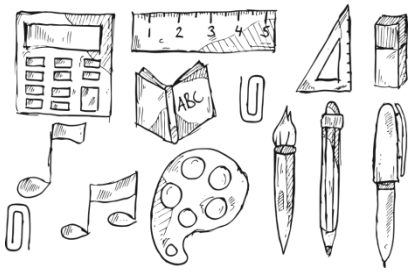
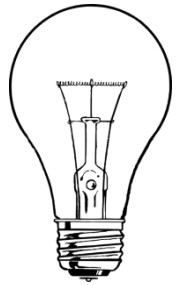
R – Project
도서 분야 예측 시스템

아이티윌 빅데이터 31기
김동환

INDEX

Enjoy your stylish business and campus life with BIZCAM





배경

Enjoy your stylish business and campus life with BIZCAM

목표

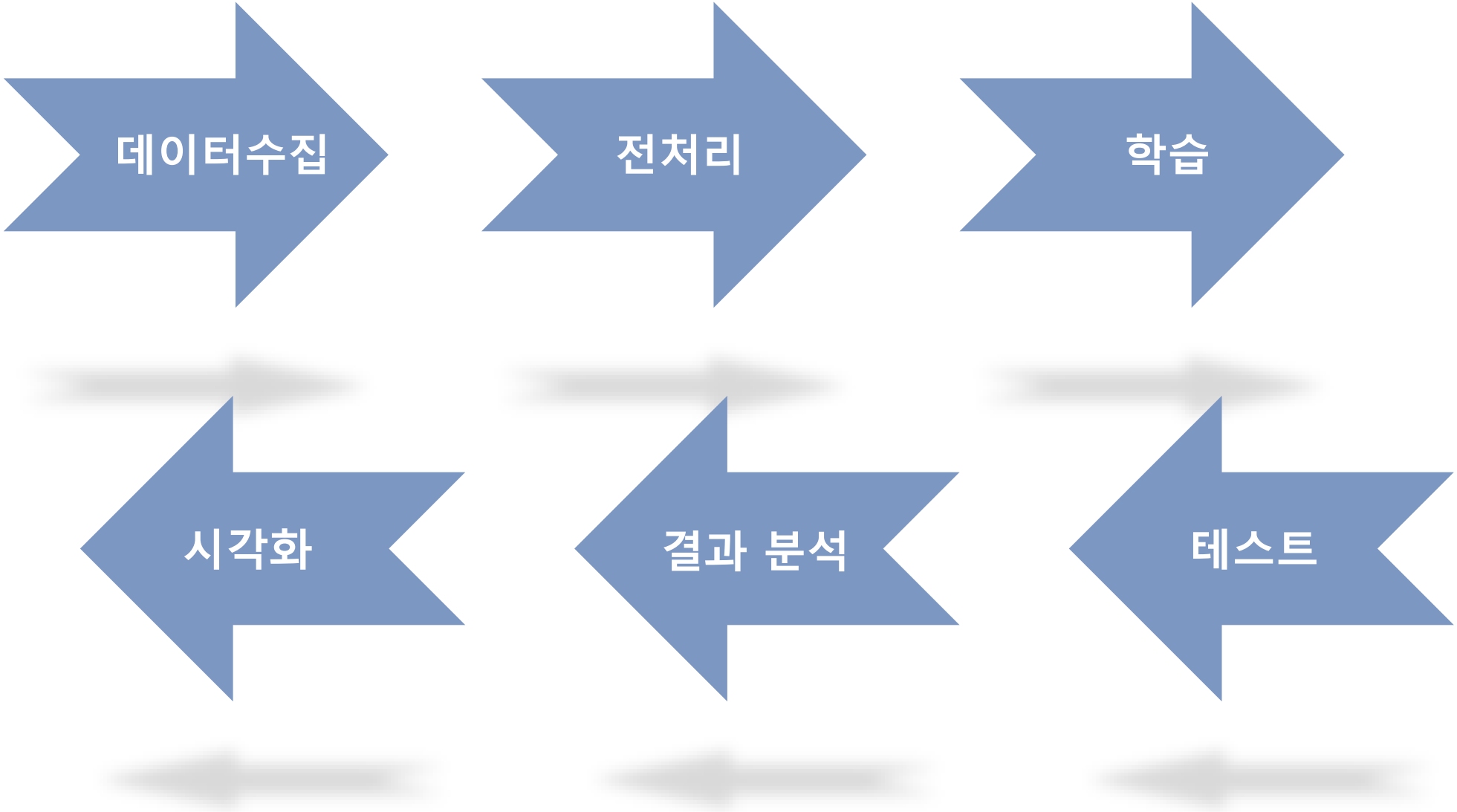
Enjoy your stylish business and campus life with BIZCAM

