

# 자연어 처리 기술 기반대본 감정 분석

- First Goal

: word2vec, 다중 감성 분류 모델, 감성 사전 등의 자연어 처리/분석 기술을 통한 대본 감정 분석

- Issue

1) 균일하지 않은 대본 형태 & 어미 변화가 다양하며, 불용어 분리가 힘든 한국어 특성

=> '주요 인물의 대사' 데이터셋을 추출하기 위해 총 167화 대본 전처리 수작업으로 진행

=> konlpy(형태소 분석 라이브러리), '한국어 불용어 사전' 이용해도 걸러지지 않는 불용어 수작업으로 제거

2) 한국어 대본 감성 분석에 적합한 머신러닝 모델 찾을 수 없음

=> 현존하는 한국어 감성분석 모델 대부분은 '리뷰' 데이터에 포함된 '평점' 데이터를 이용해 문장의 긍/부정을 라벨링하거나, 문장에 붙은 이모티콘의 감정을 문장의 긍/부정과 연관시켜 모델 학습

(ex. 리뷰 평점이 3이면 긍정의 강도를 3으로 학습 / 이모티콘이 '분노'일 경우 부정 강도를 5로 학습)

=> 대본은 캐릭터별 자주 사용하는 어휘가 도드라져, 쇼핑몰/영화 리뷰와는 다른 문장 특성을 지니므로 기존 모델 활용 불가

3) 기존 '한국어 감성 사전'을 통한 형태소 감성 매칭/분석 불가

=> 기존 범용 감성 사전 중 '오픈한글'은 서비스 종료, 'KNU'는 수록 어휘가 '거침없이 하이킥' 데이터셋과 상이

=> '대본 형태소 분석 - 감성 사전과 매칭' 방법으로는 한 문장 안에서 긍/부정 어휘가 함께 등장하는 케이스 분석 불가

## • Solution A

**: 거침없이 하이킥' 대본만을 위한 감정 사전 생성 => '기쁨', '슬픔', '분노', '평온', '솔직함' 5가지 감정으로 분석하는 캐릭터의 마음**

1. 감정 = 말(대사) + 행동(지시문)
2. 감정 정의에 따른 사전 생성
  - 1) 말(대사) 형태소 분석 > 빈도 수 기준으로 단어 리스팅 > 4가지 감정으로 분류
  - 2) 행동(지시문) 형태소 분석 > 빈도 수 기준으로 단어 리스팅 > 4가지 감정으로 분류=> 말, 행동 2가지 key를 가진 Emotion\_dic 생성  
ex) Emotion\_dict\_joy  
= { line : ["행복", "사랑", "기쁨", "여보", "밥", "집"...],  
action : ["쓰다듬", "웃으며", "좋아하며"...]}
3. 캐릭터 감정 알아내기
  - 1) 등장인물별로 Character\_emotion\_count 선언
  - 2) 캐릭터의 말(대사) 형태소 분석 > 감정 사전 line 키의 값과 비교 > 일치 시 해당 감정 count += 1
  - 3) 캐릭터의 행동(지시문) 형태소 분석 > 감정 사전 action 키의 값과 비교 > 일치 시 해당 감정 count += 1: 이때, if line == action이면 '솔직함' 감정 count += 1  
elif line != action이면 해당 감정 count -= 1(위에서 +1 해준 값 취소)
4. Emotion\_dict의 감정별 count 값으로 캐릭터의 5가지 감정 보여주기

## • Solution B

**: 거침없이 하이킥' 대본만을 위한 감정 사전 생성 => '기쁨', '슬픔', '분노' 3가지 감정으로 캐릭터 대사 분석**

1. 주요 캐릭터의 전체 대사 추출 > 형태소 분석 > 불용어 제거 > 빈도 수 기준으로 리스팅 > '기쁨', '슬픔', '분노'의 감정사전 제작
2. 캐릭터별 대사 추출 > 형태소 분석 > 불용어 제거 > 토큰화된 대사 데이터셋 확보 > 감정 사전의 단어와 일치할 경우 count += 1 > 감정 수치 도출