Sentiment Analysis

권기남 박수영 유민형 조수범

CONTENT

01 데이터 탐색 및 전처리

식당 이용 후기
긍정 500개
부정 500개

	V1	V2
7	Crust is not good.	Neg
8	Not tasty and the texture was just nasty.	Neg
11	Now I am getting angry and I want my damn pho.	Neg
12	Honeslty it didn't taste THAT fresh.)	Neg
13	The potatoes were like rubber and you could tell they h...	Neg
17	Would not go back.	Neg
18	The cashier had no care what so ever on what I had to s...	Neg
20	I was disgusted because I was pretty sure that was hum...	Neg
21	I was shocked because no signs indicate cash only.	Neg
23	Waitress was a little slow in service.	Neg
24	This place is not worth your time, let alone Vegas.	Neg

데이터 탐색 및 전처리 - 인코딩 문제

My fiancé and I came in the middle of the day and we were greeted and seated right away.

My fianc챕 and I came in the middle of the day and we were greeted and seated right away."

My fiance and I came in the middle of the day and we were greeted and seated right away.

강세 표시(tilde)가 인코딩이 되지 않는 문제 해결

데이터 탐색 및 전처리 - library(tm)

```
yelp_corpus <- Corpus(VectorSource(yelp$V1))
corpus_clean <- tm_map(yelp_corpus, tolower)
corpus_clean <- tm_map(corpus_clean, removeNumbers)
corpus_clean <- tm_map(corpus_clean, removeWords, stopwords())
corpus_clean <- tm_map(corpus_clean, removePunctuation)
corpus_clean <- tm_map(corpus_clean, stripWhitespace)
corpus_clean <- tm_map(corpus_clean, stemDocument, language = "english")
```

```
[1] Crust is not good.
[2] Not tasty and the texture was just nasty.
[3] Now I am getting angry and I want my damn pho.
```

```
[1] crust good
[2] tasti textur just nasti
[3] now get angri want damn pho
```

01 데이터 탐색 및 전처리 - train/test 구분

```
set.seed(123456789)
smp = c(sample(1:500, 350), sample(501:1000, 350))

yelp_raw_train <- yelp[smp, ]
yelp_raw_test  <- yelp[-smp, ]

yelp_dtm_train <- yelp_dtm[smp, ]
yelp_dtm_test  <- yelp_dtm[-smp, ]

yelp_corpus_train <- corpus_clean[smp]
yelp_corpus_test  <- corpus_clean[-smp]
```

```
> prop.table(table(yelp_raw_train$V2))
  0   1
0.5 0.5
```

```
> prop.table(table(yelp_raw_test$V2))
  0   1
0.5 0.5
```

데이터 탐색 및 전처리 - 긍정/부정 주요 단어 확인

최소 빈도수 5 이상으로 DocumentTermMatrix 생성

02 기초 분석 - 나이브 베이즈

actual	predicted 0	1	Row Total
0	114	36	150
	0.722	0.254	
	0.380	0.120	
1	44	106	150
	0.278	0.746	
	0.147	0.353	
Column Total	158	142	300
	0.527	0.473	

Accuracy = (114+106)/300 = 0.73

02 기초 분석 - 나이브 베이즈 (Laplace = 1)

actual	predicted 0	1	Row Total
0	115 0.728 0.383	35 0.246 0.117	150
1	43 0.272 0.143	107 0.754 0.357	150
Column Total	158 0.527	142 0.473	300

Accuracy = (115+107)/300 = 0.74

기초 분석 - KNN (K=10)

actual	predicted 0	1	Row Total
0	143	7	150
	0.577	0.135	
1	105	45	150
	0.423	0.865	
Column Total	248	52	300
	0.827	0.173	

Accuracy = (143+45)/300 = 0.63

02 기초 분석 - KNN (K=20)

	predicted		
actual	0	1	Row Total
0	147	3	150
	0.527	0.143	
1	132	18	150
	0.473	0.857	
Column Total	279	21	300
	0.930	0.070	

Accuracy = (147+18)/300 = 0.55

02 기초 분석 - 의사결정나무(C5.0)

	predicted		
actual	0	1	Row Total
0	135	15	150
	0.584	0.217	
1	96	54	150
	0.416	0.783	
Column Total	231	69	300
	0.770	0.230	

Accuracy = (135+54)/300 = 0.63

02 기초 분석 - 의사결정나무(C5.0, 부스팅10회)

	predicted		
actual	0	1	Row Total
0	135	15	150
	0.587	0.214	
1	95	55	150
	0.413	0.786	
Column Total	230	70	300
	0.767	0.233	

Accuracy = (135+55)/300 = 0.63

02 기초 분석 - SVM(kernel=Linear)

	predicted		
actual	0	1	Row Total
0	120	30	150
	0.779	0.205	
1	34	116	150
	0.221	0.795	
Column Total	154	146	300
	0.513	0.487	

Accuracy = (120+116)/300 = 0.78

02 기초 분석 - SVM(kernel=Gaussian)

	predicted		
actual	0	1	Row Total
0	134	16	150
	0.740	0.134	
1	47	103	150
	0.260	0.866	
Column Total	181	119	300
	0.603	0.397	

Accuracy = (134+103)/300 = 0.79

기초 분석 - 정리

	나이브 베이즈		KNN		의사결정나무		SVM	
옵션	기본	Laplace=1	K=10	K=20	기본	부스팅 10회	Linear	Gaussian
정확도	0.73	0.74	0.63	0.55	0.63	0.63	0.78	0.79
KAPPA	0.47	0.48	0.25	0.10	0.26	0.27	0.57	0.58
F-score	0.73	0.73	0.45	0.21	0.49	0.50	0.78	0.77

추가 분석

"The ambiance isn't much better."
"It was packed!!"
"The atmosphere is modern and hip, while maintaining a touch of coziness."
"We are so glad we found this place."
"Delicious and I will absolutely be back!"
"Google mediocre and I imagine Smashburger will pop up."
"It is PERFECT for a sit-down family meal or get together with a few friends."
"The bathrooms are clean and the place itself is well decorated."
"I had about two bites and refused to eat anymore."
"All in all, I can assure you I'll be back."

전처리를 하지 않고 텍스트의 형식적인 특성을 파악

->> 새로운 파생 변수 생성

추가 분석 - 파생 변수

번호	변수	설명
1	a1	느낌표 !의 개수
2	a2	물음표 ?의 개수
3	a3	문자열의 길이
4	a4	마침표 .의 개수
5	a5	쉼표 ,의 개수
6	a6	대문자의 개수
7	a7	숫자의 개수
8	a8	달러 기호 $의 개수
9	a9	작은 따옴표 ‘의 개수
10	a10	띄어쓰기의 개수