각 분야별 도서의 제목과 소개 내용을 추출 및 학습하고

분석하여 새로운 도서 텍스트에 대한 분야 예측

구현 순서

Enjoy your stylish business and campus life with BIZCAM



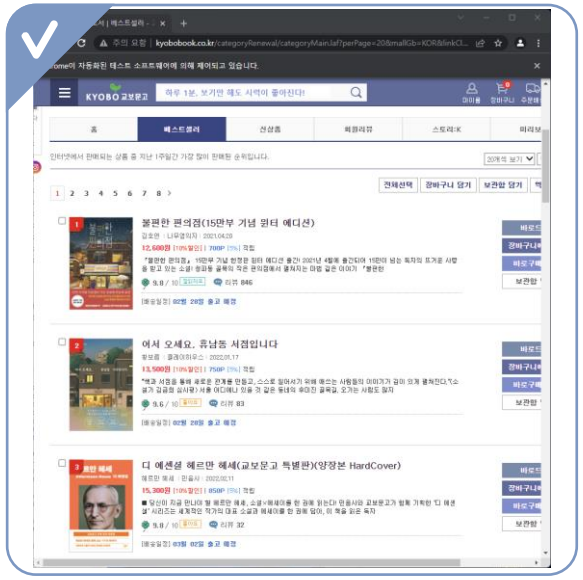
주요 라이브러리 / 함수

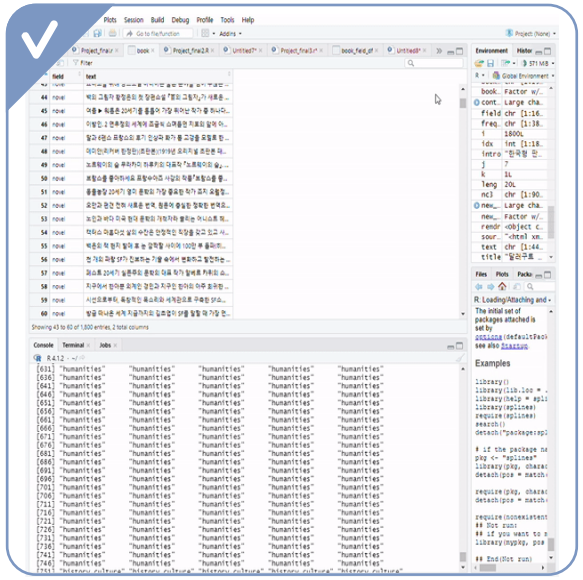
Enjoy your stylish business and campus life with BIZCAM

- rvest # 웹페이지에서 html태그를 추출하고 태그의 다양한 옵션들을 사용할 수 있는 라이브러리
  - ✓ Read_html, html_nodes, html_text
- RSelenium # 동적인 웹페이지의 태그를 추출할 수 있는 라이브러리
  - ✓ remoteDriver$(open, navigate, findElement, getPageSource, goBack)
- tm # VCorpus(말뭉치, documentTermMatrix 만드는 작업을 수행하는 라이브러리
  - ✓ VCorpus, DocumentTermMatrix, Terms, inspect, findFreqTerms
- stringr # 단어 추출,변경,제거, 공백제거 등 전처리작업에 주로 사용
  - ✓ str_replace_all, str_match_all,
- RcppMeCab # 일본어, 한국어 품사 태깅 라이브러리
  - ✓ pos
- e1071 # naiveBayes모델을 사용할 수 있는 라이브러리
  - ✓ naiveBayes
- gmodels # 두 개의 질적 자료 간의 관련성을 교차표로 나타내줄 수 있는 함수 CrossTable을 사용할 수 있는 라이브러리
  - ✓ CrossTable

