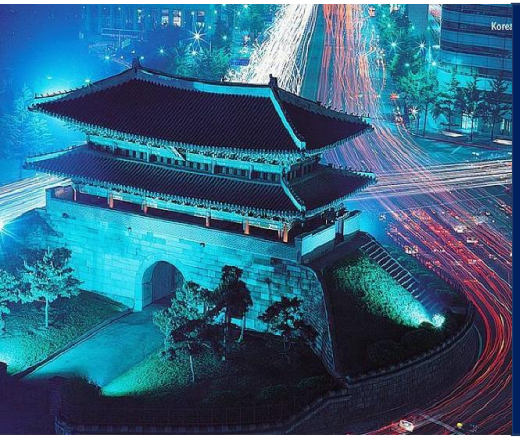
고려대학교
KOREA UNIVERSITY



Validity Extraction and Evaluation on Tourist Attraction with the review data

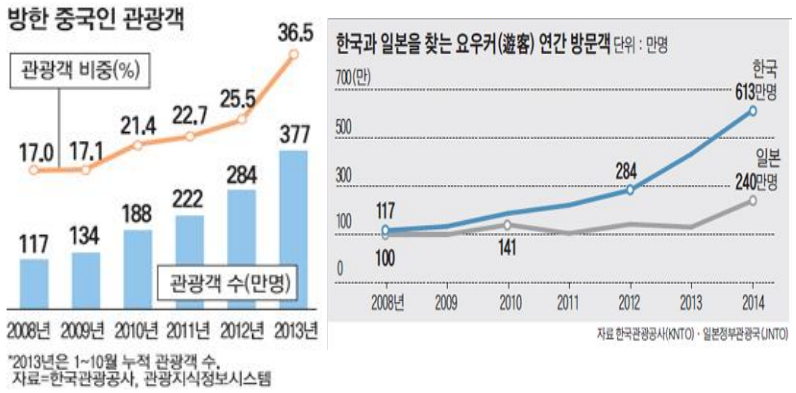
3조 Final

김보섭, 박민식, 조수현

발표자: 박민식

연구배경

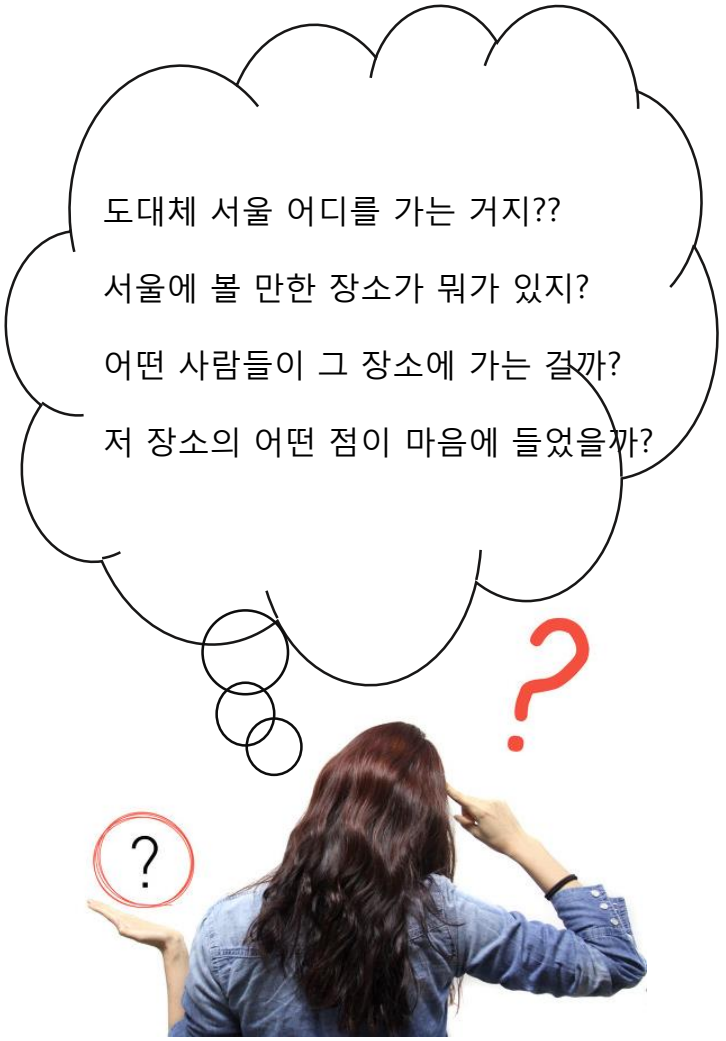
외국인 관광객 추이



매년 증가하는 외국인 관광객의 수

한국 여행 시 주로 서울을 방문

관광객을 보는 한국인의 시선



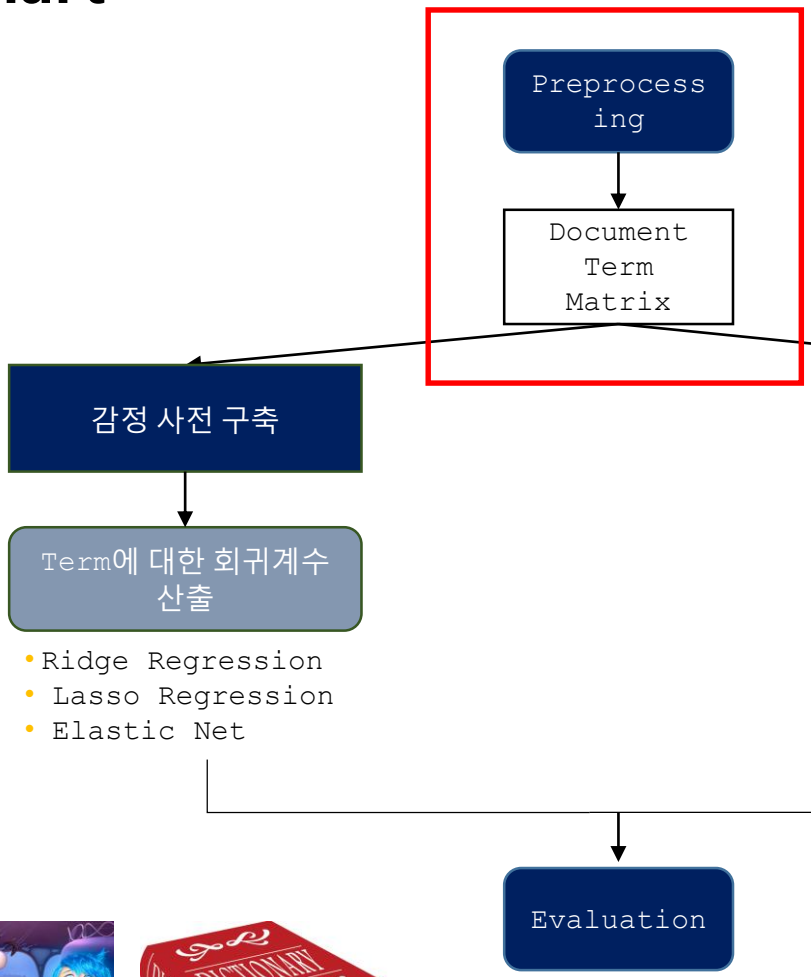
목적

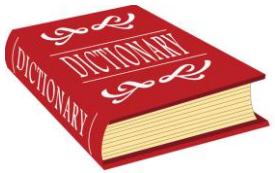
- 여행지를 평가하는 평가 요인은 명확하지 않음. (사람마다 기준이 다를 수 있음)
- 여행지를 평가하는 평가 요인을 사람들의 의견(Comment)을 통해 얻어내고
- 도출된 각각의 평가 요인을 감정 분석을 통해 평가하는 방법론 제안

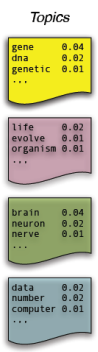
ID	Comment	Rate
Emily K	I recommend ever..	5
FOL-1003	For us Europeans it is a too much too extensive….	4
Andrew J	This is probably one of the best museums…..	5
⋮	⋮	⋮

