INTRO

*** Mission : 주어진 잉여 데이터(영화 리뷰 데이터)를 가지고 있으나 활용 방안과 비즈니스 연결이 명확하지 않다. 주어진 잉여 데이터를 활용할 수 있도록 분석 프로젝트를 기획하고 비즈니스 모델을 설계한다.**

- 이전 movie recommendation 프로젝트에 관련 내용이 포함되어 있으므로, mash-up은 고려하지 않고 주어진 데이터만을 이용한다.

*** 분석 process :**

1) 데이터 가공 및 전처리 :

- 불필요한 column(user_pk, user_id, review_count, title) 제거
- 점수에 따라 0~4 : B(Bad), 5~7 : A(Average), 8~10 : G(Good)으로 라벨링

2) 모델 설계

- B/A/G에서 가장 많이 사용된 단어 100개를 추출하여 리스트들을 만들고, TF가 높은 단어(공통적으로 자주 관측되는)를 제거한다.
- B리뷰, A리뷰, G리뷰에서 각 리스트들의 단어들이 출현하는 확률 테이블을 만든다.
- 확률 테이블을 이용해 B,A,G로 라벨링될 확률을 구하고, MAP에 따라 라벨링한다.

3) 결과 분석 및 결론 도출

*** 모델 Description :**

- Input : Review data {pk, user_pk, user_id, review_count, title, score, content}
- Output : Result {g_prob, a_prob, b_prob, predicted, actual}
- 스코어 없이 텍스트 데이터만을 이용해 해당 리뷰를 “Good”, “Average”, “Bad”로 분류할 수 있는 이진 나이브 베이지안 분류기 설계 및 훈련
- 환경/인자를 자유롭게 바꿔보고 그에 따른 정확도 변화도 함께 파악하기 위해, 라이브러리를 사용하지 않고 직접 설계

pk	user_pk	id	review_count	title	score	content
1	1706	thfk****	43	프리즌	5	재밌는데 지루함 멋있는데 지루함 정말 딱뭣도 아닌..난이걸얼마나 기대했던가..
2	1706	thfk****	43	23 아이덴티티	1	아진짜일바쫌쓰지마세요——방금보고나왔는ㄷ ㅔ내돈주고본영화2위임 쓰레기 1위는 미스터메레그린
3	1706	thfk****	43	트리플 엑스 리턴즈	10	후기:처음부터 끝까지 액션 진짜 액션은 진리다 끝.그래도 나쁘지않았어요ㅋㅋㅋㅋ스토리는10% 액션은90%

〈 Input - Review data 〉

BODY

* Result :

-

movie_sa.R × movie_sa2.R ×

Source on Save | Run | Source

```r
#install.packages("KoNLP")
#install.packages("rJava")
library(KoNLP)
#useSejongDic()
useNIADic()

#---------------------#file input#---------------------#
#setwd("/Users/hodong/Desktop/review_senti_analysis/r_code/")
mp <- data.frame(read.csv("../raw_data/movie_prepared.csv",header=TRUE))
dp <- data.frame(read.csv("../raw_data/data_prepared.csv",header=TRUE))
dtr <- data.frame(read.csv("../raw_data/data_train.csv",header=TRUE))
dte <- data.frame(read.csv("../raw_data/data_test.csv",header=TRUE))
#---------------------#file input#---------------------#


#---------------------#functions#---------------------#
#refining reviews to text&sentiment
sen_divide <- function(x){
  if(x<=4){ return("B") }
  else {
```

112:1 (Top Level) R Script

〈 R code 〉

movie_sa.R × movie_sa2.R ×

Source on Save | Run | Source

```r
#install.packages("KoNLP")
#install.packages("rJava")
library(KoNLP)
#useSejongDic()
useNIADic()

#---------------------#file input#---------------------#
#setwd("/Users/hodong/Desktop/review_senti_analysis/r_code/")
mp <- data.frame(read.csv("../raw_data/movie_prepared.csv",header=TRUE))
dp <- data.frame(read.csv("../raw_data/data_prepared.csv",header=TRUE))
dtr <- data.frame(read.csv("../raw_data/data_train.csv",header=TRUE))
dte <- data.frame(read.csv("../raw_data/data_test.csv",header=TRUE))
#---------------------#file input#---------------------#


#---------------------#functions#---------------------#
#refining reviews to text&sentiment
sen_divide <- function(x){
  if(x<=4){ return("B") }
  else {
```

112:1 (Top Level) R Script

〈 Output - prediction result 〉

BODY

* limitation :

1) CPU 리소스의 부족.

- 실제 크롤링한 데이터는 100만개였지만 분석에 활용할 수 있는 것은 10만개 미안이었다. 데이터 양을 늘릴 경우, 정확도가 올라갈 것으로 예측될 뿐만 아니라, 분류기를 장르별/국가별/등급별로 세분화할 수 있다.
- 어절 단위 분석이 제한되어 단어 자체만으로 분석해야 해서, 의미가 희석되었고, 이 때문에 분석 방법의 선택폭이 줄어들었다.

  eg. “좋은 스토리, 연출은 별로”라는 리뷰가 있을 때, 단어 단위로 추출할 경우, {별로, 스토리, 연출, 좋다}라는 vector를 얻게 되는데, 스토리가 별로인지 연출이 별로인지에 대한 의미가 희석된다.

