IMDB의 영화 리뷰 텍스트 분석 및 리뷰의 유용성 지표 개발

2015580012 통계학과 박수영
2015120039 행정학과 변주연
2015280079 경제학부 정지원
2015920061 컴퓨터과학부 현아연

목차

Ⅰ. 서론

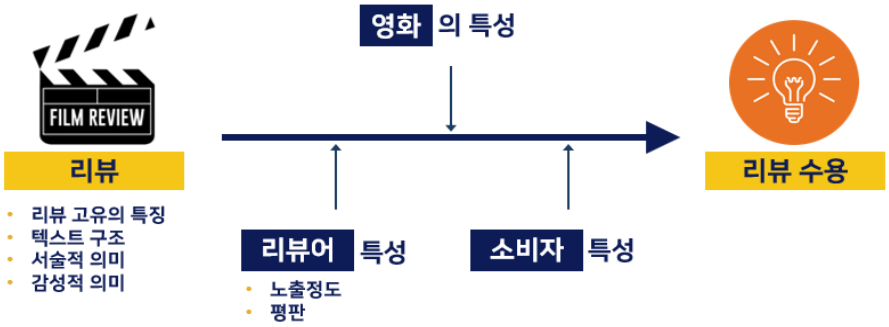
IMDb는 Internet Movie Database의 줄임말로, 영화뿐만 아니라 드라마, 배우, TV쇼 등을 망라하는 정보를 제공하는 데이터베이스이다. 2018년 12월 기준 380만여개의 리뷰가 게시되어 있는 세계에서 가장 큰 규모의 리뷰 사이트이기도 하다. IMDb의 통계자료에 따르면, 매달 250만명 이상의 사용자가 영화에 관한 정보를 얻기 위해 IMDb를 방문한다.

<그림 1>

IMDb에서는 <그림 1>에서 볼 수 있듯이, 소비자들이 일부 기준을 바탕으로 리뷰를 순차적으로 정렬하여 선별적으로 읽는 것이 가능하다. 그러한 기준 중 하나가 “Helpfulness”였는데, 이는 리뷰를 읽고 이 리뷰가 유용했는지 평가하는 시스템에서 나타난 것이다. 따라서 IMDb 웹사이트 내에서는 한 영화에 대하여 이용자들로부터 가장 많은 ‘helpful’하다는 표를 받은 순으로 리뷰를 정렬해 볼 수 있다.

리뷰를 읽는 사람들을 소비자라고 칭한다면, 소비자들은 효율적으로 리뷰를 읽고자 한다. 즉, 소비자는 수 많은 리뷰 중에서 자신에게 도움이 될 만한 리뷰를 읽고 싶어한다. 그러므로 리뷰를 직접 읽기 전 리뷰의 유용성을 알 수 있다면, 소비자는 시간이라는 비용을 아낄 수 있게 되는 것이다. 소비자의 측면뿐만 아니라, IMDb는 유용성을 바탕으로 리뷰를 선별할 수 있게 된다면, 질적 우수함을 갖춘 영화 리뷰 데이터를 구축할 수 있게 되어, 더 많은 사용자를 유인할 수 있을 것이다.

소비자들은 리뷰를 읽음으로써 자신의 영화 선택의 근거를 얻을 수 있고, 영화 관람 이전에 영화에 대한 정보를 얻게됨으로써 자신의 선택이 실패할 확률이 줄일 수 있다. 영화 리뷰가 단순히 영화의 내용만을 포함하는데에서 그치지 않고, 특정 영화에 대한 전반적인 관람자의 평을 담는 글이다. 또, 리뷰를 쓰는 사람, 즉 리뷰어의 특성도 드러날 것이다. 예를 들어, 리뷰어의 작성자가 영향력있는 평론가라면 리뷰어의 특성이 리뷰를 읽는 사람의 판단에도 영향을 미칠 것이다. 이처럼 리뷰에는 문자적인 특성뿐만 아니라 리뷰 작성자의 특성 및 리뷰의 대상인 영화의 특성이 복합적으로 담기게 되는데, 소비자들은 이러한 점을 인식하고 리뷰를 수용한다.



유용성의 지표가 마련된다면, 리뷰를 읽는 소비자들은 유용한 리뷰를 선별할 수 있게 되어 효율적인 영화 선택의 기준을 마련할 수 있을 것이다. 영화의 활발한 리뷰 작성을 위해서, 리뷰어의 입장에서는 유용하다고 평가되는 리뷰의 가이드라인이 제시될 수 있으며 IMDb 측에서는 이를 이용하여 향후 프리미엄 리뷰어, 믿을만한 리뷰어와 같이 리뷰어의 등급을 매길 수 있다는 데에서 리뷰의 유용성에 대한 그 의의를 둘 수 있다. 따라서 본 연구는 이 기준에서 착안하여 영화 리뷰를 이용하여 리뷰의 유용성에 미치는 요인들을 탐색하고, 이를 바탕으로 유용성을 지표화 하고자 한다.

Ⅱ. 본론

1. 분석 방법 소개

본 연구에서는 주어진 리뷰 데이터에 대하여 충분히 탐색하여 리뷰 및 리뷰의 텍스트가 가지고 있는 특성 및 패턴을 파악하고자 한다. 따라서 텍스트의 단어들을 통해 단순 빈도 분석, 토픽 분석, 감성 분석을 실시하였다. 단순 빈도 분석에서는 단어구름을 생성하였고, treemap을 통해 각 단어별 긍부정 빈도를 확인하였다. 여기서 영화의 구성 요소를 확인하기 위해 토픽 분석을 실시하였고 이를 기반으로 영화 구성요소에 대하여 알 수 있었다. 따라서 이들의 사전화 작업을 수행하였다. 또한 기계학습 알고리즘을 이용하여 긍부정을 분류해냈고 이들을 통해 파악할 수 있는 리뷰 데이터의 특징을 파악하였다. 결론적으로 리뷰를 그 리뷰가 담고 있는 기술적 측면, 표현적 측면 그리고 감성적 측면으로 총 3가지의 리뷰의 특성을 파악할 수 있었다.

위에서 파악한 리뷰 텍스트의 특징을 이용하여 리뷰의 유용성에 영향을 끼치는 요인들을 알아내기 위해 회귀 모형을 적합시켰다. 여기서 유용성은 각 변수들의 선형결합으로 파악이 되며 회귀 분석을 통해 각 변수들의 유의성 여부와 유의한 경우 그에 대한 영향력을 알아보고자 하였다.

2. 데이터 및 변수 설명

본 연구에서는 UCI Machine Learning Repository의 imdb의 영화 리뷰 데이터를 사용하였다. 해당 데이터는 1000개의 짧은 영화리뷰들이 담겨있는 변수(text)와 리뷰에 대한 긍부정 감정을 나타내는 변수(point)를 가진다. point는 리뷰가 긍정일 때 1, 부정일 때 0을 값으로 가진다. 총 1000개의 리뷰 중 긍정의 의미를 가진 리뷰의 수는 500개이며, 부정 리뷰의 수 또한 500개이다. 따라서 이 데이터는 동비율을 가진 데이터에 대하여 긍부정의 감성을 분류해내는 것이 주 목적인 데이터이다.