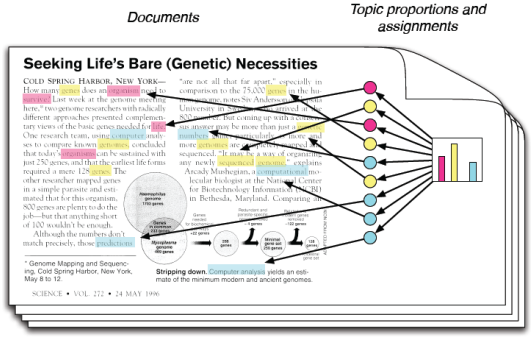
ID	Price	…	Kindness	Rate
Emily K				5
FOL-1003				4
Andrew J				5
⋮				⋮

Flow chart

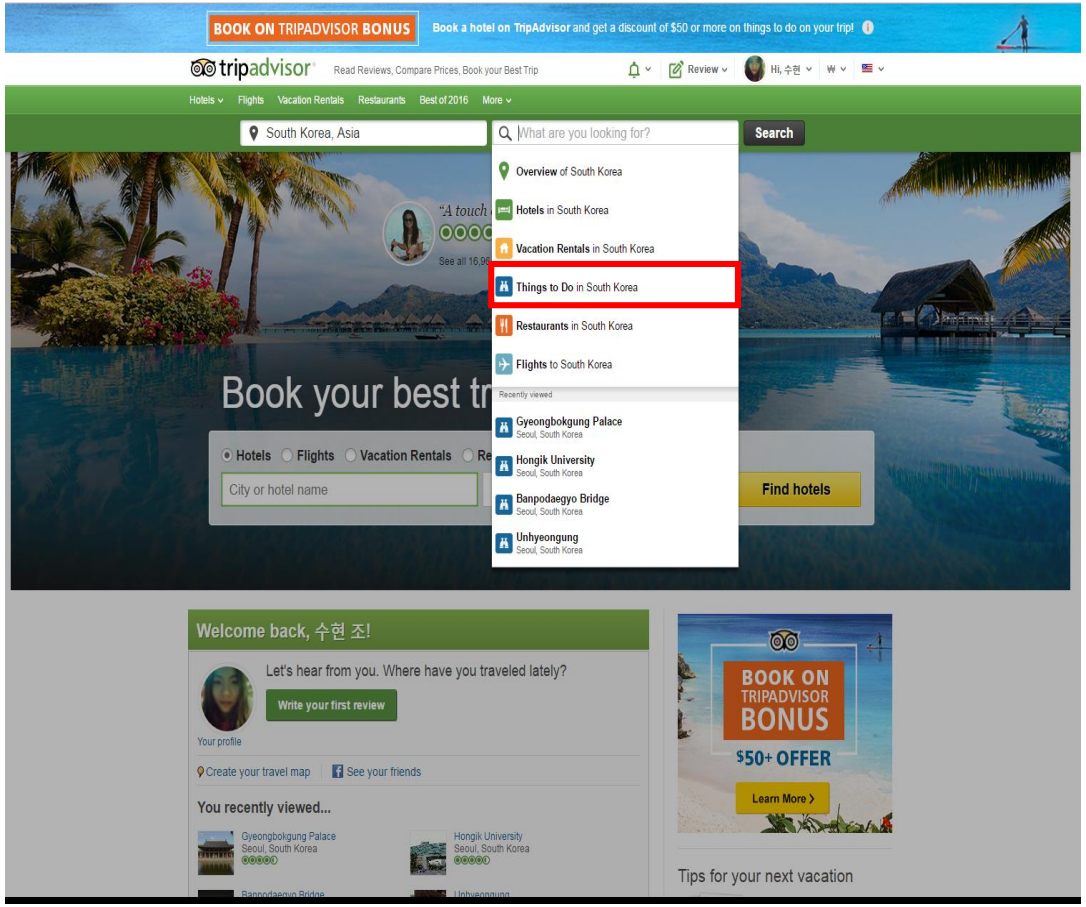

Preprocess ing
Document Term Matrix
감정 사전 구축
평가 요인 산출
Term에 대한 회귀계수 산출
Topic Modeling
• Ridge Regression
• Lasso Regression
• Elastic Net
Evaluation


DICTIONARY
DICTIONARY


Topics
gene 0.04
dna 0.02
genetic 0.01
life 0.02
evolve 0.01
organism 0.01
brain 0.04
neuron 0.02
nerve 0.01
data 0.02
number 0.02
computer 0.01