2) 한국어 사전의 한계

- SejongDic보다 NAIDic이 다루는 단어가 많아 정확도는 기본적으로 높아졌지만, 여전히 포괄하지 못하는 신조어와 철자 오류 등을 포괄할 수 없었다.

* Expected Improvement :

1) 인구통계학적 정보 mash-up :

- 인구통계학적 정보는 분류기의 종류를 다양화하는데 사용될 수 있다. 영화에 대한 평가는 영화 자체의 특성(장르/국가 등) 뿐만 아니라, 평가하는 평가자의 특성(성별/나이 등)에 따라서도 영향을 많이 받기 때문이다.

2) Python의 pre-trained된 vector map을 이용 :

- word2vec의 pretrained embedded word vector를 이용해 추출된 단어들과 유사도가 높은 단어도 분류에 함께 사용하여 정확도를 높일 수 있다.

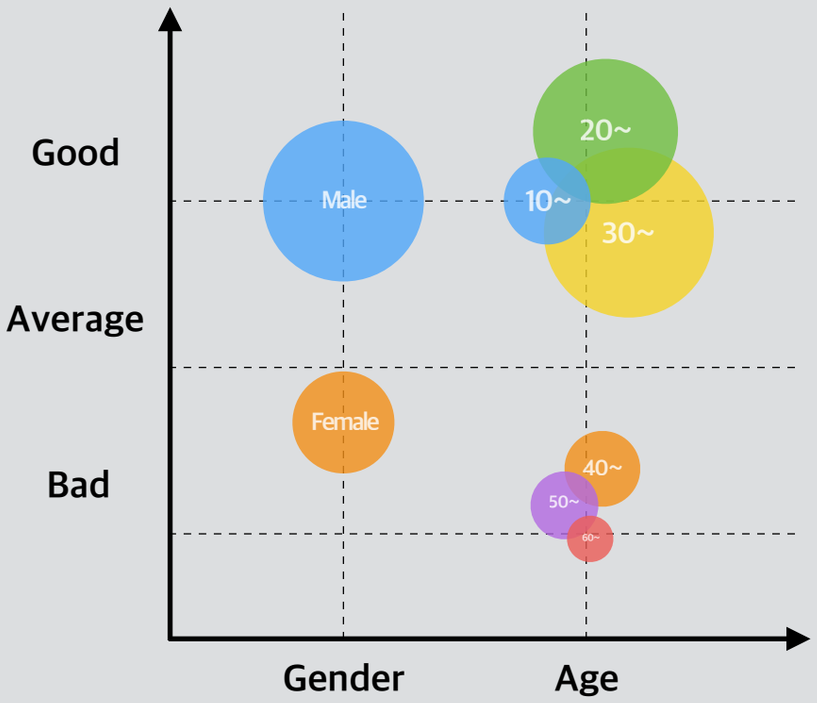
Conclusion

* Business application :

1) 집단 경향성 분석을 마케팅 전략에 활용

네이버는 영화 페이지 뿐 아니라 동일 플랫폼 내에서 텍스트 데이터(리뷰)를 구할 수 있는 곳이 많고, 서비스를 제공하는 네이버는 리뷰를 작성한 사용자의 인구통계학적 정보도 가지고 있기 때문에 집단 경향성 분석도 가능하다. 집단 경향성 분석은 세그먼테이션에 인사이트를 제공하고 마케팅 효율을 높일 수 있다.

< Preference to Action genre >



이와 유사하게 컨텐츠 자체의 속성과 사용자의 속성에 따라 세그먼테이션의 선호도 등에 대한 경향성 분석을 진행할 수 있다.

액션 장르 컨텐츠의 마케팅 타겟 설정에 활용

- Main target : 20~30 Male
- Sub target : 10 Male

Conclusion

* Business application :

2) 학습된 분류기를 이용해 교차 플랫폼의 데이터 추출 및 마이닝에 활용

필요에 따라서 소위 '리뷰 알바'를 활용하기도 하지만, 네이버 뿐만 아니라 영화를 제작/배급사 입장에서도 영화에 대한 진짜 경향성에 대한 정보는 필요하다. 그런 입장에서 거대 플랫폼의 영화 리뷰는 활용에 제한되는 측면이 있다. 반면, 영화 페이지의 리뷰로 학습된 분류기로 뉴스/블로그/트위터의 영화 평을 라벨링하거나, 뉴스/블로그/트위터 댓글의 경우 '추천/반대'로 학습 데이터를 추가해 라벨링을 할 경우, 단순히 영화 페이지의 리뷰만으로 경향성을 분석하는 것보다 상대적으로 노이즈가 적은 데이터를 구할 수 있다.

*** 분류기를 이용해 트위터 & 인스타그램에 올라온 영화의 리뷰를 라벨링해 활용할 수 있다.**