데이터 수집

Enjoy your stylish business and campus life with BIZCAM





1

각 장르별 베스트셀러의 url 추출

2

각 책의 url에 들어가서 제목과 소개글을 추출

3

정답라벨, 추출한 텍스트 데이터프레임으로 저장

텍스트 전처리

Enjoy your stylish business and campus life with BIZCAM

```
#말뭉치로 변형작업
book_corpus <- tm::VCorpus(VectorSource(book$text))

#태깅,전처리 작업
#토큰으로 만들 함수
mecab_words <- function(doc){
  # 태깅
  tagging <- RcppMeCab::pos(base::enc2utf8(as.character(doc)))

  # 필요한 품사 NNG(일반명사), NNP(고유명사), XR(어근 : 실질적으로 의미를 L
  word_noun <- str_match_all(tagging,'(\\w+)(/NNG|/NNP|/XR|/SL)')[[1]][,2]

  # 전처리
  word_noun <- str_replace_all(word_noun,'C','씨언어')
  word_noun <- str_replace_all(word_noun,'R','알프로그래밍')
  word_noun <- str_replace_all(word_noun,'NFT','엔에프티')
  word_noun <- str_replace_all(word_noun,'ETF','이티에프')
  word_noun <- str_replace_all(word_noun,'LH','한국토지주택공사')
```

학습 / 테스트 데이터 분류

Enjoy your stylish business and campus life with BIZCAM

DocumentTermMatrix

```
book_dtm <- DocumentTermMatrix(book_corpus,control=list(tokenize=mecab_words,
                                                        wordLengths=c(2,Inf),
                                                        weighting=weightBin))
```

```
<<DocumentTermMatrix (documents: 1800, terms: 14330)>>
Non-/sparse entries: 98369/25695631
Sparsity           : 100%
Maximal term length: 15
Weighting          : binary (bin)
Sample             :
      Terms
Docs   내용 독자 문제 사람 세계 시작 이야기 이해 자신 저자
  1103    1    0    1    0    0    0      0    1    0    0
  1124    0    0    1    1    1    1      1    1    1    1
  1159    0    0    1    1    1    0      0    0    1    1
  1182    0    0    1    1    0    1      1    0    1    0
  1417    1    1    1    1    1    0      1    1    0    1
  3       0    1    0    0    1    1      1    0    1    0
  36      0    1    1    1    1    1      1    0    1    0
  5       0    1    0    0    1    0      1    0    1    0
  504     0    0    1    1    1    1      1    0    1    1
  867     0    0    1    1    1    0      1    0    0    1
```

학습 / 테스트 데이터 분류할 숫자 복원 추출로 랜덤 입력 sample

```
idx <- sample(2,nrow(book_dtm),replace=TRUE,prob=c(0.9,0.1))
```

```
idx
  [1] 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2
 [51] 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 1
[101] 1 1 1 1 2 1 1 2 1 1 1 1 2 1 1 2 2 1 1 1 1 1 2 1 2 1 1 1 1 2
[151] 1 1 1 1 1 1 1 2 2 1 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2
[201] 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1
[251] 2 1 1 1 1 1 1 1 1 1 2 1 1 1 2 1 2 1 1 1 1 1 1 1 2 1 1 1 1 1
[301] 1 1 1 1 1 1 1 2 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1
[351] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 2 1 1 1 2 1 1 1
[401] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1
[451] 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1
[501] 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[551] 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[601] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[651] 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[701] 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[751] 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1
[801] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 2 1 2 1
[851] 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1
[901] 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1
[951] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 2 1 1
[ reached getOption("max.print") -- omitted 800 entries ]
```