3. EDA

3.1 데이터 탐색 및 전처리

첫째, 원 데이터를 R을 이용하여 불러오는 과정에서 일부 영문자가 한글로 읽히는 인코딩 오류가 발생하여 해당 문자에 대한 전처리 작업을 실시하였다. 원본 데이터와 비교 결과, 빈칸에 대한 인코딩 오류와 'Québec'의 'é', 'å' 등의 라틴어 계열 강세(tilde) 표시에 따른 오류가 있었다. 따라서 이들을 적절히 삭제 혹은 변환하는 과정을 거쳤다.

전처리 전	전처리 후	원 텍스트	오류 텍스트
혯,혰,혚		… in all of them a true…	… in all of them 혯 a true…
챕	e	"movie business" of Québec	"movie business" of Qu챕bec
책	a	with Trond Fausa Aurvåg	with Trond Fausa Aurv책g

리뷰의 내용 파악을 위한 토픽 분석을 위해 어간을 분리하게 되면 원데이터의 의도와는 다른 의미로 변환될 수 있으므로 어간분리작업은 진행하지 않고, 아래 그림과 같이 올바른 변환 작업을 위한 원래의 의도에 영향을 주지 않는 파생변수사전을 만들었다.

전처리 전	전처리 후	전처리 전	전처리 후
scenes	scene	characters	character
sucked	suck	minutes	minute
sucks	suck	acting	act
things	thing	acted	act
/10	point	actors	actor

out of 10	point	liked	like
points	point	likes	like
10.10	tenpoint	films	film
makes	make	movies	movie
effects	effect		

out of 10, /10 등 점수를 표현하는 방법이 다양했기에 모두 point로 통일하였다. 또한 10.10처럼 특수문자 및 숫자제거 시 사라지는 단어를 위해 위와 같이 대체하였다. 뒤에 등장할 영화 구성 요소별 빈도 파악의 정확성을 위하여 effects, acted 등의 단어를 단어원형으로 대체하였다.

can't	cannot	couldn't	mustn't	isn't
aren't	wasn't	weren't	hasn't	haven't
hadn't	doesn't	don't	didn't	won't
wouldn't	shan't	shouldn't	never	neither
not	nor	nobody	nothing	

토픽분석 및 단어 빈도 분석에서는 내용을 파악하는 것이 주 목적이므로 의미가 'no'와 같은 위 단어들을 모두 'no'로 대체하였다.

one	movie	film	will	just	can	ever

또한 위 단어들은 내용 파악에 영향을 주지 않다고 판단하여 제거하였다.

3.2 단어들의 단순 빈도 분석

(1) wordcloud

전체 리뷰 데이터에서 빈번하게 나타난 단어들이 무엇인가 파악하기 위해서, 사용된 단어들에 대한 단순 빈도 분석을 실시하였다.

위의 워드클라우드에서는 중심에서 가까울수록 단어의 이용 빈도가 높은 것을 의미한다. 'bad'라는 두 단어의 사용빈도가 매우 높다는 것을 파악할 수 있었다. 'just', 'like', 'act', 'character', 'time'과 같은 단어의 사용이 빈번한 것을 확인해 볼 수 있었다.

term	bad	like	good	character	act	time	really	great	even	see	scene	well
freq	67	59	56	55	46	42	41	40	39	36	33	31

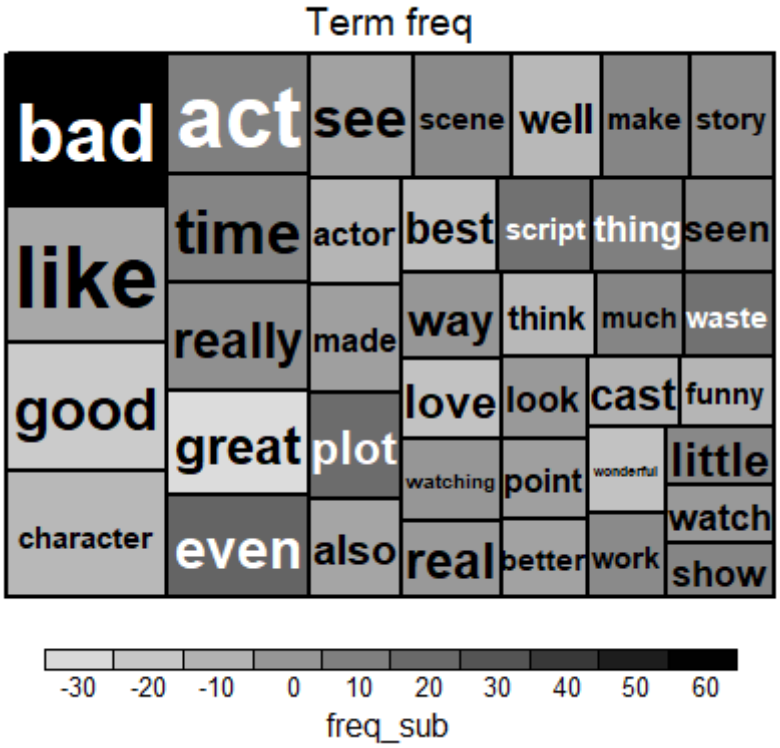
최소 30번 이상 사용된 단어들을 내림차순으로 정렬한 결과는 위와 같다. 워드클라우드에 나타났던 단어들의 언급 빈도가 구체적인 수치로 표현되어 단어의 빈도를 더욱 정확하게 파악할 수 있었다. 가장 빈번하게 찾아진 단어는 최대 70번이 넘게 사용된 것을 확인했다.

(2) ngram

```
> sort(bigramlist, decreasing=T)[1:10]
     waste time       ever seen        one best        ive ever           s ever
             11              10               7               6               6
       act bad     avoid costs      dont think     dont waste dont waste time
              4               4               4               4               4
```

의미 있을 것이라 판단되는 n-gram을 확인해보았다. n개의 어절과 음절을 연쇄적으로 분류하여 그 빈도를 확인해 보았다. 가장 많이 발견되는 것은 'waste time'으로 영화를 보는 것이 시간낭비일만큼 영화에 대하여 부정적인 의견을 나타낸 것을 확인하였다. 뒤를 이어서는 'ever seen'과 'one best'가 각각 10번과 7번으로 높은 빈도를 보였다. 특히 'ever'은 ive나 s와 같이 여러 번 n-gram으로 등장하는데 소비자가 지금까지 본 영화 중에서 좋았다 혹은 나빴다는 의미로 영화에 대하여 평가할때 강조하는 표현으로 많이 사용된 어구 임을 확인해 볼 수 있었다. 뒤를 이어서는 'act bad'가 등장했는데, 이는 배우들의 연기가 별로였다는 식의 표현을 act bad라는 표현으로 자주 쓴 것을 찾아볼 수 있었다. 'waste'는 'dont waste time', 'dont waste'와 같이 다른 단어와 결합되어 많이 쓰이는 것을 다시 한번 확인해 볼 수 있었는데, 이를 통해서 소비자는 영화를 평가함에 있어서 영화에 시간을 소비하는 것으로 보기 때문에 그 시간이 낭비였다라고 연관지어 생각하는 것을 확인해 볼 수 있었다.