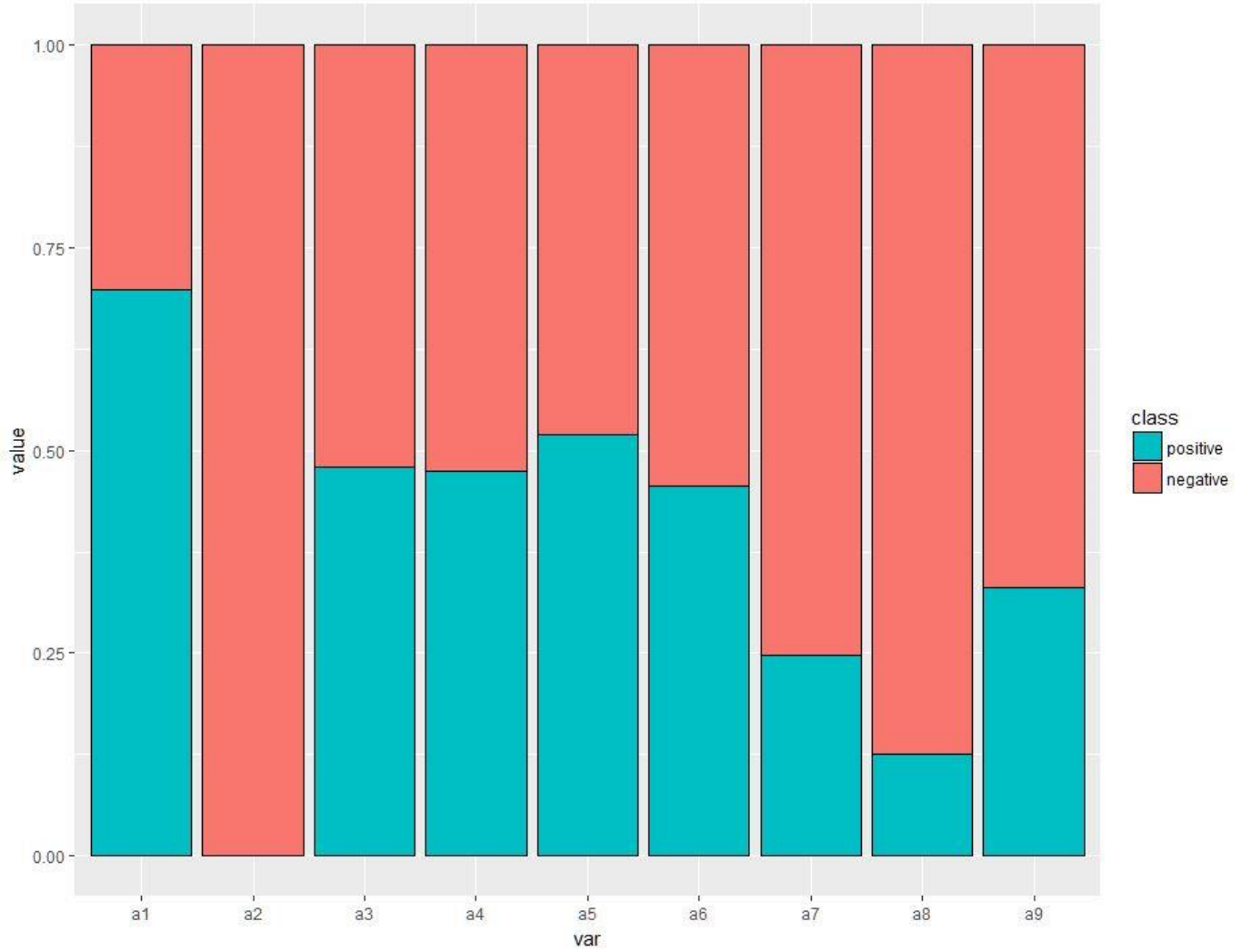
추가 분석 - 파생 변수

번호	변수	설명
11	a11	괄호 열기 (의 개수
12	a12	괄호 닫기)의 개수
13	a13	&의 개수
14	a14	큰 따옴표 “의 개수
15	a15	콜론 :의 개수
16	a16	하이픈 -의 개수
17	a17	문자열에서 공백을 뺀 길이
18	a18	-ly로 끝나는 단어의 개수
19	a19	세미 콜론 ;의 개수
20	a20	the의 개수

추가 분석 - 파생 변수

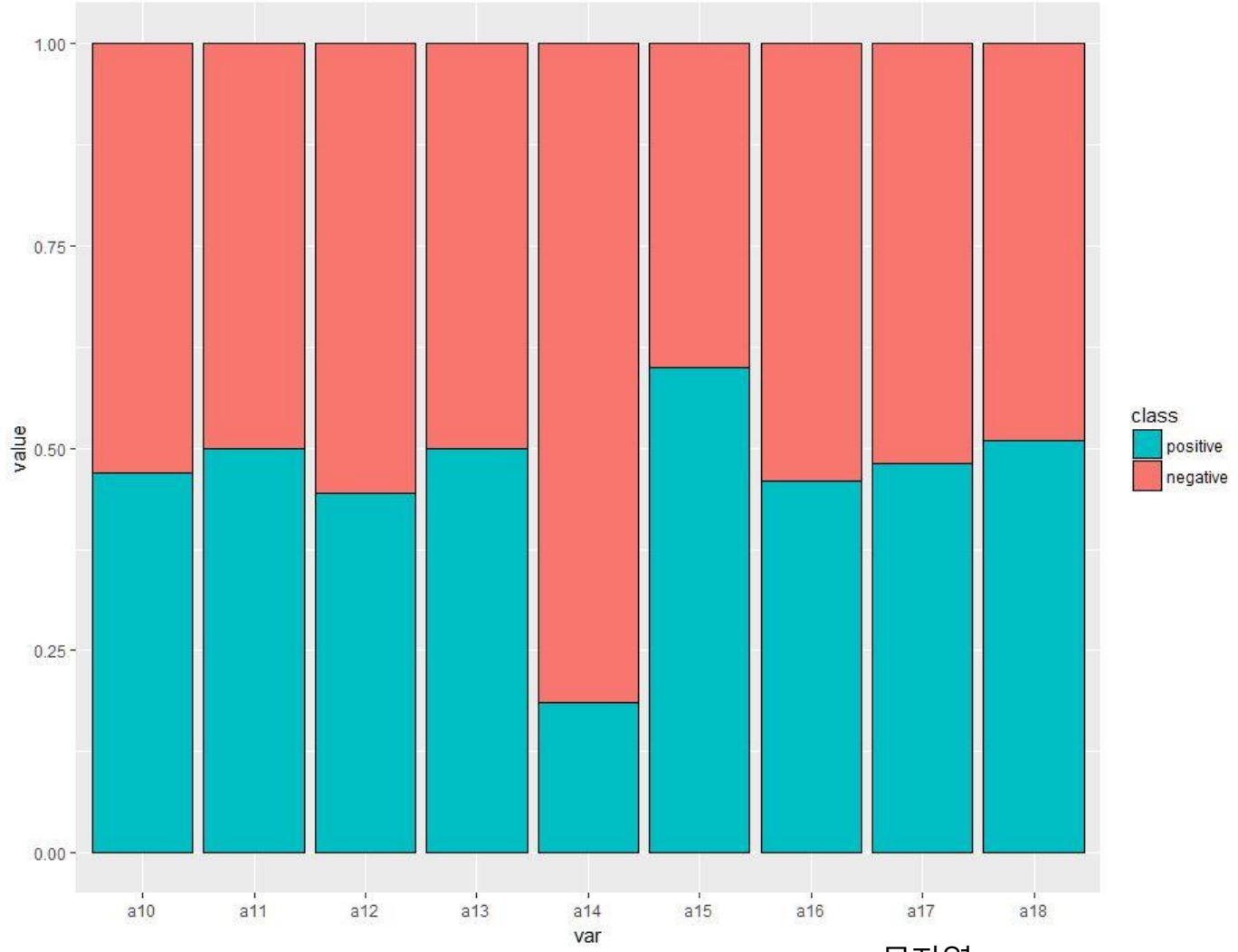
번호	변수	설명
21	a21	부정 표현의 개수
22	a22	미래 표현의 개수
23	a23	so의 개수
24	a24	much의 개수
25	a25	too의 개수
26	a26	this의 개수
27	a27	my의 개수

추가 분석 - 파생 변수



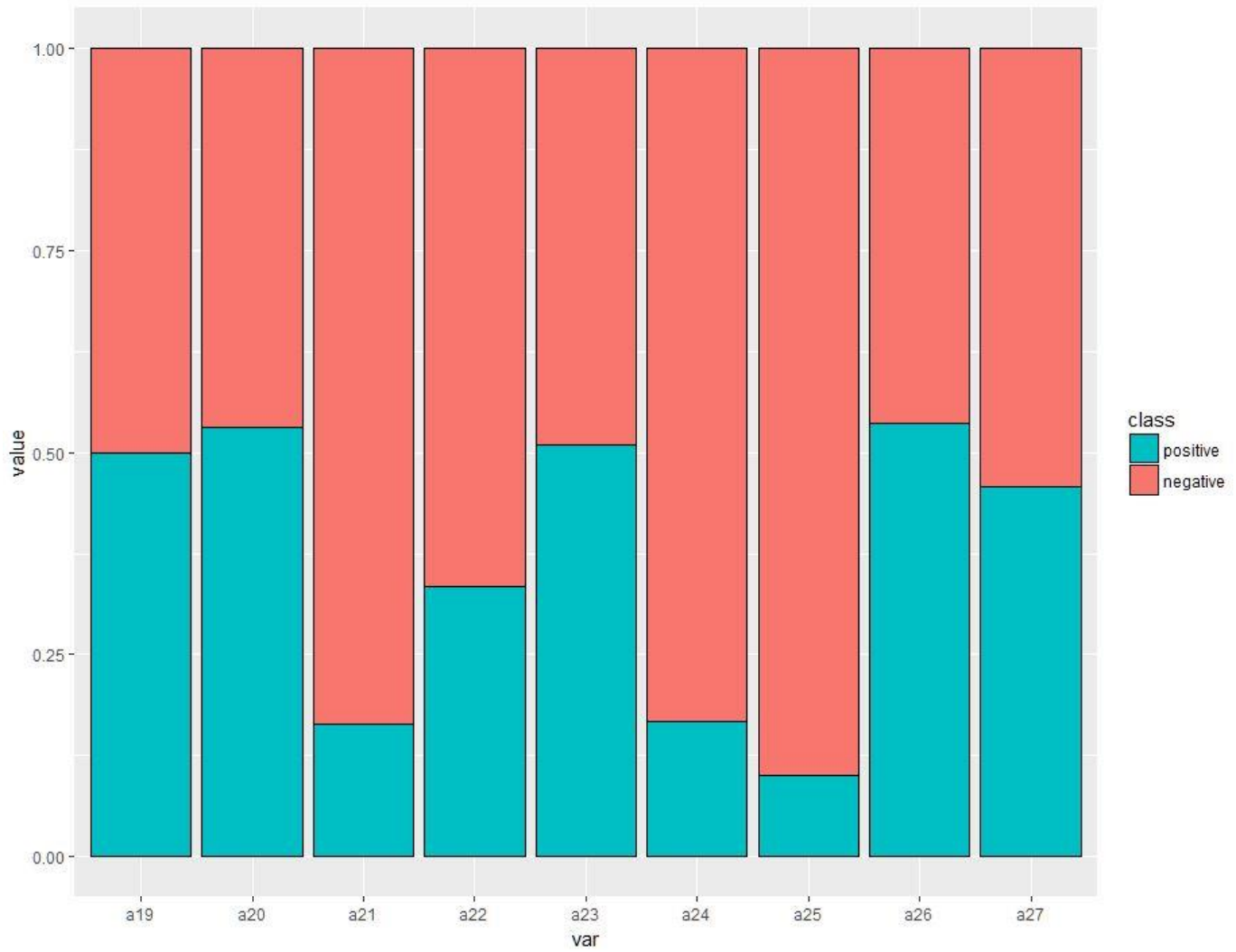
! ? 문자열 . , 대문자 숫자 $ '