학습 / 테스트 데이터 분류

Enjoy your stylish business and campus life with BIZCAM

학습데이터

```
<<DocumentTermMatrix (documents: 1624, terms: 14330)>>
Non-/sparse entries: 89223/23182697
Sparsity           : 100%
Maximal term length: 15
Weighting          : binary (bin)
Sample             :
      Terms
Docs   내용 독자 문제 사람 세계 시작 이야기 이해 자신 저자
  1103    1    0    1    0    0    0       0    1    0    0
  1124    0    0    1    1    1    1       1    1    1    1
  1159    0    0    1    1    1    0       0    0    1    1
  1182    0    0    1    1    0    1       1    0    1    0
  1417    1    1    1    1    1    0       1    1    0    1
  3       0    1    0    0    1    1       1    0    1    0
  36      0    1    1    1    1    1       1    0    1    0
  5       0    1    0    0    1    0       1    0    1    0
  504     0    0    1    1    1    1       1    0    1    1
  867     0    0    1    1    1    0       1    0    0    1
```

학습데이터 개수

```
> nrow(book_dtm_train)
[1] 1624
```

학습데이터 정답라벨 개수

```
> length(book_dtm_train_labels)
[1] 1624
```

테스트데이터

```
<<DocumentTermMatrix (documents: 176, terms: 14330)>>
Non-/sparse entries: 9146/2512934
Sparsity           : 100%
Maximal term length: 15
Weighting          : binary (bin)
Sample             :
      Terms
Docs   내용 문제 사람 세계 시작 이야기 이해 자신 저자 필요
  1095    1    1    0    0    0       0    0    0    0    1
  1136    0    1    0    1    0       0    1    0    0    0
  1199    0    1    0    1    1       0    0    0    0    0
  1292    1    1    1    0    0       0    1    0    1    0
  1602    0    0    0    0    0       0    1    0    0    0
  1662    1    1    0    0    0       0    1    0    1    1
  235     1    0    0    1    1       1    0    0    0    0
  335     0    0    1    0    1       0    0    1    1    1
  346     0    0    0    1    0       1    1    0    0    0
  68      0    1    0    1    0       1    0    1    0    0
```

테스트데이터 개수

```
> nrow(book_dtm_test)
[1] 176
```

테스트데이터 정답라벨 개수

```
> length(book_dtm_test_labels)
[1] 176
```

학습 / 테스트 데이터 분류

Enjoy your stylish business and campus life with BIZCAM

apply

```
convert_counts <- function(x){
  x <- ifelse(x>0,'YES','NO')
}
```

```
book_train <- apply(book_dtm_train,MARGIN=2,convert_counts)
book_test <- apply(book_dtm_test,MARGIN=2,convert_counts)
```

```
가가   가가미   가감   가게   가격   가계
"NO"   "NO"   "NO"   "NO"   "NO"   "NO"
가급   가까이   가나안   가난   가네코   가늠
"NO"   "NO"   "NO"   "NO"   "NO"   "NO"
가독   가독성   가드   가드너   가든   가디언
"NO"   "NO"   "NO"   "NO"   "NO"   "NO"
로등   가로세로   가로수   가루   가르   가르시아
"NO"   "NO"   "NO"   "NO"   "NO"   "NO"
가명   가문   가뭄   가미   가미야   가미오카
"NO"   "NO"   "NO"   "NO"   "NO"   "NO"
ㄹ라   가비지   가사   가산점   가상   가상현실
"NO"   "NO"   "NO"   "NO"   "NO"   "NO"
가속   가속도   가수   가스   가스관   가스실
"NO"   "NO"   "NO"   "NO"   "NO"   "NO"
가시   가식   가신   가액   가야   가에
"NO"   "NO"   "NO"   "NO"   "NO"   "NO"
가운   가운데   가을   가의   가이   가이드 가이
"NO"   "NO"   "NO"   "NO"   "NO"   "NO"
```

naiveBayes, predict

Enjoy your stylish business and campus life with BIZCAM

- naiveBayes

```
book_classifier <- naiveBayes(book_train,book_dtm_train_labels)
```