(3) treemap



위 그림은 최소 15번 이상 사용된 단어들에 대해 treemap을 구성해 본 결과이다. 각 박스의 크기는 전체 데이터에서의 단어 출현 빈도의 정도를 나타내며, 각 박스의 색이 검을수록 부정리뷰에서 더 많이 나타남을 나타낸다. 바탕에 씌어진 글씨의 경우 검은색이면 부정리뷰보다 긍정리뷰에서 더욱 많이 나타났으며, 하얀 글씨면 긍정리뷰보다 부정리뷰에서 더욱 많이 나타났음을 의미한다.

예를 들어 가장 큰 부분을 차지하는 'bad'의 경우 부정적 리뷰로 분류된 경우가 우세하였기에 검은 배경에 하얀글씨가 써진 것을 확인해 볼 수 있다, 'act'의 경우 'bad'와 같이 흰색으로 글씨가 씌였기에 부정적 리뷰에서 많이 확인이 되지만, 배경이 'bad'보다는 연한 색에 가깝기 때문에 'bad'보다는 긍정적리뷰에서도 많이 관찰된 단어라는 것을 알 수 있다. 'good'과 'like'의 경우 검은색 글씨로 씌어져 긍정적 리뷰에서 많이 찾아지는 것을 공통적으로 확인해 볼 수 있었으나 'good'의 배경이 더 밝은색을 가지고 있기 때문에 'like'는 'good'보다 부정적인 리뷰에 높은 출현빈도를 보이는 것이라 짐작해 볼 수 있다.

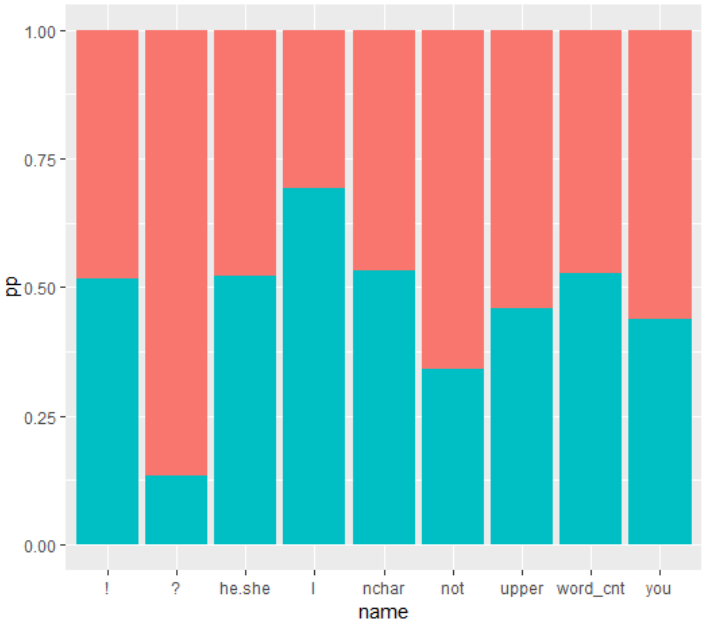
그외에도 'time'을 주목해 볼 수 있었는데 'time'은 다소 부정적인 리뷰에서 자주 찾아지기에 배경이 어두운 색을띄지만, 글씨는 밝은 하얀색이 아니라 검은색을 보이기 있기 때문에 긍정적인 리뷰에서도 많이 관찰이 되는 단어라는 것으로 이해해 볼 수 있었다. 'really'역시 'time'과 비슷한 모습을 보였다. 이를 통해 볼 때, 'really'를 이용한 강조의 표현은 긍정과 부정에서 골고루 찾아지는 단어라는 것을 이해해 볼 수 있다.

3.3 파생 변수

리뷰의 긍정 및 부정 여부를 판단하는데 있어서 스테밍을 거치게 되면 본래의 의미를 파악하기 힘들게 될 가능성이 있다. 예를 들어, "It isn't good"이란 문장에서 중요하지 않은 단어들을 제거한다면, 'good'만이 남게 되어, 긍정으로 분류될 가능성이 크다. 따라서 우리는 이러한 점을 고려하여 전처리를 실시하였다. 아래와 같이 not과 같은 부정어구의 개수에 대한 변수들을 생성하였다. 또한, 일례로 'It's BAD'의 문장의 경우에는 의미를 강조하고 싶은 단어에는 대문자를 사용한다는 것을 알 수 있다. 따라서 연속된 대문자가 몇개 사용되었는지 등을 알아내기 위해 강조용 대문자 수에 대한 변수를 생성하였다. 리뷰 데이터 중 느낌표가 있는 경우에는 대부분 강한 찬사나 강한 비난과 같은 극단적인 내용을 담고 있었다. 물음표가 담겨있는 리뷰는 부정적인 의견을 담은 경우가 많았는데, 원 데이터를 살펴보니 비꼬는 뉘앙스의 리뷰가 대다수였다. 이와 같이 리뷰 데이터의 감정적 특성을 반영하는 파생변수를 생성하였다. 또한, 리뷰데이터의 외형적 특성을 보여주는 각 리뷰 당 사용된 단어 개수 및 전체 텍스트 길이를 변수로 만들었다. 또 다른 리뷰의 형식적 측면을 반영하는 파생변수는 문장의 주어를 대상으로 한다. 문장의 주어를 'I', 'you', 'she/he'로 나누어 이러한 대명사에 해당하는 주어의 개수를 의미하는 파생변수를 추가하였다. 내용적 측면에서는 토픽 분석을 통하여 알아낸 영화의 주요 요소 3가지에 대한 내용이 들어가있는지, 말하고자 하는 대상이 무엇인지를 알아보기 위한 변수들을 생성하였다. 위 결과를 거쳐 추가한 파생변수는 아래 표와 같다.