Documents
Topic proportions and assignments
Seeking Life's Bare (Genetic) Necessities

데이터 소개 및 수집

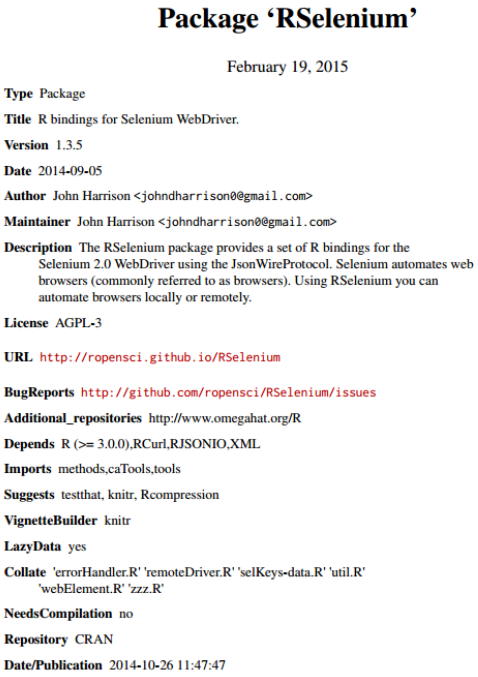


- 103개의 장소에 대한 data
- 영어로 작성된 Review만 수집
- 21620 observations
- 7개의 variables
  - 장소
  - id
  - location
  - 방문 시기
  - 제목
  - Review
  - 평점

데이터 소개 및 수집

❖ 데이터 수집

- 웹사이트를 테스트 할 때 자동으로 수행할 목적으로 만든 Open source
- JavaScript로 만들어진 사이트는 인터넷 창에서 화면을 클릭해줘야 데이터를 크롤링 가능
- tripadvisor.com는 JavaScript로 되어있어서 크롤링을 하기 위해서 Selenium이 필요
- R에서는 'RSelenium'이라는 패키지를 이용, JavaScript로 만든 사이트의 데이터 크롤링 가능



Package 'RSelenium'

February 19, 2015

Type Package

Title R bindings for Selenium WebDriver.

Version 1.3.5

Date 2014-09-05

Author John Harrison <johndharrison0@gmail.com>

Maintainer John Harrison <johndharrison0@gmail.com>

Description The RSelenium package provides a set of R bindings for the Selenium 2.0 WebDriver using the JsonWireProtocol. Selenium automates web browsers (commonly referred to as browsers). Using RSelenium you can automate browsers locally or remotely.

License AGPL-3

URL http://ropensci.github.io/RSelenium

BugReports http://github.com/ropensci/RSelenium/issues

Additional_repositories http://www.omegahat.org/R

Depends R (>= 3.0.0),RCurl,RJSONIO,XML

Imports methods,caTools,tools

Suggests testthat, knitr, Rcompression

VignetteBuilder knitr

LazyData yes

Collate 'errorHandler.R' 'remoteDriver.R' 'selKeys-data.R' 'util.R' 'webElement.R' 'zzz.R'

