

Department of Software Convergence

Data Analysis Capstone Design Project   
Weekly Progress Report

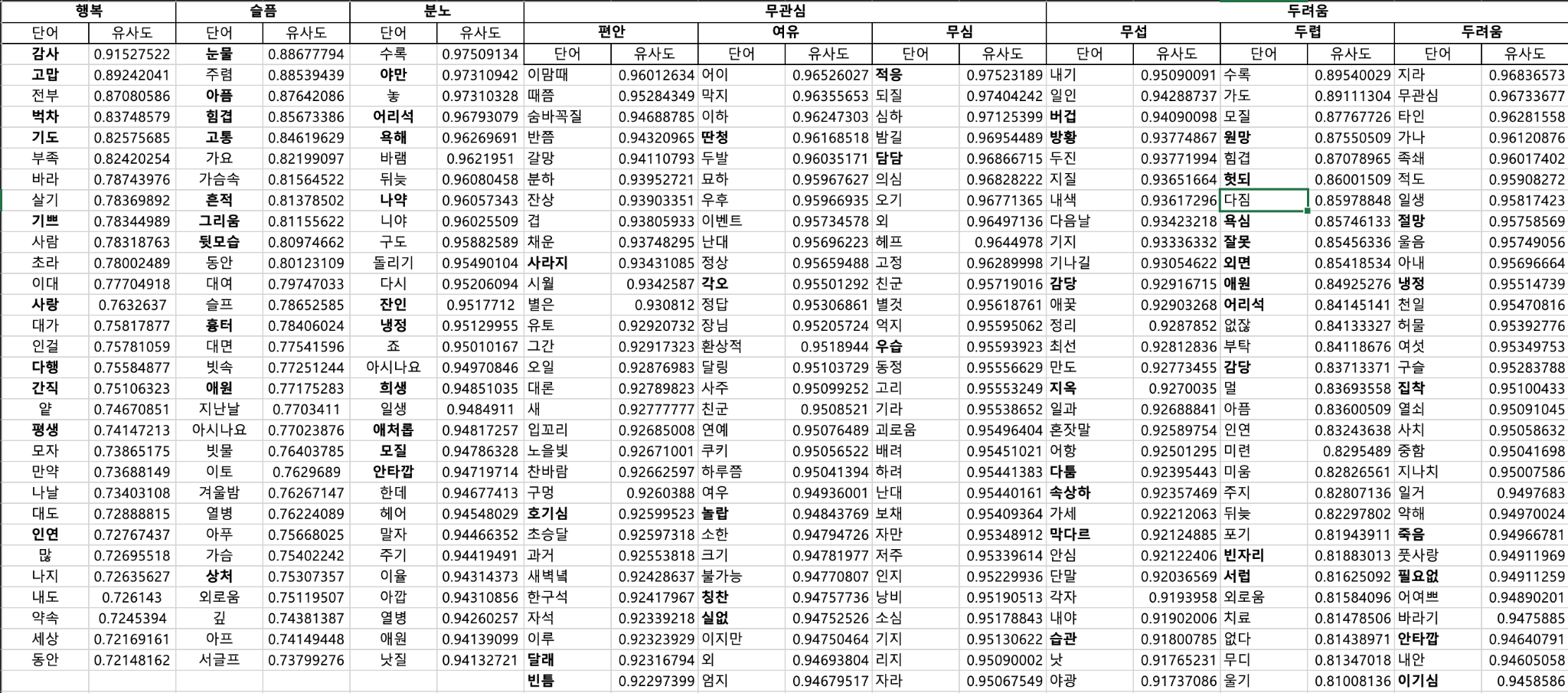
| Project Title | 딥러닝을 활용한 노래 가사 분석 및 추천 시스템 |
| --- | --- |
| Student / ID | 김은비 / 2019100858  유창현 / 2018102121  이은경 / 2019100896  이인석 / 2018110659 |
| Reporting Week | 2022-04-15 ~ 2022-04-21 |
| Project Manager of Reporting Week | 유창현 |
| Faculty Supervisor | 유창현 |