추가 분석 - 파생 변수



띄어쓰기 ( ) & " : - 문자열 - 공백 -ly

추가 분석 - 파생 변수



; the 부정 미래 so much too this my

추가 분석 - 파생 변수를 이용한 분석

	나이브 베이즈		KNN		의사결정나무		SVM	
옵션	기본	Laplace=1	K=10	K=20	기본	부스팅 10회	Linear	Gaussian
정확도	0.59	0.59	0.55	0.50	0.67	0.69	0.69	0.53
KAPPA	0.17	0.17	0.11	0.01	0.33	0.39	0.37	0.07
F-score	0.70	0.70	0.59	0.52	0.71	0.71	0.74	0.57

04 DTM 모델 개선 - 나이브 베이즈

```
> m
Naive Bayes

1000 samples
 187 predictor
   2 classes: 'Neg', 'Pos'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 1000, 1000, 1000, 1000, 1000, 1
Resampling results across tuning parameters:

  usekernel  Accuracy   Kappa
  FALSE           NaN         NaN
   TRUE      0.4994125  0.01474478

Tuning parameter 'fL' was held constant at a value of 0
Tuning parameter 'adjust' was held constant
 at a value of 1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were fL = 0, usekernel = TRUE and adjust = 1.
```

```
> ctrl = trainControl(method = "cv", number = 10, selectionFunction = "best")
> grid_nb = expand.grid(.fL = 0, .usekernel = TRUE, .adjust = 1)
> set.seed(123456789)
> m_nb = train(cl~., data = a, method = "nb", metric = "Accuracy",
+              trControl = ctrl, tuneGrid= grid_nb)
> m_nb
Naive Bayes

1000 samples
 187 predictor
   2 classes: 'Neg', 'Pos'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 900, 900, 900, 900, 900, 900, ...
Resampling results:

  Accuracy  Kappa
  0.54      0.08

Tuning parameter 'fL' was held constant at a value of 0
Tuning parameter 'usekernel' was held
 constant at a value of TRUE
Tuning parameter 'adjust' was held constant at a value of 1
```

04 DTM 모델 개선 - KNN

```
> m_knn0
k-Nearest Neighbors

1000 samples
1533 predictors
   2 classes: '0', '1'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 1000, 1000, 1000, 1000,
Resampling results across tuning parameters:

  k  Accuracy   Kappa
  5  0.6429754  0.2836535
  7  0.6312580  0.2594069
  9  0.6201921  0.2365566

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 5.
```

```
> m_knn
k-Nearest Neighbors

1000 samples
1533 predictors
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 900, 900, 900, 900, 900, 900, ...
Resampling results:

  Accuracy  Kappa
  0.648     0.296

Tuning parameter 'k' was held constant at a value of 5
```

04 DTM 모델 개선 - 의사결정나무

```
> model_tree
C5.0

1000 samples
1532 predictors
   2 classes: 'Neg', 'Pos'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 1000, 1000, 1000, 1000, 1000, 1000,
Resampling results across tuning parameters:

  model  winnow  trials  Accuracy   Kappa
  rules  FALSE     1      0.6854952  0.3686468
  rules  FALSE    10      0.6395343  0.2813725
  rules  FALSE    20      0.6514717  0.3045806
  rules   TRUE     1      0.6494275  0.2963550
  rules   TRUE    10      0.5943169  0.1939169
  rules   TRUE    20      0.5919957  0.1873321
  tree   FALSE     1      0.6532955  0.3040713
  tree   FALSE    10      0.6500390  0.3009451
  tree   FALSE    20      0.6702155  0.3381568
  tree    TRUE     1      0.6393742  0.2756725
  tree    TRUE    10      0.6338196  0.2665551
  tree    TRUE    20      0.6416121  0.2802826

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were trials = 1, model = rules and winnow = FALSE.
```

```
> model_tree_1
C5.0

1000 samples
1532 predictors
   2 classes: 'Neg', 'Pos'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 900, 900, 900, 900, 900, 900, ...
Resampling results:

  Accuracy  Kappa
  0.672     0.344

Tuning parameter 'trials' was held constant at a value of 1
Tuning parameter 'model' was held constant at a value of rules

Tuning parameter 'winnow' was held constant at a value of FALSE
```

04 DTM 모델 개선 - SVM

```
> model_svm
Support Vector Machines with Linear Kernel

1000 samples
1532 predictors
   2 classes: 'Neg', 'Pos'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 1000, 1000, 1000,
Resampling results:

  Accuracy   Kappa
  0.7520721  0.5045104

Tuning parameter 'C' was held constant at a value of 1
```

```
> model_svm_tree1
Support Vector Machines with Linear Kernel

1000 samples
1532 predictors
   2 classes: 'Neg', 'Pos'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 900, 900, 900, 900, 900, 900,
Resampling results:

  Accuracy  Kappa
  0.765     0.53

Tuning parameter 'C' was held constant at a value of 1
```

04 파생 변수 모델 개선 - KNN

```
k-Nearest Neighbors

1000 samples
  27 predictor
   2 classes: 'Neg', 'Pos'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 1000, 1000, 1000, 10
Resampling results across tuning parameters:

  k  Accuracy   Kappa
  5  0.5311346  0.06235186
  7  0.5258168  0.05187247
  9  0.5204480  0.04129601

Accuracy was used to select the optimal model
The final value used for the model was k = 5.
```

	Neg	Pos
a21	100.000000	100.000000
a9	41.011401	41.0114
a1	38.833150	38.8331
a4	34.709683	34.7096
a10	32.470033	32.4700
a3	22.834657	22.8346
a17	20.708036	20.7080
a7	19.784268	19.7842
a20	18.538158	18.5381
a24	8.372521	8.3725
a6	6.344976	6.3449
a18	6.155199	6.1551
a25	5.581681	5.5816
a26	5.118401	5.1184
a14	4.201610	4.2016
a2	4.186261	4.1862
a8	4.186261	4.1862
a27	3.534599	3.5345
a22	2.790840	2.790840
a5	2.225695	2.225695
a12	1.395420	1.395420
a15	1.395420	1.395420
a13	1.378675	1.378675
a23	1.353558	1.353558
a16	1.322858	1.322858
a11	0.000000	0.000000
a19	0.000000	0.000000

```
k-Nearest Neighbors

1000 samples
   9 predictor
   2 classes: 'Neg', 'Pos'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 900, 900, 900, 900, 900, 900, ..
Resampling results:

  Accuracy  Kappa
  0.529     0.058

Tuning parameter 'k' was held constant at a value of 5
```

부정 표현, 작은 따옴표, 느낌표, 마침표, 띄어쓰기, 문자열, (문자열-공백), 숫자, the

04 파생 변수 모델 개선 - 의사결정나무

```
C5.0

1000 samples
  27 predictor
   2 classes: 'Neg', 'Pos'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 1000, 1000, 1000, 1000, 1000, 1000
Resampling results across tuning parameters:

  model  winnow  trials  Accuracy   Kappa
  rules  FALSE    1      0.6091589  0.2185481
  rules  FALSE   10      0.6154051  0.2308123
  rules  FALSE   20      0.6081192  0.2161601
  rules   TRUE    1      0.6039679  0.2080358
  rules   TRUE   10      0.6032810  0.2076126
  rules   TRUE   20      0.6126663  0.2259257
  tree   FALSE    1      0.6060520  0.2113378
  tree   FALSE   10      0.5985266  0.1969660
  tree   FALSE   20      0.6002792  0.2006134
  tree    TRUE    1      0.6026271  0.2044731
  tree    TRUE   10      0.6055522  0.2114446
  tree    TRUE   20      0.5986978  0.1974068

Accuracy was used to select the optimal model using the larg
The final values used for the model were trials = 10, model = rules and winnow = FALSE.
```

	Overall
a1	100.0
a21	100.0
a20	94.3
a26	84.1
a5	79.5
a18	73.5
a7	69.1
a6	53.9
a10	45.7
a9	44.9
a17	40.4
a23	23.8
a3	7.8
a27	3.6
a16	3.2
a2	0.6
a4	0.0
a8	0.0
a11	0.0
a12	0.0
a13	0.0
a14	0.0
a15	0.0
a19	0.0
a22	0.0
a24	0.0
a25	0.0

```
C5.0

1000 samples
  11 predictor
   2 classes: 'Neg', 'Pos'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 900, 900, 900, 900, 900, 900, ...
Resampling results:

  Accuracy  Kappa
  0.64      0.28

Tuning parameter 'trials' was held constant at a value of 10
Tuning parameter 'model' was held constant at
 a value of rules
Tuning parameter 'winnow' was held constant at a value of FALSE
```

느낌표, 부정 표현, the, this, 쉼표, -ly, 숫자, 대문자, 띄어쓰기, 작은 따옴표, so

04 파생 변수 모델 개선 - SVM

```
Support Vector Machines with Linear Kernel

1000 samples
  27 predictor
   2 classes: 'Neg', 'Pos'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 1000, 1000, 1000, 1000, 1000,
Resampling results:

  Accuracy   Kappa
  0.6663635  0.3327082

Tuning parameter 'C' was held constant at a value of 1
```