NeedsCompilation no

Repository CRAN

Date/Publication 2014-10-26 11:47:47

데이터 소개 및 수집

❖ 데이터 수집

- R selenium을 이용한 데이터 수집

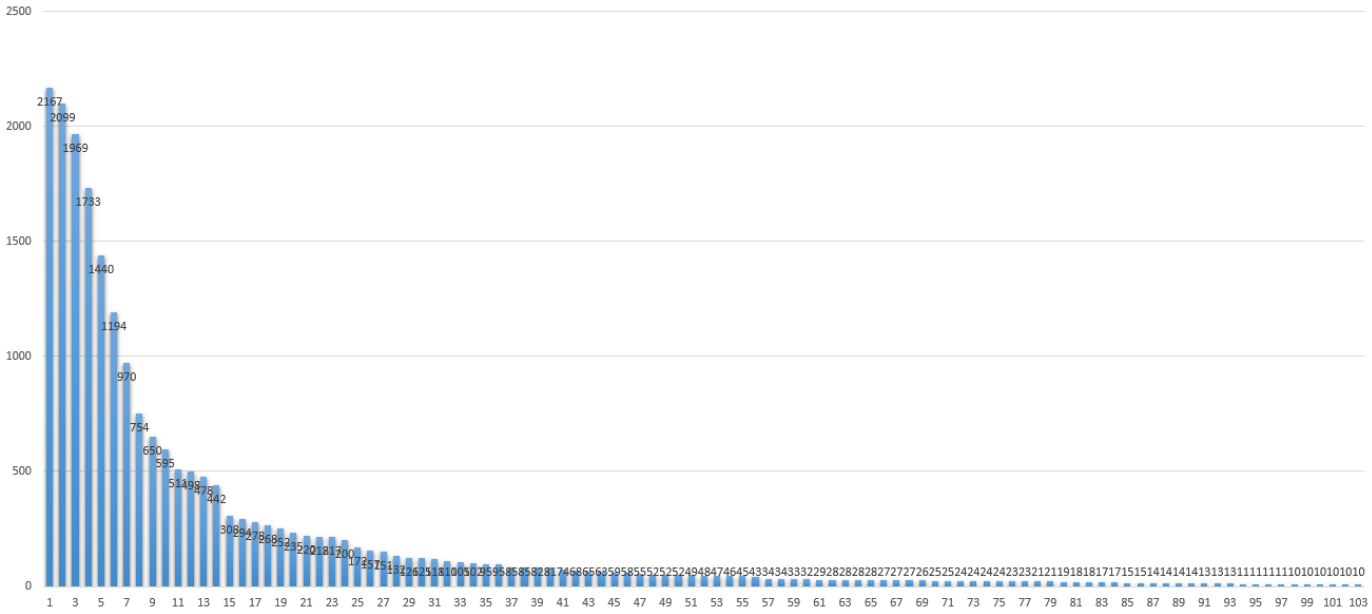
site	id	location	date	title	entry	rate
Bukhansan_National_Park	RSReynoldsTravel	Victoria, Canada	15149	National Park on Seoul's Doorstep	I am shocked that there are only two reviews of the ...	5
Korea_House	Columbus9	Columbus, Ohio	15163	What a treasure!	I got to visit the Korea House twice on my recent trip...	5
Namsan_Park	dsong11	Memphis, Tennessee	15170	Escape the City	Escape the hustle and bustle of 25 million people wit...	5
Changdeokgung_Palace	IgaMotylska	Johannesburg, South Africa	15173	The Secret Garden within Changdeokgung	I have been to Changdeokgung and the Secret Gard...	4
Insadong	Jennikimi	Cincinnati, Ohio	15175	The Never Ending Streets of Delight	Insa-dong is one of the popular districts in South Kor...	5
The_Blue_House_Cheong_Wa_Dae	JerryThirteen	Kuala Lumpur, Malaysia	15179	A good insight into The Blue House	A rare opportunity to see The President's Office com...	4
Bukhansan_National_Park	xamhayana	South Pole	15188	Fantastic nature retreat within easy reach from Seoul	Bukhansan is a deeply enchanting forest park, and v...	5
N_Seoul_Tower	Diane_Nguyen	Denver, Colorado	15194	Lover's Locks	At the base of the tower, lovers write their names on ...	3
Noryangjin_Fish_Market	mlddrss	Vienna, Austria	15196	Must-do in Seoul	Besides buying fish you can choose fish, oysters (or...	5
Insadong	NoelBella	Malaysia	15198	Interesting.	For any tourist who visits seoul, this is the place to g...	4
Changdeokgung_Palace	Hormuz	Mumbai, India	15203	The Amazing Royal Palaces of Seoul, South Korea.	Seoul was made the capital city of Korea over six ce...	5
Jeongdong_Theater	DutchCarioca	Singapore, Singapore	15206	Great show	In the Chongdong Theater the show running in Augu...	4
The_National_Folk_Museum_of_Korea	The_3_Black_Cats	Camberley, UK	15209	A charming insight into Korean Culture	The National Folk Museum of Korea is situated in the ...	4
Gyeongbokgung_Palace	chequered_8787	Kuala Lumpur, Malaysia	15222	Over-rated	I'd like to start off by saying that I'm a history buff - es...	2
Namsan_Park	Brendan O	Seoul, South Korea	15234	One of the best things about Seoul	I have lived in Seoul for several years now, and I hav...	5
Hangang_Park	poopshooter	Petaling Jaya, Malaysia	15239	nice recreational area...	The Han River spans 500 plus kms and is 1km wide w...	3
Seodaemun_Prison_History_Hall	poopshooter	Petaling Jaya, Malaysia	15239	worthwhile stop...	This is a great place to visit for a piece of Korean ind...	4
Changdeokgung_Palace	travelr0	Vancouver, WA	15240	Fascinating place	This palace, now in Seoul, was originally built in the ...	5
Trickeye_Museum	Azza80	Christchurch NZ	15253	Fun, quirky Art gallery	The Trick Eye Museum is a gallery of optical illusion ...	4
Hangang_Park	ms_scotsgirl	Washington DC, District of Columbia	15254	My favourite day in South Korea was cycling along t...	Rent bikes, cross to the north bank of Mapo Bridge, c...	5
Olympic_Park	GManSeoul	Seattle, Washington	15255	Nice get away for a few hours	The park is located just outside the main areas of Se...	4

