

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 | 2022-03-18 ~ 2022-03-24 |
| 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)

-기존 논문 분석을 통한 사용 기법 논의

-분석 대상 키워드 결정 및 크롤링할 노래 선정 방식 토의

-노래 가사 크롤링 방향 논의

# **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)

기존 논문 분석을 통한 사용 기법 논의

-Word2vec: 연속적 워드 임베딩 학습 모형이다. 비슷한 의미를 가지는 단어는 같은 맥락에 출현한다는 것을 전제로 한다. TF-IDF알고리즘과 다르게 단어 간의 의미적 유사도를 벡터로 표현할 수 있다. 그러나 문서 내에서 단어의 순서로 인해 생기는 문맥을 고려할 수 없다는 단점이 있다.

해당 논문에서는 word2vec을 사용해 일반적인 TF-IDF방식보다 단어를 의미 있게 표현하고 doc2vec을 통해 문맥을 고려하여 문서를 벡터로 표현한다. 추출된 특징 벡터를 분류하기 위해서는 순환 신경망(RNN) 분류기를 사용했다. 모든 품사를 문서의 특징으로 사용하면 불필요한 정보가 다수 포함되기 때문에 명사를 문서의 주요 특징으로 보고 형태소 분석을 통해 명사만을 추출했다. 추출된 명사 중에서도 다수의 문서에서 출현하는 단어는 문서 간의 분별력을 낮추므로 제거(빈도수가 50%가 넘는 어절은 불용어로 처리)했다. Word2vec(일반적으로 성능이 더 우수하다고 알려진 스킵그램 방식)을 사용하여 단어 벡터를 생성하고 doc2vec을 사용하여 문서를 벡터화 한다. 마지막 단계로 문서 벡터를 RNN의 한 방법인 GRU 분류기를 통해 문서를 분류했다. 사용 데이터는 영어와 한국어로 된 신문기사이다. 성능은 TF-IDF < word2vec < word2vec + doc2vec 순으로 좋은 실험결과를 보였으며, 한글 기사 데이터에서도 정확도가 90%를 넘긴 것으로 보아 한글 문서에서도 단어의 의미와 문맥을 고려한 word2vec, doc2vec방법이 효과가 있다는 것을 확인할 수 있었다. ( 단어의 의미와 문맥을 고려한 순환신경망 기반의 문서 분류(2018))

-LSTM을 활용한 이별(or 사랑) 가사 감정 분류 모델 : 해당 논문에서는 이별 노래 가사에 나타난 감정을 분석하고 이별 가사에 나타나는 단어 간 유사도를 Word2Vec 학습을 통해 사전을 구축한 후 LSTM 을 사용하여 노래를 학습시켜 유사한 감정의 가사를 추천해주는 이별 가사 감정 분류 모델을 제안한다.

분석 대상 키워드 결정 및 크롤링할 노래 선정 방식 토의

-분석 대상 키워드 : ‘사랑’을 키워드로 하여 분석

-크롤링을 위한 노래 선정 방식 : 멜론 사이트의 멜론 DJ 플레이리스트 중 ‘#사랑’ 태그와 관련된 플레이리스트들에 있는 노래를 선정

노래 가사 분석과 유사도 측정 방식의 세부사항 결정

‘사랑’ 키워드에 대한 세부 감정들을 결정하고, 가사를 학습한 Word2Vec 모델을 통해 세부 감정들과 유사도가 높은 단어들을 뽑습니다.

뽑은 단어들을 데이터 라벨링하여 딥러닝 모델(예 : LSTM)을 활용하여 분류 모델을 구현하고, 분류된 그룹 내의 유사도를 측정합니다.

추천시스템 : 컨텐츠 기반 필터링(Contents-Based Filtering)

음악 간에 측정한 유사도를 바탕으로 사용자가 듣거나 선택한 음악과 유사도가 높은 음악을 추천

예) 유튜브의 본문란, 자막을 통해 추천, 쇼핑몰에서 비슷한 이미지의 옷 추천

사용하는 기법 : Word2Vec, doc2vec, Term-Based 기법(문서의 유사성), Similarity 알고리즘, 문서 다중 분류 등

# **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)

1. 라벨링은 어떻게?

특정 키워드와 관련된 노래 가사 데이터를 수집한 후에 그 가사의 내용과 의미에 따라 세부적인 감정을 분류하려면 그 기준이 존재해야 한다. 따라서 선행연구에서 진행했던 것처럼 세부 감정을 결정한 후 그 감정에 맞는 감성사전을 구축하여 각 노래가사마다 라벨링을 진행하기로 결정했다.

1. 노래 가사 크롤링은 어떤 방향으로 진행해야 하나?

기존에 사람들이 많이 진행했던 방식은 멜론의 top100과 같은 랭킹 내 노래 정보를 크롤링하는 것인데 우리는 한가지 키워드에 맞는 노래 가사를 수집해야 하기 때문에 그 방식을 결정하는데 어려움이 있었다. 논의 끝에 멜론에서 제공하는 ‘멜론DJ’메뉴의 해시태그 기능을 이용하여 노래 가사를 크롤링하기로 결정했다.

단어의 의미와 문맥을 고려한 순환신경망 기반의 문서 분류(word2vec+ doc2vec 활용)

VS LSTM을 활용한 이별(or 사랑) 가사 감정 분류 모델

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

-노래 가사 크롤링

여러 패키지들과 코드를 활용하여 음원 스트리밍 서비스 사이트(멜론)의 #사랑 키워드를 포함하는 음악 플레이리스트들의 음악을 크롤링하여 노래와 가사 데이터를 얻음

-데이터 전처리와 세부 감정 결정

수집한 가사 데이터에서 영어, 특수문자 등 불용어를 제거하고 단어를 토큰화하고 품사를 파악 등의 전처리 과정을 시작하고 ‘사랑’ 키워드에서 느낄 수 있는 세부 감정들을 관련 자료 수집과 회의를 통해 결정해야 함.