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

* 구축된 감성사전의 검토와 결과 수정

# **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)

* 지난주에 Word2vec을 활용하여 도출한 각 감정별 감성사전에 연관성이 없어보임에도 유사도가 높게 나온 단어(예: 무섭 - 긷)들이 있었고, 여러 품사들을 모두 활용하다보니 주요 키워드를 선정과 단어를 구별하는데도 어려움이 있었다. 이번주에는 이런 문제를 개선할 방법에 대해 팀원들과 논의를 진행하였다. 논의를 통해 나온 방법은 1) word2vec의 파라미터를 값을 더 다양하게 바꾸어 실험하면서 결과를 개선한다. 2) 기존에는 품사 제한 없이 한국어 형태소는 모두 사용하였는데, 이를 일반명사와 형용사로 제한한다. 3) 기존 감성사전 구축에 사용한 word2vec의 most\_similar함수는 코사인 유사도를 통해 유사도를 산출하므로 이를 개선하기 위해 다른 유사도 계산 방식을 활용한다. 4) 워드 임베딩 방식을 Word2vec에서 다른 방식으로 변경한다.
* 또한, 기존에 감정을 7가지(행복, 편안함, 두려움, 놀람, 분노, 슬픔, 무관심)로 분류하였는데, 감성사전을 구축하는 과정에서 놀람, 편안함 2가지 감정에 대한 키워드를 선정하는 것에 어려움이 있고, 도출된 결과에도 무관한 단어들이 다수 포함되었다. 이러한 이유로 토의를 진행한 끝에 ‘놀람’이라는 감정은 사랑이라는 주제에서 그리 중요한 감정이 아니며, 노래가사에 놀람과 관련된 단어가 많이 등장하지 않는다는 사실을 확인하였기 때문에 배제하기로 결정하였다. 편안함 역시 무관한 단어들이 결과에 미치는 영향이 크고 행복, 무관심 감정에 포함될 수 있는 감정 중 하나라고 판단하여 삭제하였다. 따라서, 최종적으로 확정된 5가지 감정(행복, 두려움,슬픔, 분노, 무관심)에 대한 감성사전 구축을 재진행하였다.
* 우선 첫번째로 채택한 방식은 사용 품사를 일반명사와 형용사로만 한정하는 방법이다. pos태깅을 통해 일반명사(NNG)와 형용사(VA)만을 추출해서 활용해 보았다.



하지만 수정된 방식 또한 애매하고 모호한 단어들로 채워지고, 의미의 실제 연관성이 매우 적은 경우가 대부분이며, 실제 활용되는 단어의 수가 매우 적었기 때문에 다른 방법을 찾아보기로 하였다.

* 두번째로 시도한 방법은 기존의 활용한 word2vec설계에서 그 파라미터를 바꾸어가며 결과를 개선하는 것이다. 기존의 word2vec설계는 CBOW 방식(sg = 0), size = 100, window = 5, iter = 10, min\_count = 3이었다. 여기서 iter와 min\_count를 주로 바꾸어가며 실험해 보았고, 유사도 도출에 사용하는 키워드도 변경해가며 결과를 확인했다. 사용된 키워드는 다음과 같다.

1. 행복(즐거움, 설렘 등): 행복
2. 두려움(무서움 등): 두렵
3. 슬픔(우울한, 비참한, 괴로운 등): 슬픔
4. 분노(미워하는, 짜증이 난 등): 화
5. 무관심(피곤한, 태연한): 무심



* min\_count = 1로 수정하여 새롭게 유사도를 도출



* 결과적으로 애매하거나 감정에 적절하지 않은 단어는 거의 없음, 몇 개 안되는 적절치 못한 단어들에 대한 처리 방식은 고민 중이다.

# **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)

* 품사를 제한하여 사용하기에는 활용 단어의 수가 급격하게 적어진다는 문제점이 있었고 문맥을 제대로 파악할 수 없을 것 같다는 우려도 있었다. 그래서 word2vec의 설계와 키워드로 바꾸어가며 결과를 도출해 보았지만, 저번 주에 교수님이 지적해주셨듯이 적지만 여전히 애매하거나 맞지 않은 단어가 존재하기 때문에 이를 해결할 방법이 필요하다고 생각한다. 따라서 이번주에 교수님께 맞지 않는 키워드를 인위적으로 삭제 괜찮을지 혹은 배제할 수 있는 다른 방법이 있는지 조언을 구할 계획이다.

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

* 감성사전 최종 확정 및 이를 활용한 곡별 감정 라벨링 진행
* 감정 분류에 활용할 딥러닝 모델 결정 및 설계 진행