1	느낌표	7	He/She
2	물음표	8	부정어구(not,nor 등)
3	연속 대문자(강조)	9	단어 개수
4	문자열의 개수	10	감성어 단어수
5	I	11	영화 요소 단어수
6	You	12	

아래의 그래프는 위의 표의 일부인 느낌표, 물음표, 연속 대문자(강조), 문자열의 개수, I, you, he/she, 부정어구를 대상으로 긍정과 부정의 표현에 해당하는 비율을 시각화한 것이다. 이 파생변수들의 기초통계량은 아래와 같다.




```
> summary(model[,c(8:15)])
       !                ?                upper             nchar                I                you              he.she            not        
 Min.   :0.000   Min.   :0.000   Min.   :0.0   Min.   :  7.00   Min.   :0.000   Min.   :0.000   Min.   : 0.0   Min.   :0.0  
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.0   1st Qu.: 41.00   1st Qu.:0.000   1st Qu.:0.000   1st Qu.: 0.0   1st Qu.:0.0  
 Median :0.000   Median :0.000   Median :0.0   Median : 68.00   Median :0.000   Median :0.000   Median : 1.0   Median :0.0  
 Mean   :0.085   Mean   :0.015   Mean   :0.1   Mean   : 82.27   Mean   :0.013   Mean   :0.103   Mean   : 1.3   Mean   :0.4  
 3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0.0   3rd Qu.:109.00   3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.: 2.0   3rd Qu.:0.0  
 Max.   :2.000   Max.   :3.000   Max.   :6.0   Max.   :479.00   Max.   :1.000   Max.   :4.000   Max.   :11.0   Max.   :6.0  
```

본격적인 파생변수의 활용에 앞서 기존의 데이터에 새롭게 추가한 파생변수들에 대한 기초 통계자료 및 긍부정별 출현에 대한 간단한 EDA분석을 실시해 보았다. 그래프에서 붉은색은 부정적 리뷰이고 푸른색은 긍정적 리뷰에서 발견된 파생변수임을 알려준다. 먼저 가장 눈에 띄는 것은 '?'의 출현이다. '?'는 가장 많이는 3번까지 발견되었는데, 80%가 넘는 부분이 모두 부정적인 표현에서 찾아졌음을 확인해 볼 수 있었다. 'Is it possible for a movie to get any worse than this?'와 같은 표현이 물음표를 사용한 대표적인 리뷰였는데, 자신이 영화에 대한 부정적인 감정을 표현하고 그것을 "다수의 사람들에게 너도 그렇게 생각하지?" 라는 식의 동의를 구하는 방법으로 물음표를 사용하거나, 강조를 위하여 많이 사용한 것을 확인해 볼 수 있다. 한편, 물음표와 비슷한 역할로 강조를 의미하는 느낌표는 물음표와 달리 긍정과 부정에 비슷한 비율로 사용된 것을 확인해 볼 수 있었다. 이를 통해 느낌표는 단지 자신의 의사를 강조하는 것이기 때문에 긍부정 모두에게서 높은 빈도로 찾아졌다고 추론할 수 있었다.

사람들을 리뷰를 쓸 때 전달하고자 하는 바에 따라서 지칭대명사를 다르게 사용한다. 예를 들면, 자신이 이 영화에 대하여 어떤 감정을 느꼈는지를 표현하기 위하여 'I'를 사용하기도 하며, 이 영화를 자신의 리뷰를 읽는 독자인 당신이 꼭 봐라 혹은 보지말라를 이야기하기 위해서는 'you'를 사용하기도 한다. 또한 영화에 등장하는 배우들을 이야기하거나 영화 자체를 이야기하기 위하여 'he'혹은 'she'를 사용하게 된다. 이러한 관점에서 위의 파생변수를 이해해 볼 수 있다. 먼저 'I'는 부정적 표현보다는 긍정적 표현에서 많이 찾아졌다. 나는 이 영화에 대하여 좋게 느꼈다고 평가하는 경우가 많은 것이다. 상대방에게 이 영화를 볼 것을 추천하는 경우 혹은 추천하지 않는 경우에 해당하는 'you'는 긍정과 부정 모두에서 비슷한 비율로 찾아지는 것을 확인해 볼 수 있었고, 영화의 주인공을 이야기하는 'he'나 'she'역시 긍부정에 대하여 비슷한 비율로 찾아볼 수 있었다. 이를 통해 볼 때 영화 리뷰를

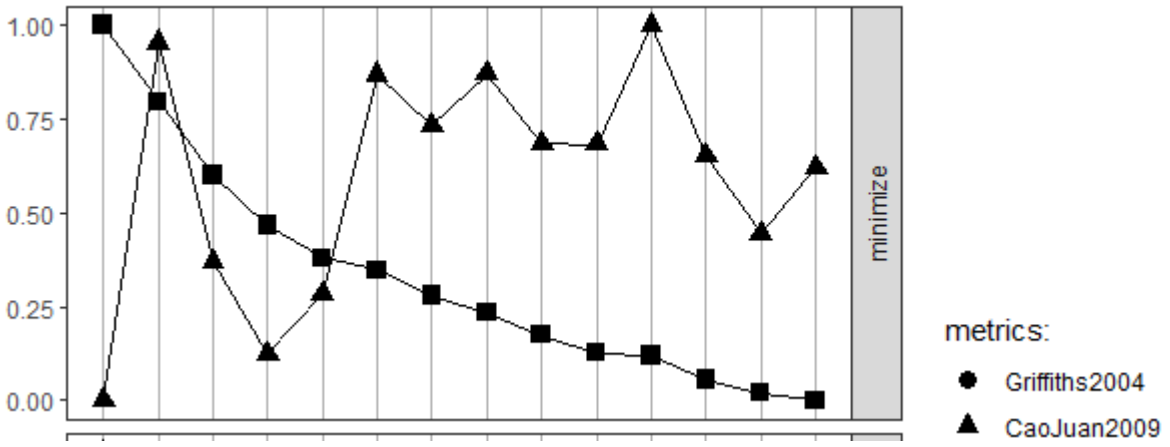
작성할 때 사람들은 영화에 대하여 긍정적인 평가를 할 때 자신이 영화에 대하여 어떻게 느꼈는지를 'I'라는 주어를 사용하여 서술하는 것을 알 수 있다.

upper_cnt는 연속된 대문자의 개수를 세어본 파생변수이다. 영어의 사용에 있어서는 문장의 시작을 대문자로 서술하여야 하지만, 중간에 사용된 연속된 대문자들의 등장은 의견을 강조하고 싶을 때 사용한 것으로 해석해야한다. 'VERY', 'BAD'와 같이 강조를 위하여 서술한 경우를 쉽게 찾아볼 수 있었다. 이러한 강조적 표현은 부정에서 많이 찾아질 것이라 흔히들 생각할 수 있지만, 긍정에서도 비슷한 비중으로 발견되었다. nchar는 문자열 개수를 의미하는 파생변수인데, 전체적으로 파악해 봤을 때는 부정적 리뷰라고 해서 글이 긍정적인 글보다 월등히 길거나 짧지 않았음을 확인해 볼 수 있었다.

4. 분석을 기반으로 한 EDA

4.1 토픽 분석(LDA/CTM)

(1) LDA 튜닝 알고리즘



CaoJuan(2009)은 LDA(Latent Dirichlet allocation)에서 최적의 k값을 찾기 위해 베이지안의 비모수적 방법에 기반한 HDP(hierarchical Dirichlet process)와의 혼합비율로 계산한다. 여기서 너무 많은 토픽을 가진다면 단어 간 의미가 서로 겹치게 되므로 식별 가능성을 유지하기에는 너무 많은 상관 관계가 발생할 수 있다. 따라서 k의 수가 너무 적어서 동질성이 떨어질 위험과 반대로 너무 많아서 군집 간 의미가 서로 겹치는 위험을 줄이기 위해, 이 방법을 바탕으로 본 분석에서는 적절한 k값을 3으로 설정하였다.