흑역사

book_dtm_train_labels	NO	YES
art_popularculture	1.00000000	0.00000000
computer_it	1.00000000	0.00000000
economy	1.00000000	0.00000000
history_culture	0.98591549	0.01408451
humanities	1.00000000	0.00000000
novel	1.00000000	0.00000000
poem_essay	1.00000000	0.00000000
politics_society	1.00000000	0.00000000
religion	1.00000000	0.00000000
selfdevelope	1.00000000	0.00000000
sience	1.00000000	0.00000000
technology_engineering	1.00000000	0.00000000

힌두교

book_dtm_train_labels	NO	YES
art_popularculture	1.00000000	0.00000000
computer_it	1.00000000	0.00000000
economy	1.00000000	0.00000000
history_culture	0.98496241	0.01503759
humanities	1.00000000	0.00000000
novel	1.00000000	0.00000000
poem_essay	1.00000000	0.00000000
politics_society	1.00000000	0.00000000
religion	0.99270073	0.00729927
selfdevelope	1.00000000	0.00000000
sience	1.00000000	0.00000000
technology_engineering	1.00000000	0.00000000

힌트

book_dtm_train_labels	NO	YES
art_popularculture	1.000000000	0.000000000
computer_it	1.000000000	0.000000000
economy	0.992537313	0.007462687
history_culture	1.000000000	0.000000000
humanities	1.000000000	0.000000000
novel	1.000000000	0.000000000
poem_essay	0.992366412	0.007633588
politics_society	1.000000000	0.000000000
religion	1.000000000	0.000000000
selfdevelope	0.978571429	0.021428571
sience	1.000000000	0.000000000
technology_engineering	1.000000000	0.000000000

- predict

```
book_test_predict <- predict(book_classifier,book_test)
```

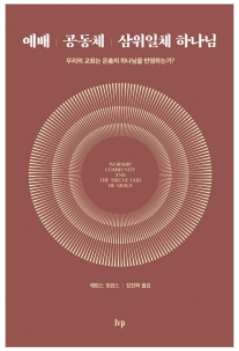
```
> book_test_predict
  [1] poem_essay           novel                novel                politics_society
  [5] novel                poem_essay           novel                novel
  [9] novel                history_culture      novel                novel
 [13] novel                poem_essay           poem_essay           poem_essay
 [17] art_popularculture   poem_essay           poem_essay           poem_essay
 [21] poem_essay           poem_essay           poem_essay           history_culture
 [25] poem_essay           humanities           poem_essay           poem_essay
 [29] selfdevelope         economy              economy              economy
 [33] economy              economy              economy              sience
 [37] politics_society     economy              selfdevelope         economy
 [41] economy              politics_society     economy              economy
 [45] economy              economy              economy              humanities
 [49] poem_essay           religion             selfdevelope         selfdevelope
 [53] selfdevelope         humanities           selfdevelope         humanities
 [57] selfdevelope         poem_essay           selfdevelope         selfdevelope
 [61] religion             politics_society     selfdevelope         politics_society
 [65] humanities           humanities           selfdevelope         humanities
```

