

Department of Software Convergence

Data Analysis Capstone Design Project   
Weekly Progress Report

| Project Title | 딥러닝을 활용한 노래 가사 분석 및 추천 시스템 |
| --- | --- |
| Student / ID | 김은비 / 2019100858  유창현 / 2018102121  이은경 / 2019100896  이인석 / 2018110659 |
| Reporting Week | 2022-03-25 ~ 2022-03-31 |
| Project Manager of Reporting Week | 이인석 |
| Faculty Supervisor | 유창현 |

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

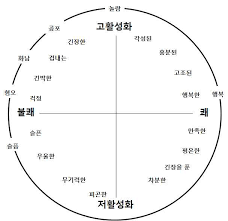
* 분석 대상 키워드를 통해 노래 가사 직접 크롤링
* 회의와 논문을 참고하여 세부 감정 선택

# **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. ’사랑’의 세부 감정 결정

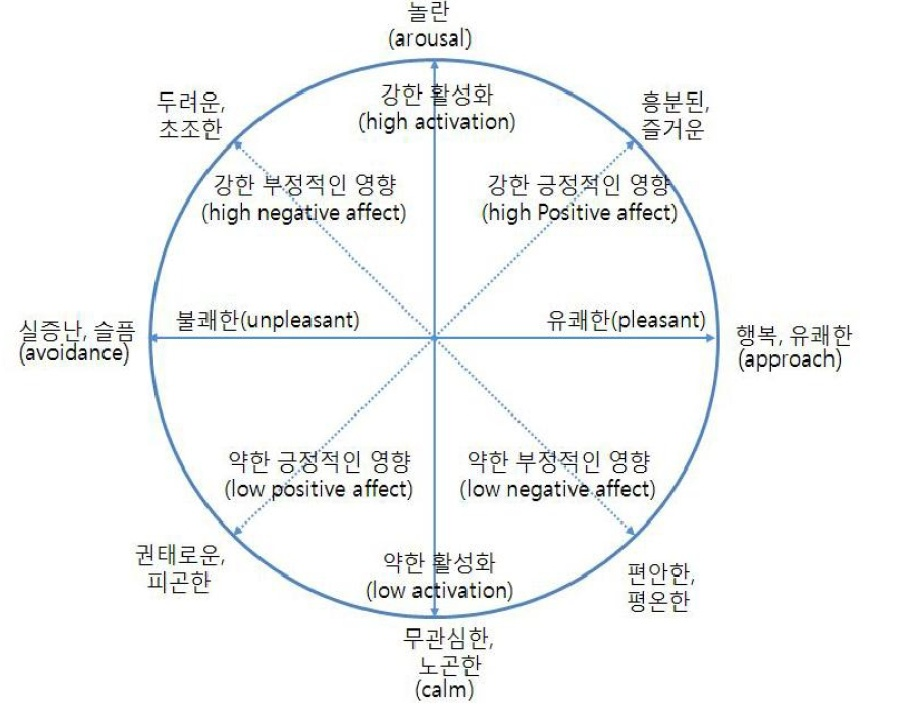
* Russell(1980) 모델(4가지)

Russell은 각성 수준과 정서가(긍정/부정)를 기준으로 정서를 네 가지 하위 차원으로 분류했다. 그 구체적인 분류는 고각성 긍정정서(신이 난, 흥분한, 놀라운, 아주 기뻐하는, 행복한 등), 저각성 긍정정서(기쁜, 만족스러운, 흡족한, 고요한, 차분한, 여유있는 등), 고각성 부정정서(불안해 하는, 두려워 하는, 긴장한, 화난, 괴로워 하는, 짜증이 난 등), 저각성 부정정서(비참한, 우울한, 슬픈, 따분한, 의기소침한 등)가 있다.



* Tellegen-Watson Clark 모델(8가지)

Tellegen-Watson Clark 모델은 긍정과 부정(positive/negative affect), 유쾌함과 불쾌함 (pleasantness/unpleasantness), 활성화 정도를 활용해 이전 감성 분류 모델들보다 더 다양한 감성을 분류하고 있다.



‘사랑’의 세부 감정을 결정하기 위해 기존 연구들에서 활용된 감정분류 모델을 찾아보았고, 위 두가지 모델이 가장 참고하기에 적합다고 생각했다. 어색한 번역투와 ‘사랑’이라는 키워드에 부합하지 않는 감정은 제외하고 팀원들간의 토의를 통해서 다음과 같은 사랑의 세부 감정 7가지를 직접 선정했다.

* 놀람 두려움 행복 슬픔 분노 무관심 편안함

1. 데이터 크롤링

자동화 도구인 Selenium 라이브러리를 활용해 가사 데이터 크롤링을 진행했다. 우선 멜론DJ 메뉴에서 ‘사랑’ 이라는 키워드를 해시태그로 검색했고 인기순으로 정렬되어 있는 4개의 페이지에서 노래 가사 데이터를 추출하기로 결정했다. 팀원 4명이 각 1페이지씩 담당하여 각 DJ의 플레이리스트에 접근해서 곡 별 제목과 가사를 추출했다. 이때 국내 가요가 아닌 팝이나 가사가 없는 곡들의 플레이리스트는 배제했다.

# **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)

BeautifulSoup 패키지를 활용해 가사 데이터를 크롤링 하려고 했으나 <a> 태그의 href 속성이 URL 절대 주소로 명시되어 있는것이 아닌 javascript로 구현되어 있어서 크롤링 하는데 어려움이 있었다. 따라서 다른 방법으로 크롤링을 시도하였다.

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

* 데이터 전처리

수집한 가사 데이터에서 영어, 특수문자 등 불용어를 제거하고 단어를 토큰화하고 품사를 파악 등의 전처리.