	Neg	Pos
a21	100.000000	100.000000
a9	41.011401	41.011401
a1	38.833150	38.833150
a4	34.709683	34.709683
a10	32.470033	32.470033
a3	22.834657	22.834657
a17	20.708036	20.708036
a7	19.784268	19.784268
a20	18.538158	18.538158
a24	8.372521	8.372521
a6	6.344976	6.344976
a18	6.15	
a25	5.58	
a26	5.11	
a14	4.20	
a2	4.18	
a8	4.18	
a27	3.53	
a22	2.79	
a5	2.22	
a12	1.39	
a15	1.39	
a13	1.37	
a23	1.35	
a16	1.32	
a11	0.00	
a19	0.00	

느낌표, 띄어쓰기, 느낌표, 마침표,
대문자, 문자열, (문자열-공백), the

```
Support Vector Machines with Linear Kernel

1000 samples
   9 predictor
   2 classes: 'Neg', 'Pos'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 900, 900, 900, 900, 900, 900, ...
Resampling results:

  Accuracy  Kappa
  0.661     0.322

Tuning parameter 'C' was held constant at a value of 1
```

04 파생 변수 모델 개선 - 로지스틱

```
Generalized Linear Model

1000 samples
  27 predictor
   2 classes: 'Neg', 'Pos'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 1000, 1000, 1000, 1000, 1000
Resampling results:

  Accuracy   Kappa
  0.6561904  0.314399
```

	Overall
a1	30.333142
a2	0.000000
a3	9.646927
a4	3.438
a5	24.801
a6	21.310
a7	32.672
a8	2.068
a9	5.458
a10	12.493
a11	4.271
a12	5.996
a13	1.821
a14	9.629
a15	7.576
a16	13.790719
a18	6.804186
a19	2.392533
a20	26.050578
a21	100.000000
a22	11.879283
a23	13.017631
a24	25.697315
a25	21.013832
a26	17.740757
a27	2.546372

```
Generalized Linear Model

1000 samples
  13 predictor
   2 classes: 'Neg', 'Pos'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 900, 900, 900, 900, 900, 900, ...
Resampling results:

  Accuracy  Kappa
  0.677     0.354
```

느낌표, 쉼표, 대문자, 숫자, 띄어쓰기, 하이픈, the, 부정표현, 미래표현, so, much, too, this

04 파생 변수 모델 개선 - 신경망

```
Neural Network

1000 samples
  27 predictor
   2 classes: 'Neg', 'Pos'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 1000, 1000, 1000, 1000, 1000, 1000, ...
Resampling results across tuning parameters:

  size  decay  Accuracy   Kappa
  1     0e+00  0.5227696  0.06766341
  1     1e-04  0.5406154  0.09962101
  1     1e-01  0.6365647  0.27319064
  3     0e+00  0.5917623  0.19111624
  3     1e-04  0.5974889  0.19578314
  3     1e-01  0.6238221  0.24790471
  5     0e+00  0.5865971  0.17636676
  5     1e-04  0.6024680  0.20828910
  5     1e-01  0.6110864  0.22369703

Accuracy was used to select the optimal model using the largest value
The final values used for the model were size = 1 and decay = 0.1
```

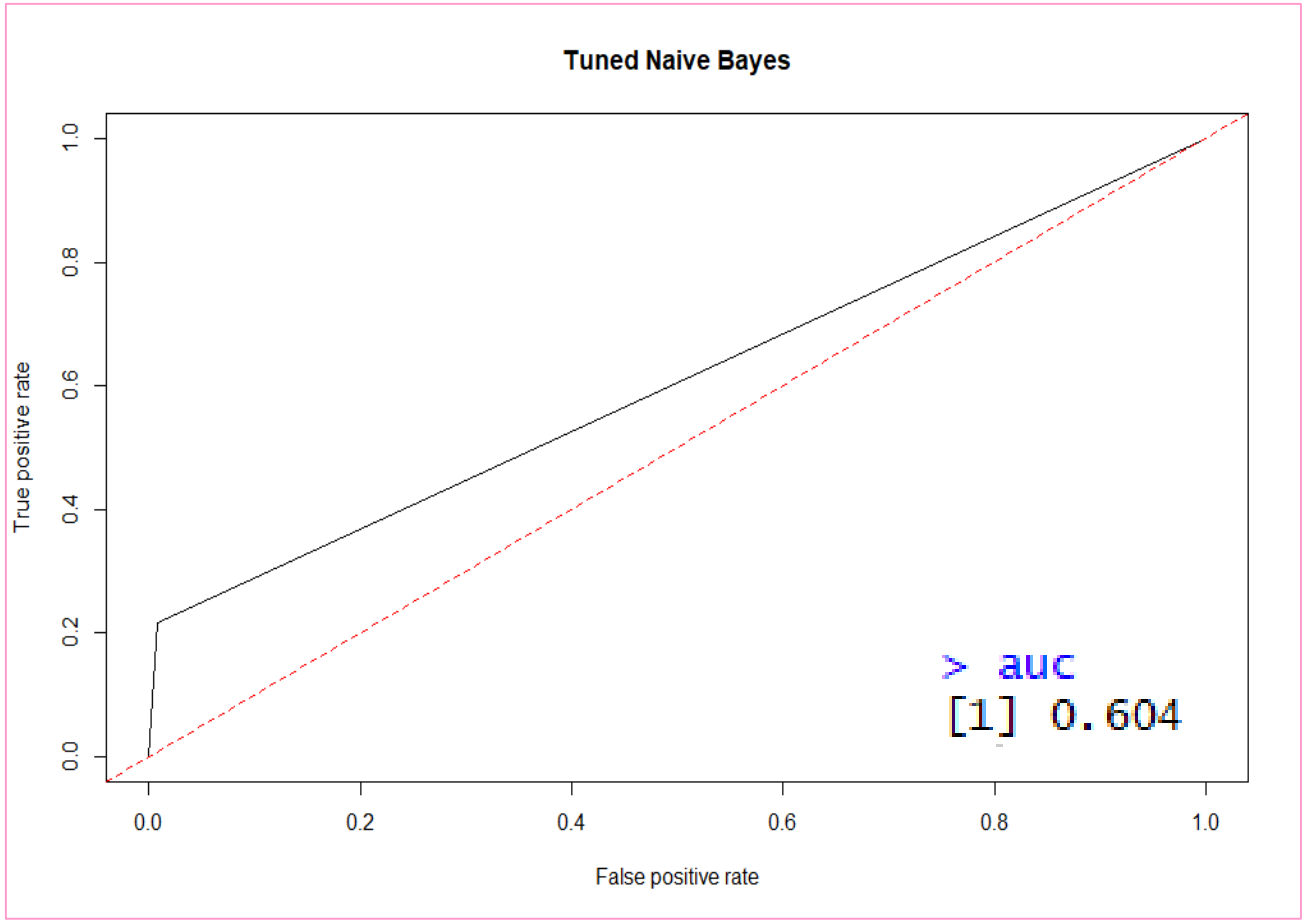
	Overall
a1	29.0448155
a2	100.0000000
a3	0.0000000
a4	0.87
a5	11.60
a6	1.75
a7	19.02
a8	33.80
a9	1.92
a10	0.87
a11	7.62
a12	11.22
a13	8.41
a14	12.57
a15	15.64
a16	12.21
a17	0.33
a18	4.29
a19	1.4234327
a20	9.1584486
a21	71.6257274
a22	17.3271159
a23	9.2203223
a24	83.7776188
a25	53.4807844
a26	12.4951409
a27	1.4593612

```
Neural Network

1000 samples
  15 predictor
   2 classes: 'Neg', 'Pos'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 900, 900, 900, 900, 900, 900, ...
Resampling results:

  Accuracy  Kappa
  0.667     0.334

Tuning parameter 'size' was held constant at a value of 1
Tuning parameter 'decay' was held constant at a
 value of 0.1
```