book_test_predict	art_popularculture	computer_it	economy	history_culture	humanities	novel	poem_essay	politics_society	religion	selfdevelope	sience	technology_engineering	Row Total
art_popularculture	16	0	0	0	0	0	0	0	0	1	0	0	17
	102.449	2.028	1.545	1.642	0.386	1.352	1.835	1.352	1.256	0.001	0.869	1.835	
	0.941	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.059	0.000	0.000	0.097
	0.800	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.100	0.000	0.000	
	0.091	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.000	0.000	
computer_it	0	17	0	0	0	0	0	0	0	0	0	1	19
	2.159	95.746	1.727	1.835	0.432	1.511	2.051	1.511	1.403	1.080	0.001	0.539	
	0.000	0.895	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.053	0.053	0.108
	0.000	0.810	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.111	0.053	
	0.000	0.097	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.006	
economy	0	1	13	0	0	0	0	0	0	0	0	0	14
	1.591	0.269	108.058	1.352	0.318	1.114	1.511	1.114	1.034	0.795	0.716	1.511	
	0.000	0.071	0.929	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.080
	0.000	0.048	0.812	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
	0.000	0.006	0.074	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
history_culture	0	0	0	11	0	1	0	1	0	0	0	0	13
	1.477	1.551	1.182	75.618	0.295	0.001	1.403	0.001	0.960	0.739	0.665	1.403	
	0.000	0.000	0.000	0.846	0.000	0.077	0.000	0.077	0.000	0.000	0.000	0.000	0.074
	0.000	0.000	0.000	0.647	0.000	0.071	0.000	0.071	0.000	0.000	0.000	0.000	
	0.000	0.000	0.000	0.062	0.000	0.006	0.000	0.006	0.000	0.000	0.000	0.000	
humanities	1	0	0	2	3	0	2	0	1	1	0	0	10
	0.016	1.193	0.909	1.107	33.827	0.795	0.785	0.795	0.092	0.328	0.511	1.080	
	0.100	0.000	0.000	0.200	0.300	0.000	0.200	0.000	0.100	0.100	0.000	0.000	0.057
	0.050	0.000	0.000	0.118	0.750	0.000	0.105	0.000	0.077	0.100	0.000	0.000	
	0.006	0.000	0.000	0.011	0.017	0.000	0.011	0.000	0.006	0.006	0.000	0.000	
novel	0	0	0	0	0	12	0	0	0	0	0	0	12
	1.364	1.432	1.091	1.159	0.273	127.812	1.295	0.955	0.886	0.682	0.614	1.295	
	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.068
	0.000	0.000	0.000	0.000	0.000	0.857	0.000	0.000	0.000	0.000	0.000	0.000	
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
poem_essay	2	0	1	0	0	0	7	0	0	0	0	0	18
	0.001	2.148	1.63								1.943		
	0.111	0.000	0.00								0.000	0.102	
	0.100	0.000	0.00								0.000		
	0.011	0.000	0.000	0.000	0.000	0.000	0.080	0.006	0.000	0.006	0.000	0.000	
politics_society	0	0	1	2	0	0	1	7	0	0	0	1	12
	1.364	1.432	0.008	0.610	0.273	0.955	0.067	38.288	0.886	0.682	0.614	0.067	
	0.000	0.000	0.083	0.167	0.000	0.000	0.083	0.583	0.000	0.000	0.000	0.083	0.068
	0.000	0.000	0.062	0.118	0.000	0.000	0.053	0.500	0.000	0.000	0.000	0.053	
	0.000	0.000	0.006	0.011	0.000	0.000	0.006	0.040	0.000	0.000	0.000	0.006	
religion	0	0	0	1	0	1	1	0	12	1	0	0	16
	1.818	1.909	1.455	0.193	0.364	0.058	0.306	1.273	99.028	0.009	0.818	1.727	
	0.000	0.000	0.000	0.062	0.000	0.062	0.062	0.000	0.750	0.062	0.000	0.000	0.091
	0.000	0.000	0.000	0.059	0.000	0.071	0.053	0.000	0.923	0.100	0.000	0.000	
	0.000	0.000	0.000	0.006	0.000	0.006	0.006	0.000	0.068	0.006	0.000	0.000	
selfdevelope	0	0	0	0	1	0	1	0	0	6	0	0	8
	0.909	0.955	0.727	0.773	3.682	0.636	0.022	0.636	0.591	67.655	0.409	0.864	
	0.000	0.000	0.000	0.000	0.125	0.000	0.125	0.000	0.000	0.750	0.000	0.000	0.045
	0.000	0.000	0.000	0.000	0.250	0.000	0.053	0.000	0.000	0.600	0.000	0.000	
	0.000	0.000	0.000	0.000	0.006	0.000	0.006	0.000	0.000	0.034	0.000	0.000	
sience	1	0	1	0	0	0	0	1	0	0	7	2	12
	0.097	1.432	0.008	1.159	0.273	0.955	1.295	0.002	0.886	0.682	66.465	0.383	
	0.083	0.000	0.083	0.000	0.000	0.000	0.000	0.083	0.000	0.000	0.583	0.167	0.068
	0.050	0.000	0.062	0.000	0.000	0.000	0.000	0.071	0.000	0.000	0.778	0.105	
	0.006	0.000	0.006	0.000	0.000	0.000	0.000	0.006	0.000	0.000	0.040	0.011	
technology_engineering	0	3	1	1	0	0	0	4	0	0	1	15	25
	2.841	0.000	0.713	0.829	0.568	1.989	2.699	2.034	1.847	1.420	0.061	56.067	
	0.000	0.120	0.040	0.040	0.000	0.000	0.000	0.160	0.000	0.000	0.040	0.600	0.142
	0.000	0.143	0.062	0.059	0.000	0.000	0.000	0.286	0.000	0.000	0.111	0.789	
	0.000	0.017	0.006	0.006	0.000	0.000	0.000	0.023	0.000	0.000	0.006	0.085	
Column Total	20	21	16	17	4	14	19	14	13	10	9	19	176
	0.114	0.119	0.091	0.097	0.023	0.080	0.108	0.080	0.074	0.057	0.051	0.108	

