

Department of Software Convergence

Data Analysis Capstone Design Project   
Weekly Progress Report

|  |  |
| --- | --- |
| Project Title | 딥러닝을 활용한 노래 가사 감정 분석 |
| Student / ID | 김은비 / 2019100858  유창현 / 2018102121  이은경 / 2019100896  이인석 / 2018110659 |
| Reporting Week | 2022-05-06 ~ 2022-05-12 |
| Project Manager of Reporting Week | 김은비 |
| Faculty Supervisor | 유창현 |

# **Tasks Outlined in Previous Weekly Progress Report** (Provide detailed information on the tasks to be completed in this week)

* 음원 사이트 멜론에서 연도별 TOP100 가사 크롤링 진행
* 기존에 구축한 감성사전을 기반으로 새로 수집한 데이터에 대한 감정 라벨링 작업 수행
* 감정분류에 활용할 딥러닝 모델 자료조사

# **Progress Made in Reporting Week** (Provide detailed information on the progress that you made in the reporting week. Limit your write-up to no more than two page)

1. 멜론 연도별 TOP100 가사 데이터 크롤링

멜론의 차트 파인더 기능을 활용하여 연도별 TOP 100에 해당하는 곡의 가사 데이터를 수집을 진행하였다. 주간이나 월간 데이터의 경우 중복 데이터가 많을 것이라 판단하여 연도별 차트 크롤링을 통해 데이터를 수집하였다.

텍스트이(가) 표시된 사진

자동 생성된 설명

그 결과 30년치에 해당하는 3,000곡의 가사 데이터를 수집할 수 있었고, 중복 제거 등의 과정을 거쳐 총 2,803곡에 대한 감정분석을 수행하였다.

1. 기존에 구축한 감성사전 기반의 곡별 감정 라벨링 진행

2,803곡의 가사 데이터에 대한 토큰화 진행 후 기존에 구축했던 감성사전을 기반으로 감정 라벨링 작업을 수행하였다. 그 결과, 약 2,400곡 이상이 ‘행복’에 해당하여 이 방법으로도 데이터 불균형 문제를 해결할 수 없다는 결론이 도출되었다.

텍스트이(가) 표시된 사진

자동 생성된 설명

1. 감정분류에 활용할 딥러닝 모델 자료조사
   1. LSTM:  Long Short-Term Memory, 전통적인 RNN의 단점(장기 의존성 문제)을 보완하기 위해 게이트 개념을 추가한 방식이다. 은닉층의 메모리셀에 입력 게이트, 망각 게이트, 출력 게이트를 추가하여 불필요한 기억 지우고 기억해야 할 것들을 결정한다. 긴 시퀀스의 입력을 처리하는데 탁월한 성능을 보이며 셀 상태 정보를 가지고 있다. 처음 프로젝트를 계획 단계부터 사용하기로 결정한 모델이다.
   2. BiLSTM: Bidirectional LSTM, 두 개의 LSTM으로 이루어져 있는 모델이다. 하나의 LSTM은 input을 forward 방향으로 받고, 다른 하나는 backward 방향에서 받기 때문에 이전 정보와 이후 정보를 모두 담고 있다.
   3. GRU: Gated Recurrent Unit, LSTM보다 게이트의 개수가 1개 적어진 2개의 게이트(업데이트 게이트, 리셋 게이트)가 존재한다. 또한 LSTM과 다르게 셀 상태에 대한 정보가 없다. LSTM보다 간소화된 구조를 가지고 있지만 비슷한 학습성능을 보인다, 특히 학습 데이터가 적은 경우 LSTM보다 좋은 성능을 보이기도 한다.
   4. CNN: Convolutional Neural Network, 시각적 영상을 분석하는 데 주로 사용되는 모델로 자연어 처리의 경우 이미지 처리에 사용되는 2D CNN과 달리 1D CNN 사용한다. 커널의 너비를 문장 행렬에서의 임베딩 벡터의 차원과 동일하게 설정하여 설계한다.

# **Difficulties Encountered in Reporting Week** (Provide detailed information on the difficulties and issues that you encountered in the reporting week. Limit your write-up to no more than one page)

* 데이터 분균형의 문제: 노래가사 데이터의 특성상 기존에 수집한 ‘사랑’이 주제가 되는 노래나 인기도가 높은 TOP100의 노래가사 모두 행복에 해당하는 경우가 많아서 딥러닝 모델의 학습 데이터를 확보하는 것에 어려움을 겪고 있다. 그래서 다음주에는 각 감정에 해당하는 노래 가사 데이터를 수집하기 위해 해시태그 기능을 다시 활용하여 각 감정별 노래 수집을 진행할 예정이다.
* 멜론 아이피 차단: 멜론 인기차트 페이지에 오랜 시간 접속하여 관련 데이터를 수집하다 보니 사이트 자체에서 페이지 접근을 아예 차단하는 경우가 있어서 데이터 수집의 어려움이 있었다.

# **Tasks to Be Completed in Next Week** (Outline the tasks to be completed in the following week)

* 해시태그 기능을 활용한 데이터 재수집과 감성사전 수정, 라벨링 과정 완료(데이터 불균형 문제 해결)
* 논의를 통해 LSTM 외에 활용할 딥러닝 모델 확정 및 구현 시작