EDA

❖ 장소

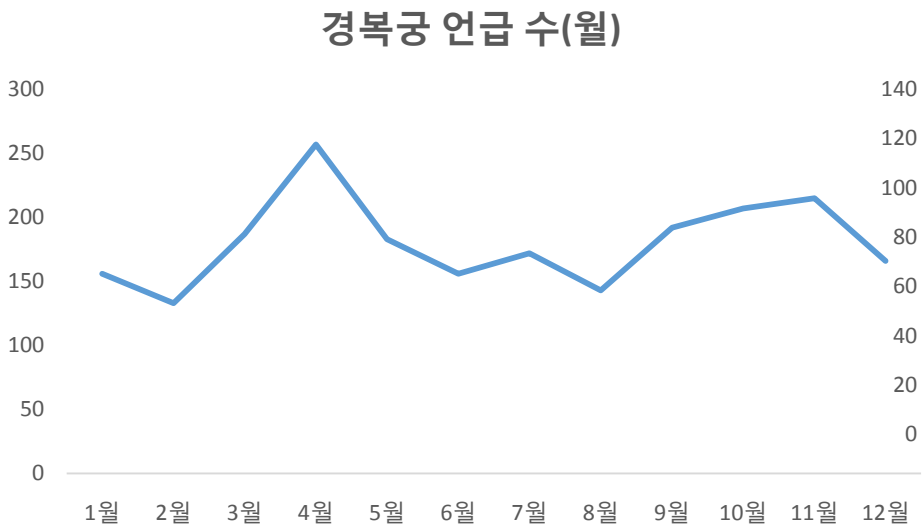
- 총 103개의 장소, 21,620개의 리뷰 데이터
- 빈도수 순서대로 아래 표와 그림으로 나타낼 수 있음

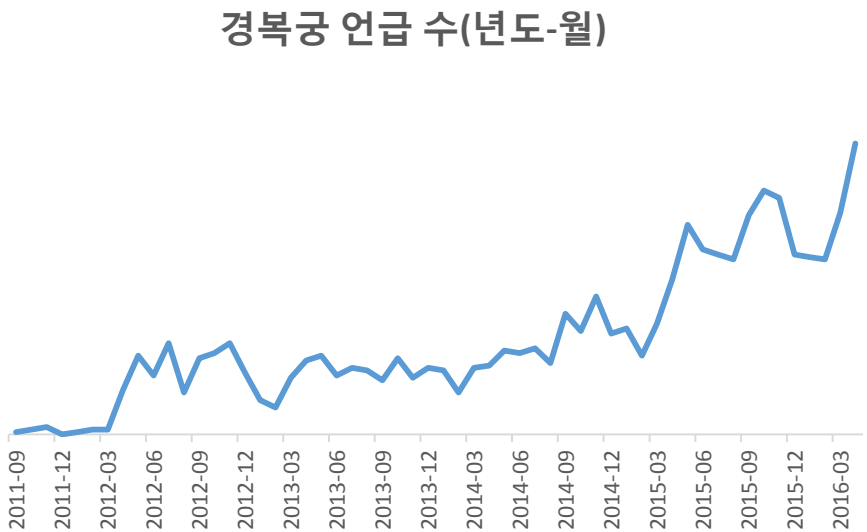
장소	빈도
경복궁	2167
남산타워	2099
명동 쇼핑거리	1969
인사동	1733
전쟁기념관	1440
창덕궁	1194
북촌	970
…	…
…	…
롯데월드 민속 박물관	10
현충원	10