(2) 토픽 분석으로 나타난 영화 구성 요소

imdb_ctm			imdb_lda		
Topic.1	Topic.2	Topic.3	Topic.1	Topic.2	Topic.3

good	like	bad	good	like	bad
time	seen	character	time	seen	character
see	scene	great	see	scene	great
watching	funny	also	watching	funny	also
even	character	story	even	character	story
act	bad	point	act	bad	point
waste	think	script	waste	think	script
like	best	really	like	best	really
make	suck	totally	make	suck	totally
love	act	excellent	love	act	excellent
awful	ending	seen	thing	real	look
totally	music	plot	great	work	like

LDA분석 및 CTM분석을 종합하여 표로 정리하였다. 앞에서도 언급했던 바와 같이, 영화리뷰를 토픽 분석하는 것의 의의는 앞으로 설명할 영화의 주요한 구성요소를 토픽분석으로 파악할 수 있다는 것에 있다. 이에 따라, LDA 및 CTM분석에서 찾아지는 명사, 그중에서도 영화의 구성요소로 생각해 볼 수 있는 것들에 주목했다.

우선 좌측의 CTM의 세 번째 토픽에서의 'script', 'story', 두 번째 토픽에서 'scene' 등이 관찰되고, LDA의 세번째 토픽에서 'story', 두번째 토픽에서 'scene'가 관찰된다. 이를 종합해 영화의 구성요소로 composition즉, 영화의 전반적인 구성을 생각해 볼 수 있다. 작가가 써낸 대본이나 영화의 전반적인 흐름 혹은 엔딩과 같이 영화의 줄거리에 관하여 영화 관람객이 어떠한 의견을 보였는지를 살펴볼 것이다.

두 번째로 찾아볼 수 있는 것은 sound이다. CTM의 분석 결과 두 번째의 토픽에 'music'이 들어가는 데, 영화를 관람하는 것에 있어, sound 즉 음악 혹은 음향으로 분류할 수 있는 단어들이 많이 발견되지 않는 것으로 보아, 실제 음악과 관련하여 리뷰를 남긴 사람들이 많이 있는 것은 아니라는 것을 예상해 볼 수 있다. 영화가 시각적인 매체인 영상을 통해서 이루어지지만, 시각이라는 감각만으로 전달될 수 없다. 따라서 시각만큼 중요한 청각적 요소인 음향에 관하여 관람객들이 리뷰를 남겼다는 것을 알 수 있다.

세 번째로 CTM, LDA분석의 모든 부분에서 공통적으로 발견되는 'character', 'act'에 주목해 보았다. 영화를 생각하면, 영화 속 등장하는 캐릭터와 이를 연기하는 배우들을 제외하고 영화를 논하는 것은 불가능하다. 배역은 어떠하고, 이를 맡은 배우들의 연기력에 관한 평은 연기적 측면인 play라는 분류로 묶어서 살펴보고자 한다.

(3) 영화 구성 요소의 사전화

영화를 구성하는 많은 요소 중 가장 큰 비중을 차지하는 청각적 효과(sound), 영화의 전반적인 구성(composition), 그리고 연기 및 배역(play)을 뜻하는 동의어들을 포함한 사전을 만들어 분석에 사용하였다. 이는 전처리에서 어근 등을 분리해내기 전 원 텍스트에서의 텍스트의 형식적인 특성을 파악하기 위함이므로 전처리 이전에 해당 단어 포함 개수를 확인하여 새롭게 변수를 파생시켰다. 각 영화 구성요소에 관하여 사용한 사전은 아래와 같다.

① sound

```
dic_sound = c("sound", "audio", "accent", "harmony", "melody", "music", "noise",
"note", "tone", "vibration", "voice", "din", "intonation", "loudness", "modulation",
"pitch", "racket", "report", "resonance", "reverberation", "ringing", "softness",
"sonority", "sonorousness", "static", "tenor", "tonality", "sonance", "sonancy")
```

```
> table(sound)
0  1  2  3
938 55  6  1
```

데이터에서 찾아지는 sound 관련 단어를 사용한 리뷰는 1번 사용이 55회 2번사용이 6회, 3회사용이 1회로 총 62개가 찾아졌다. 위의 사전을 이용하여 찾아지는 문장들은 아래와 같았다.

[14] " With great sound effects, and impressive special effects, I can't recommend this movie enough. "

[10] "Special mention should be made of the superb music score and sound effects, which are an integral element in helping to make this such a memorable and enjoyable cartoon. "
[7] "While you don't yet hear Mickey speak, there are tons of sound effects and music throughout the film--something we take for granted now but which was a huge crowd pleaser in 1928. "

[12] "Also the music by Mark Snow is possibly the best score I've ever heard. "

② composition

```
dic_composition =c("plot", "action", "direct", "composition", "design", "movement",
"visual", "narrative", "graphic", "scenario", "scene", "scheme", "story", "structure",
"theme", "development", "enactment", "events", "incidents", "outline", "picture",
"progress", "subject", "suspense", "thread", "unfolding")
```

```
> table(composition)
0   1   2   3
874 107  18   1
```

데이터에서 찾아지는 composition 관련 단어를 사용한 리뷰는 총 126개 중, 1회 사용이 107개, 2회 사용이 18개로 나타났다. 위의 사전을 이용하여 찾은 문장들은 아래와 같았다.

```
[1] "Not only did it only confirm that the film would be unfunny and generic, but it
also managed to give away the ENTIRE movie; and I'm not exaggerating - every
moment, every plot point, every joke is told in the trailer.  "

[2] "The story itself is just predictable and lazy.  "

[3] "The only suspense I was feeling was the frustration at just how retarded the
girls were.  "
[4] "And generally the lines and plot is weaker than the average episode.
```

③ play

```
dic_play = c("act", "perform", "play", "portray", "burlesque", "characterize",
"dramatize", "emote", "feign", "ham", "impersonate", "mime", "mimic", "mug",
"parody", "personate", "personify", "pretend", "rehearse", "represent", "simulate",
"star", "stooge", "strut", "be on", "bring down the house", "do a turn", "go on", "go
over", "ham it up", "lay an egg", "make debut", "portrait", "put it over", "say one's
piece", "take part", "tread the boards", "character", "clown", "comedian",
"entertainer", "performer", "star", "villain", "amateur", "barnstormer", "foil", "ham",
"headliner", "idol", "impersonator", "lead", "mime", "mimic", "pantomimist",
"soubrette", "stand-in", "stooge", "thespian", "trouper", "role", "understudy",
"ventriloquist", "walk-on", "bit player", "hambone", "ingenue", "straight person",
"thesp")
```

```
> table(play)
  0   1   2   3   4   5   6   7   8
769 115  70  26   8   7   2   1   2
```