예측률 : (16+17+13+11+3+12+14+7+12+6+7+15)/176 = 0.7556818%

추가 테스트 데이터

Enjoy your stylish business and campus life with BIZCAM



무료배송 | 소득공제

[출간예정] 예배, 공동체, 삼위일체 하나님 가?

제임스 토런스 지음 | 김진혁 옮김 | IVP | 2022년 03월 04일 출간

정가 : 12,000원

판매가 : **10,800원** [10%↓ 1,200원 할인]

혜택 : [기본적립] 600원 적립 [5% 적립]

[추가적립] 5만원 이상 구매 시 2,000원 추가적립 안내

[회원혜택] 회원 등급 별, 3만원 이상 구매 시 2~4% 추가적립 안내

책소개

이 책이 속한 분야

종교 > 기독교(개신교) > 기독교일반 > 기독교일반

예배, 성례, 교제의 참된 모습은 무엇인가?
삼위일체 하나님이 교회에 불어넣으신 참생명이 드러나다!

"예배 현장에서 교리가 나왔다. 그러나 바른 교리는 바른 예배의 초석이 된다.
이 책은 우리로 하여금 바른 예배에 눈뜨게 할 것이다!"
문화랑 고려신학대학원 예배학 교수

제임스 토런스가 삼위일체 교리를 토대로 예배 신학에 관해 강의한 1994년 디즈버리 강연에 기초하여 출간한 책이다. 예배에 대한 새로운 이해가 필요한 시대, 우리의 예전은 무엇에 기초하는가. 제임스 토런스는 전통 교리와 히브리서에 대한 깊고 탄탄한 이해를 바탕으로, 예배가 은총의 삼위일체 하나님의 존재와 성자 그리스도 예수님이 행하신 바를 반영해야 함을 밝힌다. 그리고 은총의 선물로 주어진 예배에 우리가 성령으로 참여하게 된다는 복된 소식을 전한다. 이 책은 그 분량의 간결함에도 그리스도의 대제사장직 교리, 세례와 성찬에 대한 이해와 실행, 현대 페미니즘에 대한 반응까지 신학과 성례전 해석을 부족함 없이 다룬다. 이제 교회는 예수 그리스도 안에 있는 참된 중심으로 돌아가라는 이 책의 외침에 귀 기울여야 한다.

http://www.kyobobook.co.kr/product/detailViewKor.laf?mallGb=KOR&ejkGb=KOR&linkClass=21030301&barcode=9788932819143