느낌표, 물음표, 쉼표, 숫자, 달러, 괄호 닫기, 큰따옴표, 콜론, 하이픈, 부정표현, 미래표현, much, too, this

DTM 성능 평가 – 나이브 베이즈

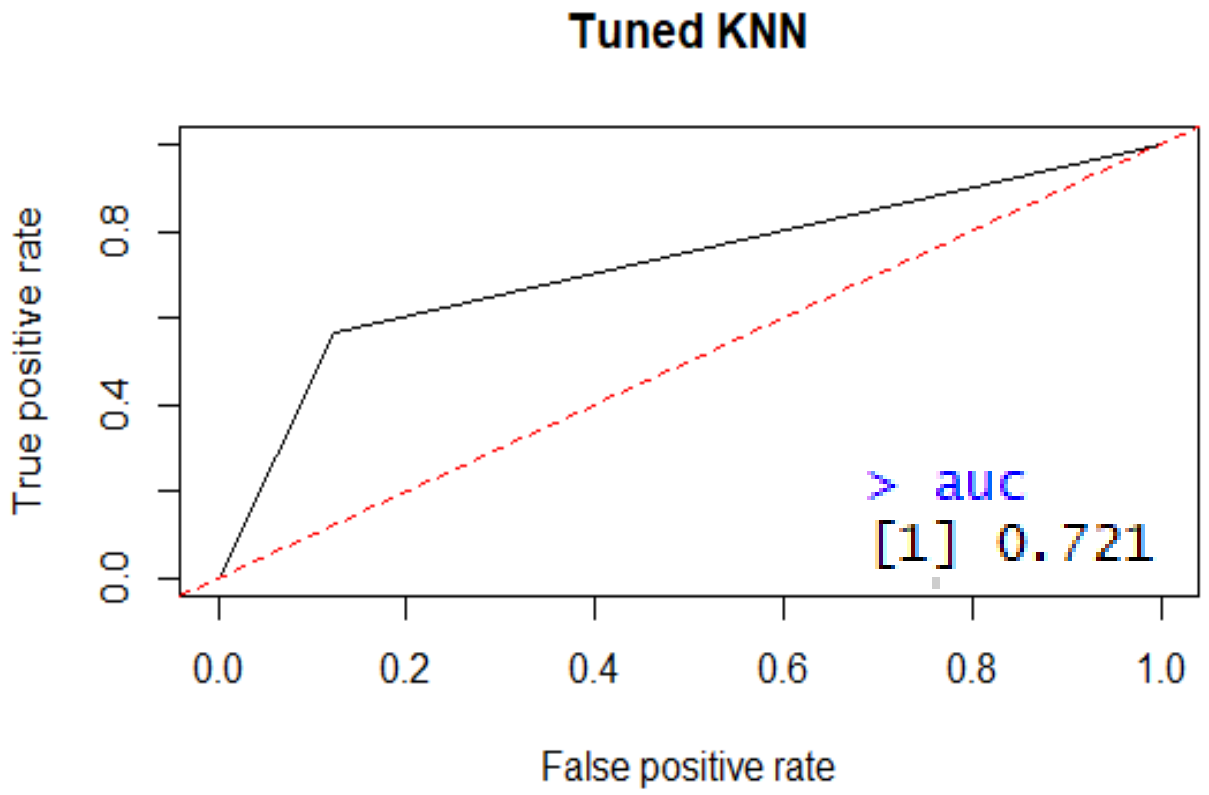
actual	predicted Neg	Pos	Row Total
Neg	496	4	500
	0.559	0.036	
	0.496	0.004	
Pos	392	108	500
	0.441	0.964	
	0.392	0.108	
Column Total	888	112	1000
	0.888	0.112	



Accuracy = (496+108)/1000 = 0.604

DTM 성능 평가 – KNN

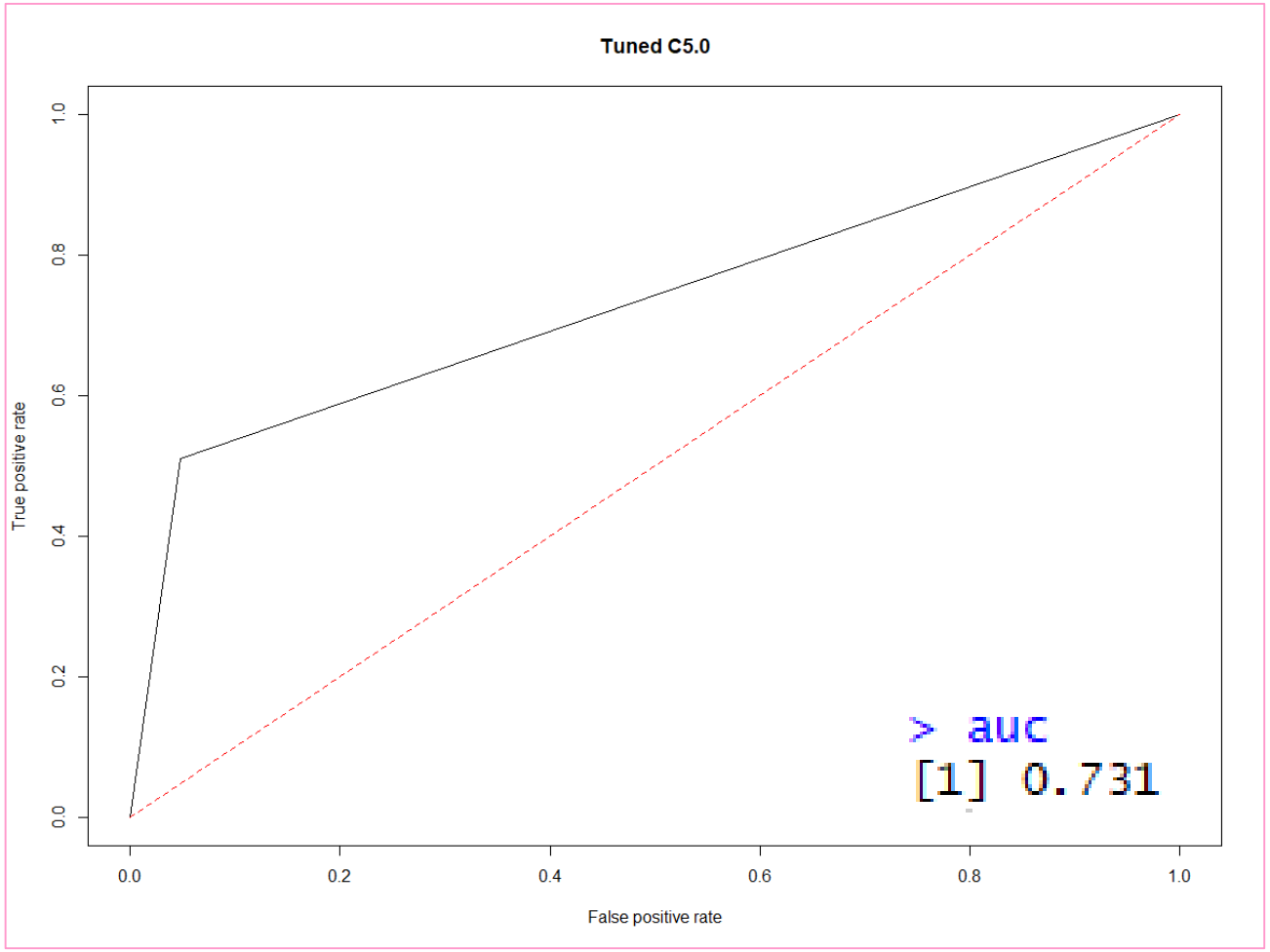
actual	predicted 0	1	Row Total
0	438	62	500
	0.669	0.180	
	0.438	0.062	
1	217	283	500
	0.331	0.820	
	0.217	0.283	
Column Total	655	345	1000
	0.655	0.345	



Accuracy = (438+283)/1000 = 0.721

DTM 성능 평가 – 의사결정나무

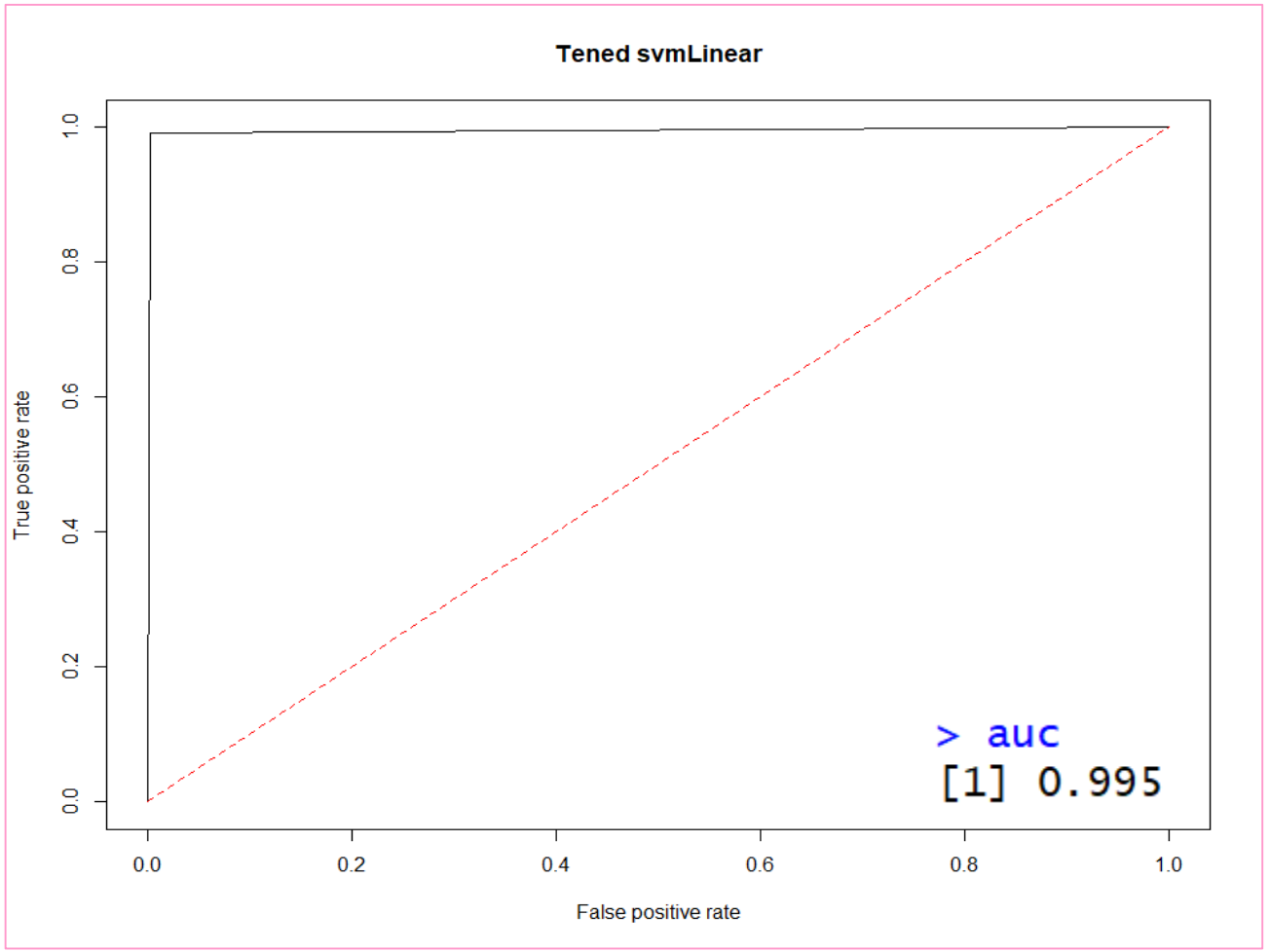
actual	predicted Neg	Pos	Row Total
Neg	476 0.660	24 0.086	500
Pos	245 0.340	255 0.914	500
Column Total	721 0.721	279 0.279	1000



