 4번, 기쁨은 5번으로 정하도록 하겠습니다.

이번 task에서 여러분들은 데이터를 추출하는 프로그램으로 'Preprocessing.py' 파일을 완성시켜야합니다. 총 3개의 task가 해당 파일 안에 있으며 이부분을 채워서 최종적으로 세 개의 파일(train.csv, valid.csv, test.csv)을 만드시는 것이 최종 목표입니다. Table 2은 실제 train.csv 파일 내용 일부를 보여주고 있습니다.

아래 step에 맞춰 task를 수행해주시면 됩니다.

3.1 Steps

1. AIHub에서 '감성 대화 말뭉치'를 다운로드 받습니다.[1]
2. Training 폴더와 Validation 폴더 내 '최종데이터'를 선택하여 압축을 풉니다.
3. json 파일들을 'Preprocessing.py' 파일이 있는 폴더로 옮깁니다.

*질의응답은 다른 학생들도 볼 수 있도록 icampus 문의게시판을 이용하시기를 권장합니다. 코드를 직접 올리지는 말아주세요.

[1]다운로드에는 회원가입 및 본인인증이 필요합니다.

```
{'profile': {'emotion': {'emotion-id': 'S05_D02_E68',
                         'situation': ['S05', 'D02'],
                         'type': 'E68'},
             'persona': {'computer': ['C01'],
                         'human': ['A02', 'G01'],
                         'persona-id': 'A02_G01_C01'},
             'persona-id': 'Pro_03719'},
 'talk': {'content': {'HS01': '아내가 드디어 출산하게 되어서 정말 신이 나.',
                      'HS02': '아 지금 정말 신이 나.',
                      'HS03': '아기가 점점 클게 벌써 기대가 되네. 내가 많이 놀아줘야지. ',
                      'SS01': '아내분이 출산을 하시는군요. 정말 축하드려요.',
                      'SS02': '잘 된 일이네요.',
                      'SS03': '좋은 아빠가 되실 거 같아요. 진심으로 축하드려요.'},
          'id': {'profile-id': 'Pro_03719', 'talk-id': 'Pro_03719_00016'}}}
```

Table 1: 감성 대화 말뭉치 데이터 예시. json 파일에서 추출함.

```
sentence,label
직장에 새로운 신입사원이 입사를 했는데 알려줄게 너무 많아.,0
집 앞에 슈퍼를 갈 때도 나를 자꾸 데려가.,0
아내가 드디어 출산하게 되어서 정말 신이 나.,5
```

Table 2: train.csv 파일 내용 일부.

4. 'Preprocessing.py'을 완성시킵니다.
    - 'extract_data' 함수를 완성시킵니다.
    - 'save_csv' 함수를 완성시킵니다.
    - Training 폴더에서 가져온 데이터를 training 데이터와 validation 데이터로 나눕니다. 이 때 'train_test_split' 함수를 이용합니다.
5. 프로그램을 실행하여 세 개의 파일(train.csv, valid.csv, test.csv)을 만듭니다.

4 Task 2

두 번째 task는 앞서 가공한 한국어 감성 데이터를 가지고 실제 감성 분류 모델을 학습시키고 그 성능을 확인하는 것을 목표로 두고 있습니다.

이번 과제에서는 한국어 사전 학습 언어 모델인 KLUE-RoBERTa[2]를 이용할 것입니다. 이를 통해 한국어 감성 분류 모델을 만들고 학습시킨 후 그 결과를 살펴보도록 하겠습니다.

페이지 URL: `https://colab.research.google.com/drive/15-nCkJUgN-WMtvnS4Q4EZEt3nf2pmBzg?usp=sharing`

위의 페이지에 가시면 과제 스켈렉톤 코드가 있습니다. 코드에는 빈 칸이 있으며 앞서 제작한 파일이 아닌 다른 파일이 올려져있습니다. 아래 step에 맞춰 task를 수행해주시면 됩니다.

4.1 Steps

1. Google Colab 페이지에 접근한 후 본인의 구글 계정으로 복사합니다.
2. Task 1에서 생성한 파일 세 개를 본인 구글 드라이브에 올린 후 링크 URL을 얻습니다.
3. Colab 페이지 안의 task들을 수행하여 완성시킵니다.
    - 올린 데이터들을 다운로드 받는 코드를 완성시킵니다.