corpus

↓

Terms

↓

DocumentTermMatrix

↓

apply 이분화

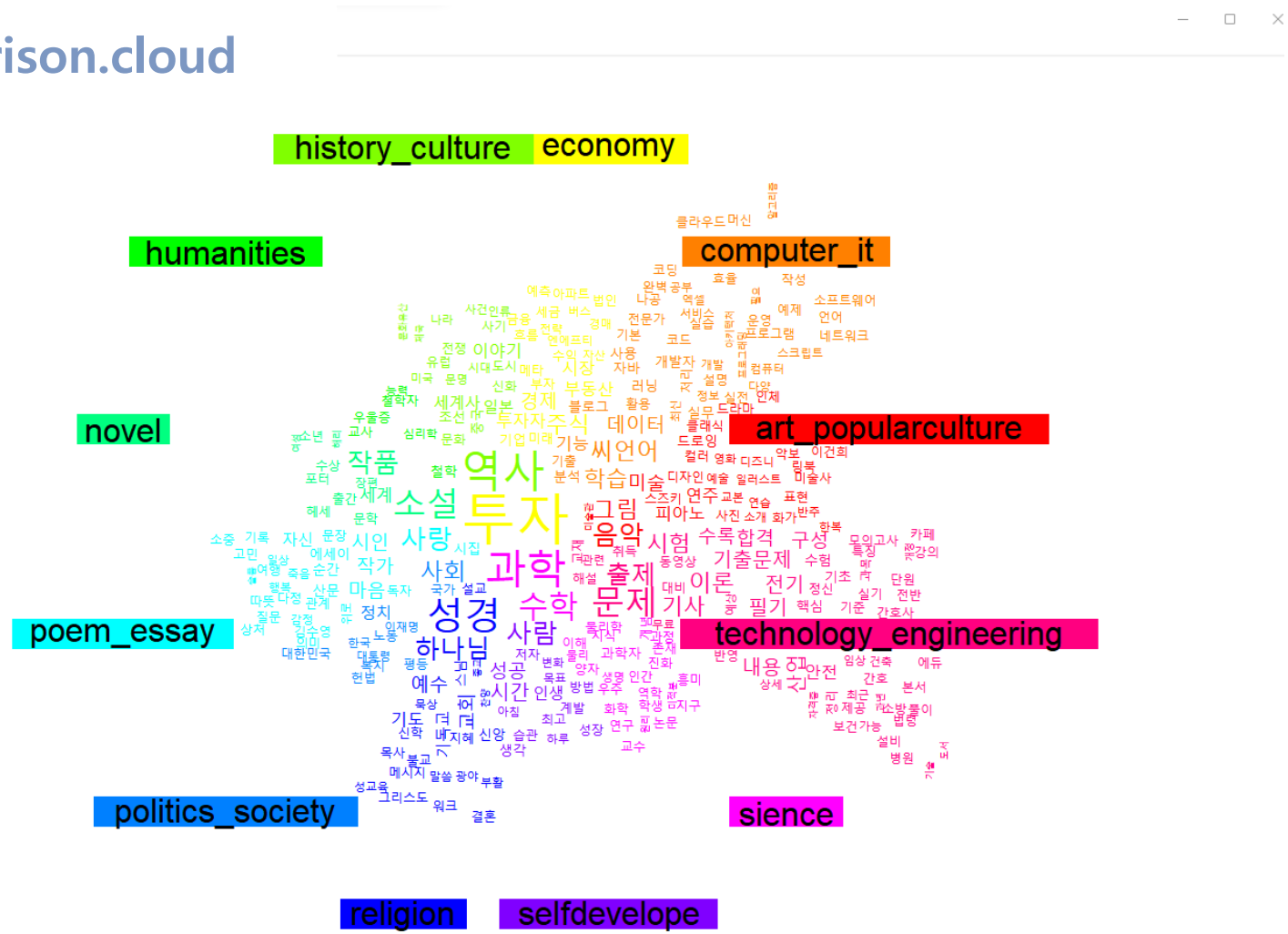
↓

```
> predict(book_classifier,t(new_dtm_test))
[1] religion
Levels: art_popularculture computer_it eco
```

분야별 단어 빈도 시각화

Enjoy your stylish business and campus life with BIZCAM

wordcloud::comparison.cloud



방향성

Enjoy your stylish business and campus life with BIZCAM

1. 기존의 낮은 예측률 문제 해결
2. 사람들의 상황, 상태에 맞춤형 책을 추천해주는 시스템 구현
   - 평점, 후기, 소개, 목차, 제목을 분석 및 학습하여 text를 입력했을 때 text와 관련된 책 추천

감사합니다