Accuracy = (476+255)/1000 = 0.731

DTM 성능 평가 – SVM

	predicted		
actual	Neg	Pos	Row Total
Neg	499	1	500
	0.992	0.002	
Pos	4	496	500
	0.008	0.998	
Column Total	503	497	1000
	0.503	0.497	



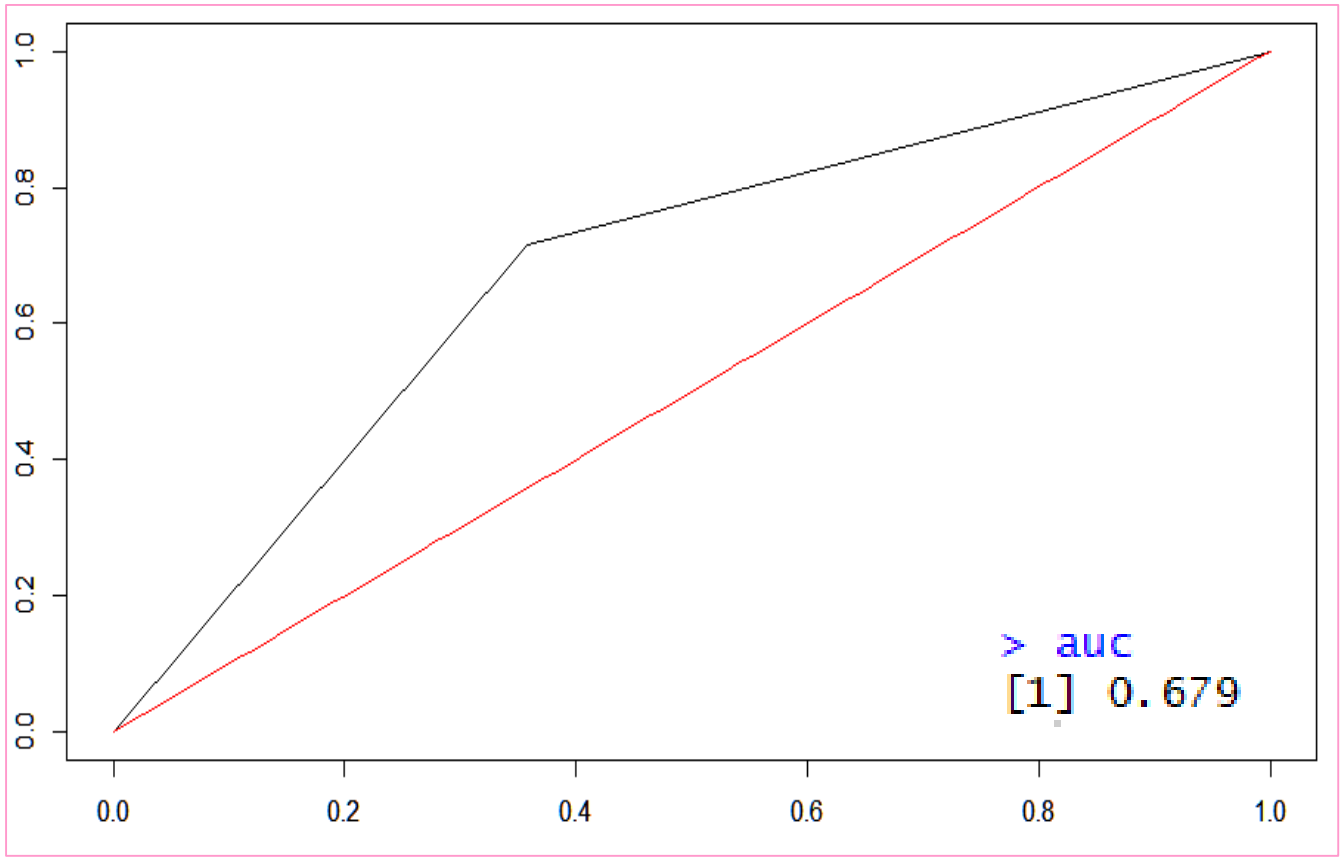
Accuracy = (499+496)/1000 = 0.995

DTM 성능 평가

	나이브 베이즈	KNN	의사결정나무	SVM
옵션	기본	K=5	기본	Linear
정확도	0.60	0.72	0.73	0.99
KAPPA	0.21	0.44	0.46	0.99
F-score	0.35	0.70	0.65	0.99
AUC	0.60	0.72	0.73	0.99

파생 변수 성능 평가 - KNN

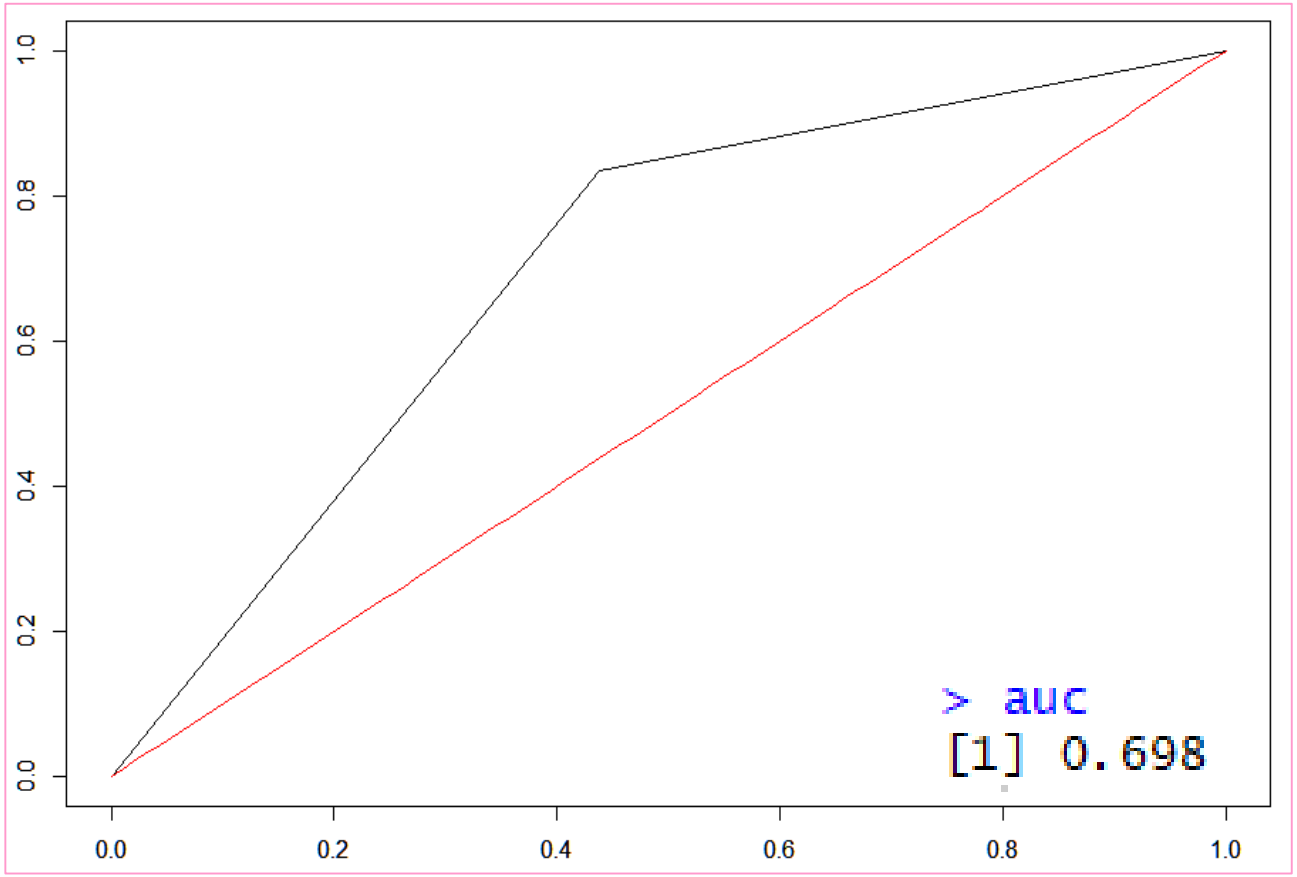
p	yelp[, 2] Neg	Pos	Row Total
Neg	329	133	462
	0.329	0.133	
Pos	171	367	538
	0.171	0.367	
Column Total	500	500	1000