- 'RoBERTaClassifier'를 완성시킵니다.
- 'KERDataset'를 완성시킵니다.

4. 프로그램을 완성시킨 후 실행하여 그 결과까지 모두 페이지에 출력이 되도록 합니다.
5. 완성된 코드와 결과까지 나온 상태에서 'ipynb' 파일로 다운로드 받습니다.
6. 코드를 살펴보며 문제점/아쉬운 점들에 대해서 논하세요. 아래 항목을 수행하면 추가 점수가 있습니다.
   - 해당 문제점에 대한 해결 방안을 제시하세요.
   - 해결 방안을 실제로 적용한 코드 작성 및 추가 결과를 첨부하세요.

Pytorch에 익숙하지 않으신 분들은 아래 페이지들을 참고하시면 도움이 되실 것입니다.

- `https://pytorch.org/tutorials/beginner/text_sentiment_ngrams_tutorial.html`
- `https://github.com/JH-lee95/Korean-Sentiments-Classification`

5 Task 3

이번 태스크는 데이터에 대한 특징을 분석 및 확인하는 것을 연습해보도록 하겠습니다.

인공지능(AI) 모델 학습에 있어 데이터는 매우 중요합니다. 어떠한 문제를 인공지능으로 풀고자 할 때 학습을 위해 관련된 데이터를 준비해야하죠. 이 때 살펴봐야할 것들은 아래와 같습니다.

- 데이터셋 충분성 - AI 모델 학습에 그 양이 충분한가?
- 데이터셋 편향성 - AI 모델 학습에 있어 일부 영역만을 가지진 않았는가?
- 데이터 정보 포괄성 - 학습에 유용한 특성(feature) 정보가 충분히 제공되고 있는가?
- 데이터 정보 변동성 - 제공하는 특성에 다양한 범위의 샘플이 확보되어 있는가?
- 원시 데이터[2] 신뢰성 - 원시 데이터의 정확성, 완전성, 일관성이 잘 갖추어졌는가?
- 라벨링 데이터[3] 신뢰성 - 라벨링 데이터의 정확성, 완전성, 일관성이 잘 갖추어졌는가?
- 개인정보 보호 - 개인정보 비식별화 등 민감정보 보호가 적절하게 처리되었는가?
- 인공지능 학습용 데이터 품질 개선을 위한 기타 검토사항

앞선 tasks에서 살펴본 '감성 대화 말뭉치'를 위의 검토 사항으로 한 번 살펴보세요. 단순히 데이터를 보는 것에서 그치지 말고 프로그램 코드를 작성하여 검토 사항을 확인해보시길 바랍니다. 예를 들어 데이터의 양이 얼마나 되는지, 데이터 내 고유한 단어 개수(사전 크기)는 얼마인지, 라벨의 개수와 그 분포는 어떠한지 등을 확인할 때 프로그래밍이 필수적일 것입니다. 특히 데이터를 직접 살펴봄으로써 만약 내가 이 데이터로 학습한다면 혹은 테스트한다면 어떤 점을 살펴봐야하는지를 생각하시면 좀 더 이해하시기 쉬우실 것입니다.

5.1 Steps

- 위의 검토 사항에 맞춰 '감성 대화 말뭉치' 데이터의 특징 및 개선점 등을 답변해주세요.

6 Submission and Evaluation

아래 파일들을 제출해주세요.

- Preprocessing.ipynb - Task 1을 완성한 코드
- Task2.ipynb - Task 2를 완성한 코드 및 결과
- Task2.pdf - Task 2의 보고서
- Task3.pdf - Task 3의 보고서

이들을 모아 'ID.zip'으로 압축해주세요. ID는 학번으로 숫자 10개로 이루어져있습니다. 그리고 압축된 파일을 icampus[3]에 올려주세요.

진행함에 있어 어려움이 있으시다면 언제라도 icampus 문의게시판을 이용해주세요. 강의자나 조교가 여러분에게 도움을 드릴 것입니다.

[2]라벨이 없는 데이터 혹은 그 부분
[3]라벨이 있는 데이터 혹은 라벨 그 자체

References

[1] AI Hub. https://aihub.or.kr, 2022.

[2] Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jungwoo Ha, and Kyunghyun Cho. Klue: Korean language understanding evaluation, 2021.

[3] SKKU i-Campus. https://icampus.skku.edu/, 2021.