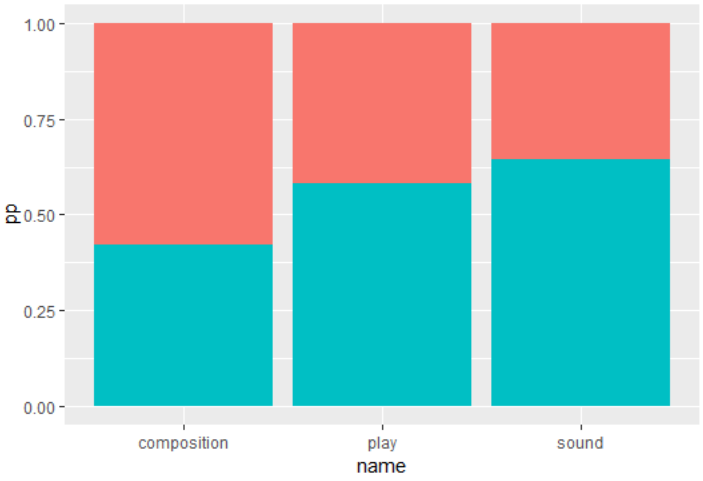
배우들의 연기 요소나 배역에 관한 요소들도 찾아볼 수 있었다. 이를 포함하는 리뷰는 총 231개였으며, 구체적인 예는 아래와 같다.

[50] "Judith Light is one of my favorite actresses and I think she does a superb job in this film! "

[44] "Great character actors Telly Savalas and Peter Boyle. "

[38] "I liked the way Dustin Hoffman's character was ready to do just about everything to stay with his son. "

[32] "There still are good actors around! "



앞서, 리뷰 데이터에서 드러난 영화의 주요한 구성 요소인 청각적 효과(sound), 영화의 줄거리 및 구성(composition), 그리고 연기 및 배역(play)이다. 이러한 요소들이 리뷰에서 각각 언급되어진 횟수를 의미하는 변수를 추가하고 시각화해보았다. 붉은색은 부정을 뜻하고 푸른색은 긍정을 뜻한다.

sound관련 사전에 포함된 단어를 인용하여 영화에 대하여 음향, 음악, 목소리에 관하여 리뷰를 남긴 사람들은 총 62명을 확인해 볼 수 있었다. 위의 구체적인 리뷰에서도 살펴볼 수 있었듯이 영화를 보는 내내 음악에 심취해 있었다는 등 음악적 요소를 칭찬하는 리뷰의 수가 조금은 더 우세한 모습을 확인해 볼 수 있다. 그러나 동시에 '음악뿐이었다' 혹은 '너무 시끄러웠다'와 같은 내용으로 부정적인 리뷰를 남긴 사람도 찾아볼 수 있었다.

영화의 전반적인 이야기나 구성(composition)에 관한 리뷰는 총126개를 찾을 수 있었다. 그래프를 통해 부정의 리뷰가 긍정의 리뷰보다 많이 관찰되는 것을 알 수 있었다. 다른 영화의 요소보다도 영화의 스토리라인이 너무 뻔해 예측가능 했다 혹은 다른 에피소드와 차별화 된것이 무엇인지 모르겠단는 내용이 담긴 부정적인 리뷰가 많이 존재했다. 내용적 요소에 긍정적인 답변을 남긴 사람들은 내용이 자연스럽게 흘러가서 좋았다 혹은 모든연령이 볼 수 있는 내용인것 같다는 식의 내용을 리뷰로 작성하였다. 이를 통해 볼 때, 영화의 요소적 측면에서 부정적인 리뷰가 많이 남겨지는 것에는 내용적인 부분이 많이 서술된다는 사실을 확인해 볼 수 있었다.

4.2 감성 분석

(1) 기계학습 알고리즘을 이용한 긍부정 분류

	SVM_LABEL	SVM_PROB	GLMNET_LABEL	GLMNET_PROB	SLDA_LABEL	SLDA_PROB	TREE_LABEL
1	1	0.7442355	1	0.5681311	1	1.0000000	2
2	1	0.5308365	1	0.5068653	2	1.0000000	2
3	1	0.9912822	1	0.8053394	1	1.0000000	1
4	1	0.8131095	1	0.9083518	1	1.0000000	2
5	2	0.8918334	2	0.8771200	2	1.0000000	2
6	2	0.9480784	2	0.8850922	2	1.0000000	2
7	2	0.8737208	2	0.7821509	2	1.0000000	2
8	2	0.5188143	1	0.5068653	2	1.0000000	2
9	2	0.8137703	2	0.7505231	2	1.0000000	2
10	1	0.5511892	1	0.8673414	1	1.0000000	2

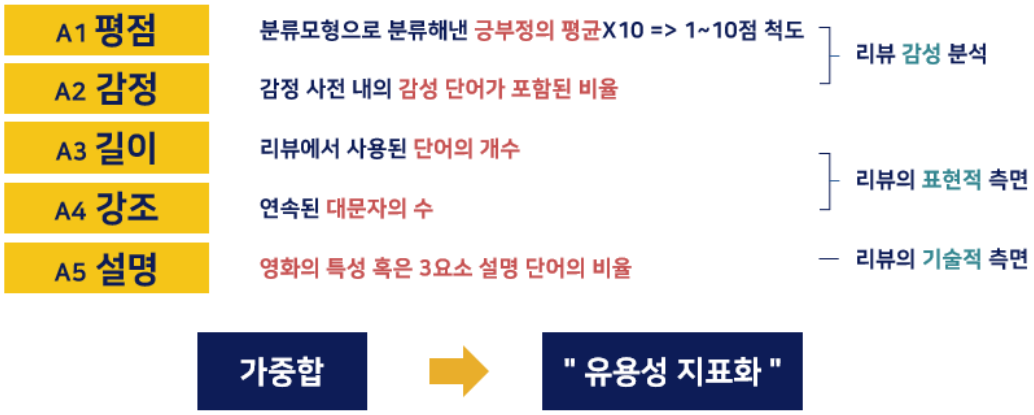
텍스트 데이터에 대하여 전처리 후 TF-IDF를 계산하여 이에 대하여 지도 기계학습을 실시한 결과는 위의 표와 같다. 여기서 알고리즘은 서포트벡터머신(SVM), 벌칙부과 일반화 선형 모형(GLMNET), 지도 LDA 모형(SLDA), 분류 및 회귀 나무(TREE), 배깅(BAGGING), 랜덤포레스트(Random Forest), 최대 엔트로피 모델링(MAXENT), 부스팅(BOOSTING), 인공신경망(NNET)을 사용하였다. 이 알고리즘들은 데이터를 각각 부정(1), 긍정(1)으로 분류하고 그 분류 확률을 보여주며 여기서 70:30의 cross validation 10-folds를 사용하였다. 예를 들면, 첫 번째 텍스트 데이터에 대하여 SVM에 의하여 긍부정을 분류한 결과, 부정이 나타났으며 이 값을 갖는 확률은 0.74였다. 이 알고리즘의 문서분류모형이 얼마나 테스트 데이터의 감정상태를 잘 예측하는지 살펴보면 정확도가 약 76.7%로 나타난다.

```
myresult$SVM_LABEL   0   1
                 1 379 112
                 2 121 388
```