Accuracy = (329+367)/1000 = 0.696

파생 변수 성능 평가 - 의사결정나무

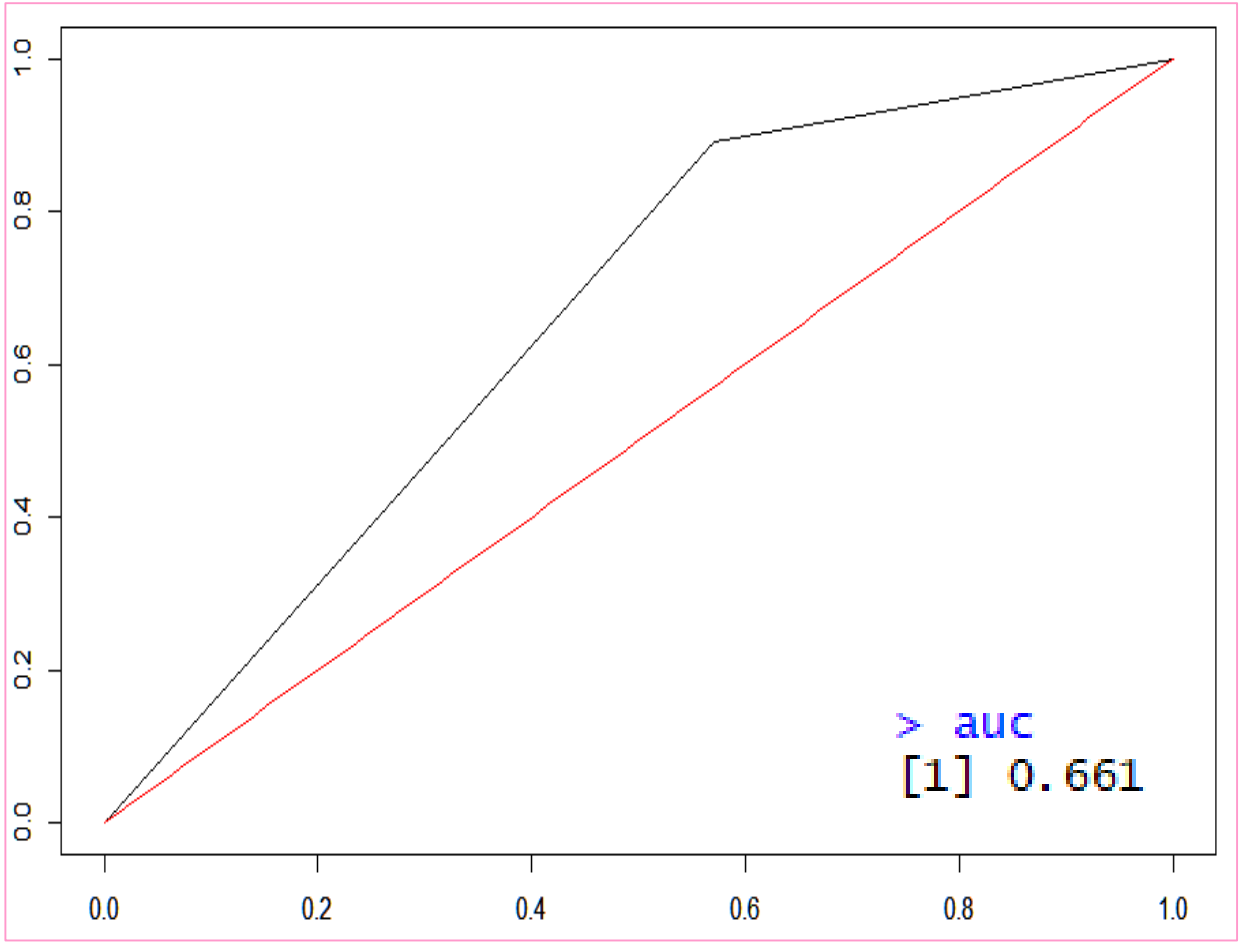
p	yelp[, 2] Neg	Pos	Row Total
Neg	321 0.321	142 0.142	463
Pos	179 0.179	358 0.358	537
Column Total	500	500	1000



Accuracy = (321+358)/1000 = 0.679

파생 변수 성능 평가 - SVM

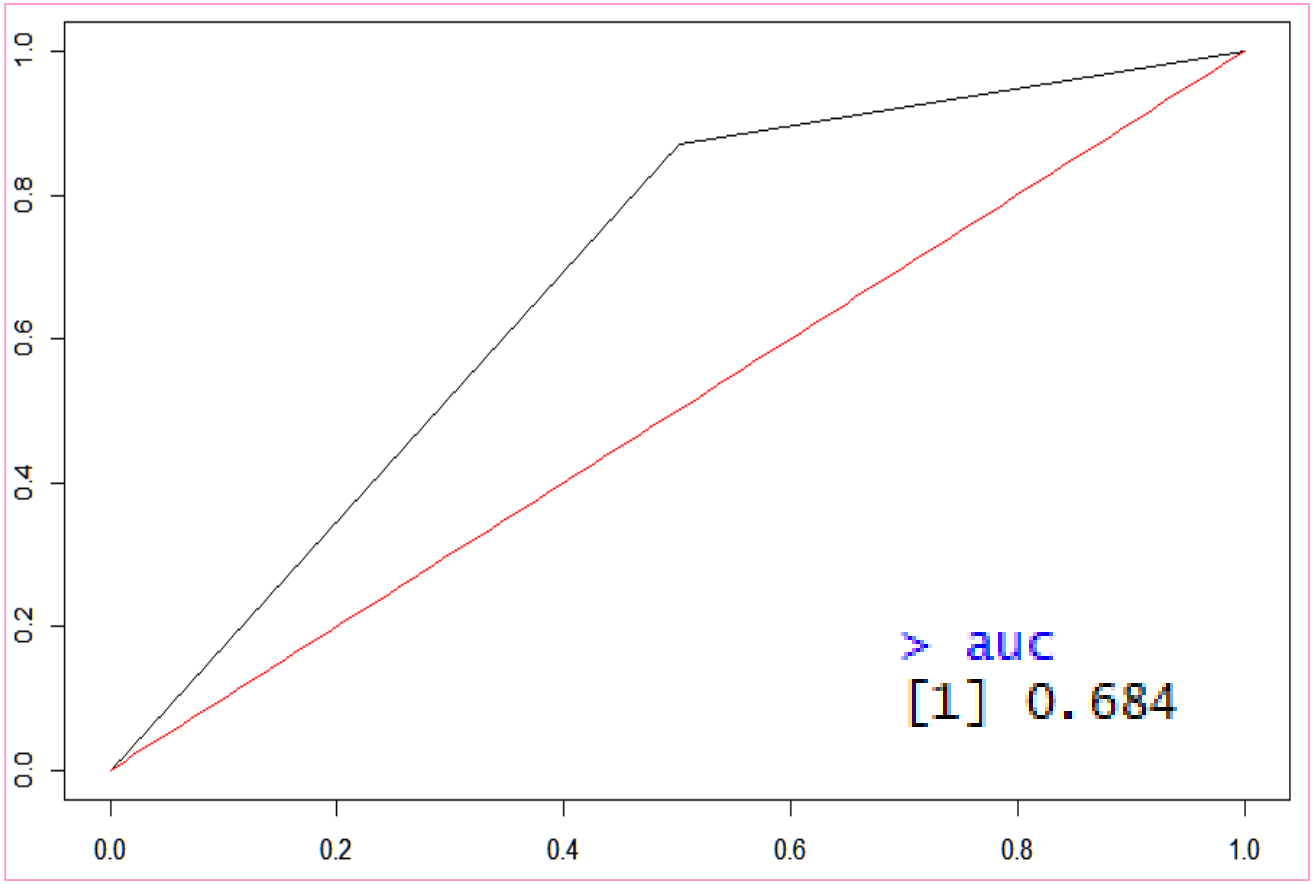
p	yelp[, 2] Neg	Pos	Row Total
Neg	281 0.281	83 0.083	364
Pos	219 0.219	417 0.417	636
Column Total	500	500	1000



Accuracy = (281+417)/1000 = 0.898

파생 변수 성능 평가 - 로지스틱 회귀분석

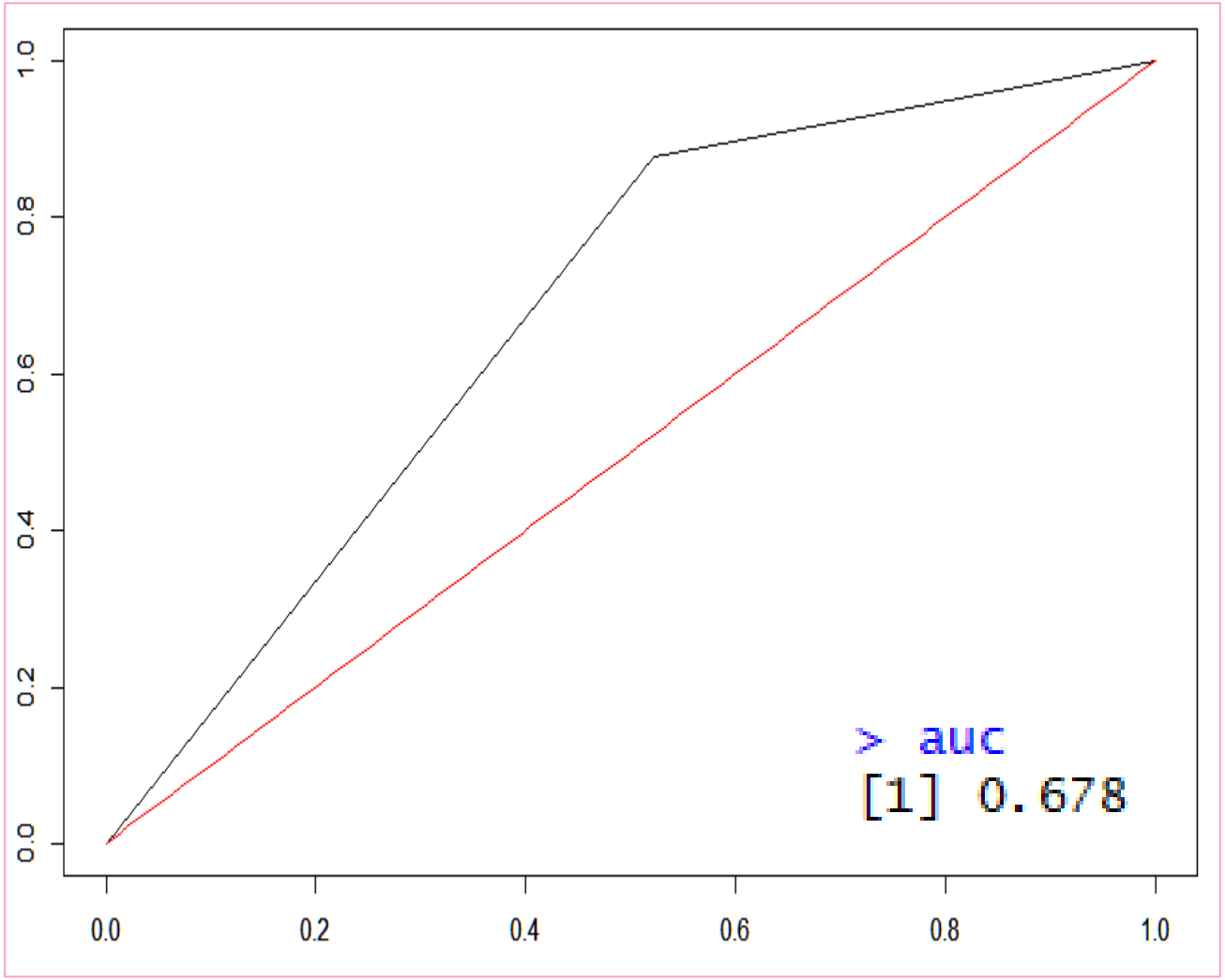
p	yelp[, 2] Neg	Pos	Row Total
Neg	249 0.249	65 0.065	314
Pos	251 0.251	435 0.435	686
Column Total	500	500	1000



Accuracy = (249+435)/1000 = 0.684

파생 변수 성능 평가 - 신경망

p	yelp[, 2] Neg	Pos	Row Total
Neg	239	61	300
	0.239	0.061	
Pos	261	439	700
	0.261	0.439	
Column Total	500	500	1000



Accuracy = (239+439)/1000 = 0.678

파생 변수 성능 평가

	KNN	의사결정나무	SVM	로지스틱	신경망
옵션	기본	K=15	Linear	기본	Node = 1
정확도	0.69	0.68	0.69	0.68	0.67
KAPPA	0.39	0.36	0.39	0.36	0.35
F-score	0.68	0.69	0.73	0.73	0.73
AUC	0.68	0.70	0.66	0.68	0.68

최종 모형 결정 – DTM + 파생변수 + SVM

```
Support Vector Machines with Linear Kernel

1000 samples
1557 predictors
   2 classes: 'Neg', 'Pos'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 900, 900, 900, 900, 900,
Resampling results:

  Accuracy  Kappa
  0.81      0.62

Tuning parameter 'C' was held constant at a value of 1
```

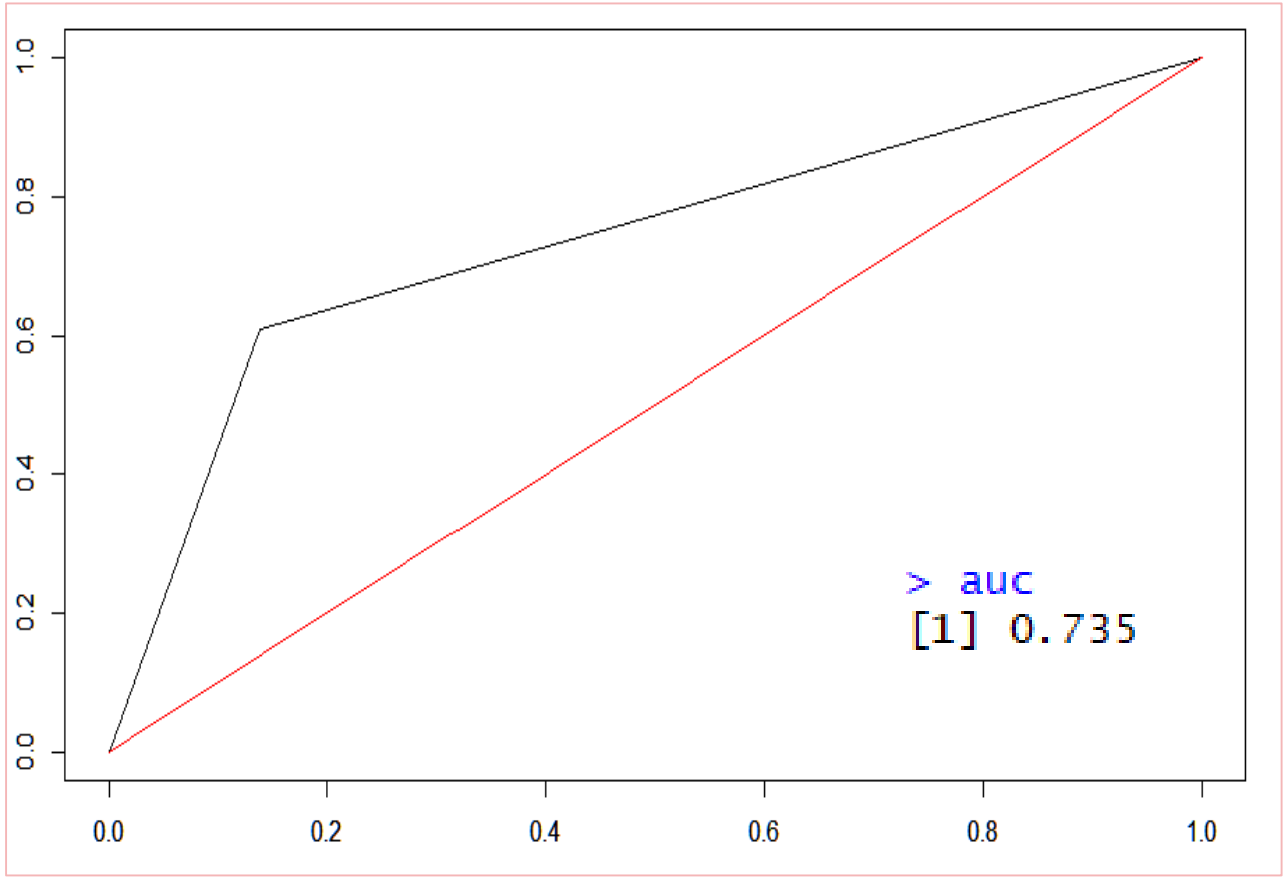
	Neg	Pos
a21	100.000	100.000
great	44.653	44.653
a9	41.011	41.011
a1	38.833	38.833
a4	34.710	34.710
good	32.884	3
a10	32.470	3
a3	22.835	2
a17	20.708	2
a7	19.784	1
love	18.838	1
a20	18.538	1
delici	16.745	1
friend	13.954	1
nice	13.256	1
amaz	13.256	1
wait	11.878	1
bad	11.861	1
never	11.175	1
minut	11.163	1
disappoint	11.163	11.163
back	11.110	11.110
perfect	10.466	10.466

```
Support Vector Machines with Linear Kernel

1000 samples
  24 predictor
   2 classes: 'Neg', 'Pos'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 900, 900, 900, 900, 900, 900, ...
Resampling results:

  Accuracy  Kappa
  0.733     0.466

Tuning parameter 'C' was held constant at a value of 1
```

부정표현, 작은따옴표, 느낌표, 마침표,
띄어쓰기, 문자열의 길이, 문자열-공백,
숫자, the

최종 모형 결정 – DTM + 파생변수 + SVM

p	yelp[, 2] Neg	Pos	Row Total
Neg	431 0.431	196 0.196	627
Pos	69 0.069	304 0.304	373
Column Total	500	500	1000



Accuracy	Kappa	F score	AUC
0.735	0.47	0.696	0.735

감사합니다