Ⅲ. 본론 2

1. 리뷰의 유용성 지표 구성 변수

리뷰의 유용성이란, 실제 IMDB에서 리뷰를 평가한 사람들 중 실제 도움이 되었다고 체크한 사람들의 비율을 뜻한다. 앞서 데이터의 EDA, 토픽분석 및 감성분석을 바탕으로, 유용성 지표에 포함해야 할 변수를 설정하였다. EDA를 통해서, 각 리뷰의 길이 및 사용 단어수를, 토픽 분석을 바탕으로 영화의 요소에 대한 내용 언급 유무, 감성 분석을 통하여 감성을 포함한 서술의 언급에 관한 것과 리뷰의 긍정 및 부정의 정도를 알아낼 수 있었다. 이를 바탕으로 감정의 정도를 평점화한 것, 감성 단어의 사용 비율, 영화의 내적 요소에 대한 설명 비율, 리뷰의 표현적 측면을 고려한 5개의 변수를 유용성 지표를 구성하는데 사용할 것이다.



첫째, 본론 1의 4.2에서 실시한 감성분석을 통하여, 분류모형으로 분류해 낸 긍정 및 부정의 평균을 구하였다. 이에 10을 곱하여, 이를 1점부터 10점으로 나타나는 척도로 만들어 이 변수를 “평점”이라고 명명하였다. 이것이 의미하는 바는, 총 8가지의 기계학습 모형에 의하여 분류된 긍부정 점수의 평균이 높을 수록, 분류 모형이 해당 텍스트를 긍정으로 분류할 비율이 높다는 뜻인데 이는 곧 텍스트가 긍정일 가능성, 즉 긍정의 강도를 의미한다고 판단하였다. 따라서 긍부정의 평균은 곧 영화의 평점에 직결되며 이를 리뷰의 감성적 측면이라고 생각하였다.

둘째, 리뷰가 담고 있는 감성이라는 측면을 알기 위하여, a-finn 사전에 포함된 단어를 리뷰에서의 언급 여부를 분석하여, 이를 전체 단어 대비 감성 단어의 사용 비율을 의미하는 “감성”이라는 변수로 만들었다. a-finn 사전을 이용하는 데 있어서, 중복으로 감성 단어의 횟수를 세는 것을 방지하기 위하여 a-finn 사전을 스테밍한 후 사용하였다. 예를 들면, ‘wasted two hours.’라는 단어에서 ‘waste’는 a-finn 사전에서 부정 단어이므로 이를 이 리뷰의 총 단어의 개수 대비 감정적 단어의 수라고 판단하여 리뷰의 감성 단어 사용 비율을 측정하였다.

셋째, 리뷰의 표현적 측면을 고려할 수 있도록 LDA를 통해서 알아낸 각 리뷰의 총 사용 단어 수 및 리뷰의 내용을 강조하는데 사용된 대문자 단어의 사용 개수를 의미하는 “길이” 및 “강조” 변수를 추가하였다. 예를 들어, 리뷰어가 단어를 많이 사용할수록 이 리뷰어가 리뷰를 통해 영화에 대하여 느낀 본인의 생각과 정보에 대하여 많이 설명하려고 하며 ‘This movie is BAD’와 같이 대문자를 통해 자신의 의견을 강조하려고 한다고 파악하였다.

마지막으로, 본론 1-4.1의 토픽 분석을 통해 파악할 수 있었던 영화의 3가지 내적인 요소의 언급이 리뷰의 유용성을 판단하는데 중요한 요소로 보았다. 이를 위해 3가지 요소를 의미하는 단어를 포함하는 영화 구성요소(factor) 사전을 구축하였다. 따라서 영화의 요소를 언급하는 단어의 비율을 포함하였다. 예를 들면, 영화의 구성요소인 'play' 사전에는 'act', 'perform'과 같은 단어들이 포함되며 실제 리뷰에서는 'The act was decidely wooden' 와 같은 텍스트가 존재한다. 이 리뷰는 영화의 연기에 대하여 지적을 하고 있고 영화의 연기에 대하여 중요시 여기는 소비자들에게는 이 리뷰가 영화의 선택에 분명히 도움이 될 것이므로 이 리뷰는 유용하게 사용될 것이다. 따라서 이 영화의 요소들에 대하여 기술하고 있는지를 고려하기 위해 이 변수를 이용하였다.

리뷰의 내용적, 형식적, 표현적 측면을 고려한 5개의 변수를 리뷰의 유용성 지표를 구성하는 변수로서 사용하고자 한다.

2. 유용성에 대한 지표화 구현

위에서 고려한 변수들로 리뷰의 유용성을 지표화해보자. 가장 단순한 형태로 가중합을 생각할 수 있다. 각 변수들에 가중치를 곱한 뒤 선형 결합의 형태로 리뷰의 유용성의 추정치를 구할 것이다. 하지만 이렇게 가중치를 두는 것에 데이터를 기반한 근거가 부족하며 논리적으로 적합하지 않다고 판단하여 새로운 데이터를 크롤링하여 실제로 'helpfulness'가 어떻게 작용하는지 확인해보았다.

3. 지표화 모형 적합

```
> summary(stp_reg)

Call:
lm(formula = helpful ~ point + word_cnt + factor_cnt)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33359 -0.07318 -0.00706  0.06425  0.28280

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.964e-01  1.144e-02  43.398   <2e-16 ***
point        1.947e-02  1.273e-03  15.292   <2e-16 ***
word_cnt     9.497e-05  3.688e-05   2.575   0.0102 *
factor_cnt  -1.095e-03  7.235e-04  -1.514   0.1305
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1024 on 630 degrees of freedom
Multiple R-squared:  0.2823,    Adjusted R-squared:  0.2789
F-statistic:  82.6 on 3 and 630 DF,  p-value: < 2.2e-16
```

위의 그림은 실제 크롤링을 통해 얻은 10108개의 데이터를 이용하여 본 연구에서 유용성을 나타낸다고 가정한 리뷰의 평점, 감정 데이터의 비중, 리뷰에 사용된 단어의 수, 강조된 단어의 수, 영화의 구성요소에 대한 설명의 비중을 독립변수로 두고, 실제 유용성을 종속변수로 하여 stepwise regression을 행한 결과이다. 위 그림에 따르면 감정 데이터의 비중과 강조된 단어의 수는 helpful을 설명하는 데에 큰 영향을 미치지 않음을 알 수 있다.

이를 다시 원 데이터에 적용시켰고 이에 대한 데이터 셋은 다음과 같이 나타났다.

	text	point	avrg.x	helpful	word_cnt	factor_cnt	sent_cnt	power_cnt
6	A very, very, very slow-moving, aimless movie about a distres...	1	0.125	0.5000661	12	0	2	0
7	Not sure who was more lost - the flat character or the audien...	1	0.125	0.4983287	18	2	4	0
8	Attempting artiness with black & white and clever camera an...	1	0.000	0.4970998	30	2	6	0
9	Very little music or anything to speak of.	1	0.500	0.5063110	7	1	0	0
10	The best scene in the movie was when Gerardo is trying to fin...	2	0.875	0.5150320	20	1	2	0
11	The rest of the movie lacks art, charm, meaning... If it's about ...	1	0.625	0.5088748	19	0	8	0
12	Wasted two hours.	1	0.125	0.4988172	2	0	1	0
13	Saw the movie today and thought it was a good effort, good ...	2	1.000	0.5164681	14	0	5	0
14	A bit predictable.	1	0.250	0.5016148	2	0	0	0
15	Loved the casting of Jimmy Buffet as the science teacher.	2	1.000	0.5167189	9	0	2	0
16	nd those baby owls were adorable.	2	0.375	0.5041919	5	0	1	0
17	The movie showed a lot of Florida at it's best, made it look ver...	2	1.000	0.5168182	14	0	4	0

Ⅳ. 결론

1. 분석 결과

단계적 변수선택법에 의한 회귀분석 실행 결과, 본 모델의 R-square 값은 약 0.28로 높지 않은 수치가 나왔다. 이는 실제 유용성을 평가하기에 해당 데이터 혹은 변수들로는 부족함이 있다고 할 수 있다. 평점의 경우, 유용성과의 유의확률이 0.05 이하로 유의미하게 나왔다. 유용성과는 양의상관관계를 가짐에 따라 평점이 높을수록 유용성이 높게 나타남을 알 수 있다. 길이 요인의 경우도 유용성과의 유의확률이 0.05 이하로 유의미하다고 할 수 있다. 하지만 그 계수는 평점보다도 낮게 나와 평점에 비해 상대적으로 영향력이 낮다는

것을 알 수 있다. 리뷰의 질을 나타내는 설명비율 변수의 경우 유의확률이 0.13으로 다소 높게 나왔으며, 계수도 음수로 나와 유용성 지표화 시의 예상과는 다른 결과를 보였다.

2. 분석에 대한 고찰

본 연구의 최종 목표는 소비자들이 리뷰의 유용성을 평가할 때 미치는 요인들을 알아보는 것이었으나, 앞서 설정한 5가지 요인과 분석 방법으로서는 소비자들의 기준을 모두 설명할 수 없었다.

우선, 데이터와 변수에 관하여 논해보자. 우선적으로 주어진 imdb_labelled.txt 데이터는 1000개의 짧은 리뷰 텍스트의 긍정 및 부정을 labelling한 것으로 실제로 머신러닝 알고리즘 등을 이용하여 긍정 및 부정을 최적으로 분류해내는 것이 주 목적인 데이터였다. 따라서 우리의 분석 목적인 리뷰의 유용성에 관하여 논하기 위해선 다른 부가적인 리뷰 고유의 특성이 필요했다. 따라서 분석 내에서도 설명력이 많이 낮아졌으며 데이터의 한계로 변수들을 더 고려하지 못하였다. 설명력을 높이기 위하여 고려하였으나, 실제 데이터가 부족하여 활용할 수 없었던 요인들은 다음과 같다. 첫째, 이 데이터 분석에서 가장 중요했던 리뷰의 유용성 평가 수이다. 실제 IMDB에서는 리뷰의 유용성에 대하여 투표를 하는 개념이 존재했다고 위에서 언급했다. 이에 대한 데이터를 받아 반응변수로 두고 다양한 회귀 방법 혹은 분류 모형을 통해 Importance of variables를 알아낼 수 있다면 이것이 본 연구가 원하는 방향으로 나아갔을 것이다. 따라서 유용성 평가 투표 수라는 분석에 매우 중요한 변수를 데이터의 한계로 고려할 수 없었다. 둘째, 리뷰어가 남긴 리뷰의 수를 고려할 수 있다. 리뷰어가 남긴 리뷰의 수는 곧 해당 리뷰어의 노출도를 뜻한다. 해당 리뷰어가 남긴 리뷰가 유용했다고 한 평가가 많았다면, 해당 리뷰어가 남긴 다른 리뷰도 유용하다 판단하기 쉬울 것이다. 따라서 리뷰어가 남긴 리뷰들의 평판 및 리뷰의 수 또한 변수로 고려할 수 있다.

다음으로 분석 방법 측면에서 논해보자. 분명 유용성에 대한 변수가 존재했다면 회귀분석 혹은 로지스틱 회귀분석을 통해 각 변수들에 대하여 유용성에 미치는 영향이 통계적으로 유의미한지, 그리고 그 영향력이 얼마나 되는지 알 수 있었을 것이다. 하지만 데이터의 한계로 알 수 없었고 실제 크롤링한 데이터에서 모형 적합을 시키고 추정된 모형에 대하여 원 데이터에 적용시킬 수밖에 없었다. 그러나 유용성에 대한 지표를 위에서 언급했던 것처럼 임의적으로 선형결합을 시켜서 두 개 혹은 세 개의 분류로 임의적으로 나눈 후 머신러닝 등의 알고리즘을 통해 분류해내는 방법 또한 존재할 수 있었으나 역시 기준의 모호함으로 이 분석에서는 고려하지 않았다. 또한 각 변수들의 거리 개념의 코사인 유사도를 고려하여 군집분석을 실시하였으나 정확도가 또한 낮았고 라벨링의 기준 또한 애매모호하여 이 연구에서 고려하지 않았다.

3. 분석에 대한 제언점

온라인 리뷰의 양이 증가하며, 그러한 방대한 양의 리뷰 중 자신에게 도움이 되는 리뷰를 찾아내는 것도 중요한 이슈가 되고 있다. 이러한 배경 속에서 본 연구를 통해서 얻은

지표화된 유용성은 실생활에서도 유의미하게 활용될 수 있을 것이다. 본 분석을 통해 도출한 지표를 활용함으로써 IMDB측에서는 잠재적 소비자들에게 양질의 리뷰를 제공할 수 있고, 소비자인 리뷰어 입장에서는 유용한 리뷰 쓰기를 위한 가이드라인을 제공받을 수 있다는 것에 분석의 의의가 있다. 또한 본 모델에서 제시한 리뷰어, 리뷰의 특징들은 다양한 포맷에서 쉽게 얻을 수 있기에 어렵지 않게 가공할 수 있는 정보라는 점에서 활용도가 높다고 판단된다. 소비자의 반응이 충분히 확보되지 않은 상황에서 중요한 리뷰를 사전에 식별하는 역할도 수행할 수 있을 것이다.