장소 방문 추세

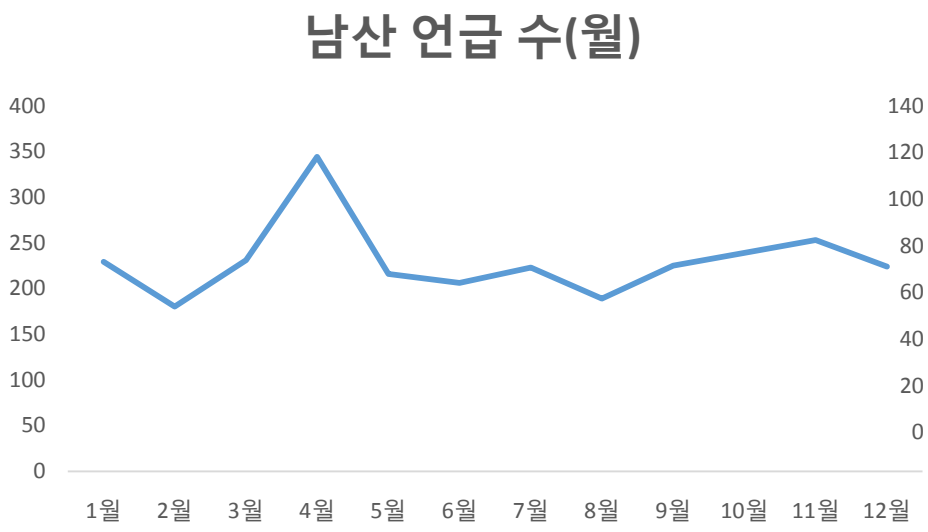
❖ 경복궁

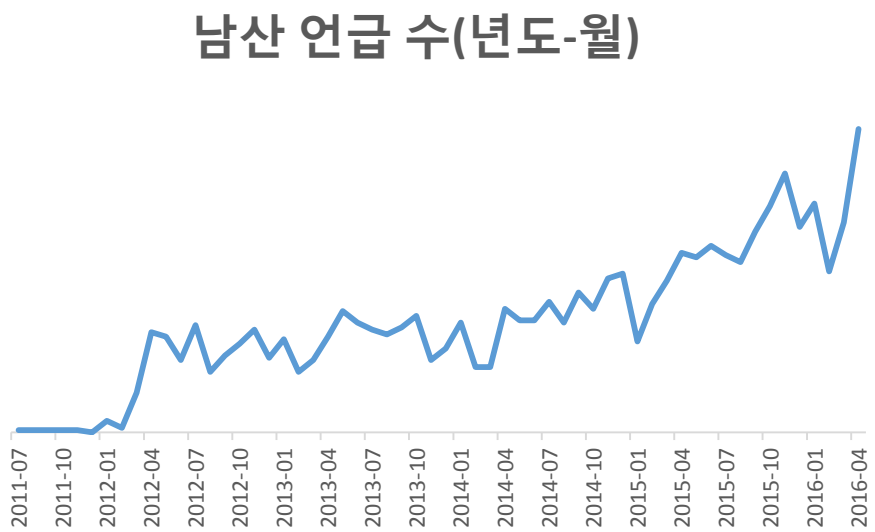

경복궁 언급 수(월)
300
250
200
150
100
50
0
1월
2월
3월
4월
5월
6월
7월
8월
9월
10월
11월
12월


경복궁 언급 수(년도-월)
140
120
100
80
60
40
20
0
2011-09
2011-12
2012-03
2012-06
2012-09
2012-12
2013-03
2013-06
2013-09
2013-12
2014-03
2014-06
2014-09
2014-12
2015-03
2015-06
2015-09
2015-12
2016-03

장소 방문 추세

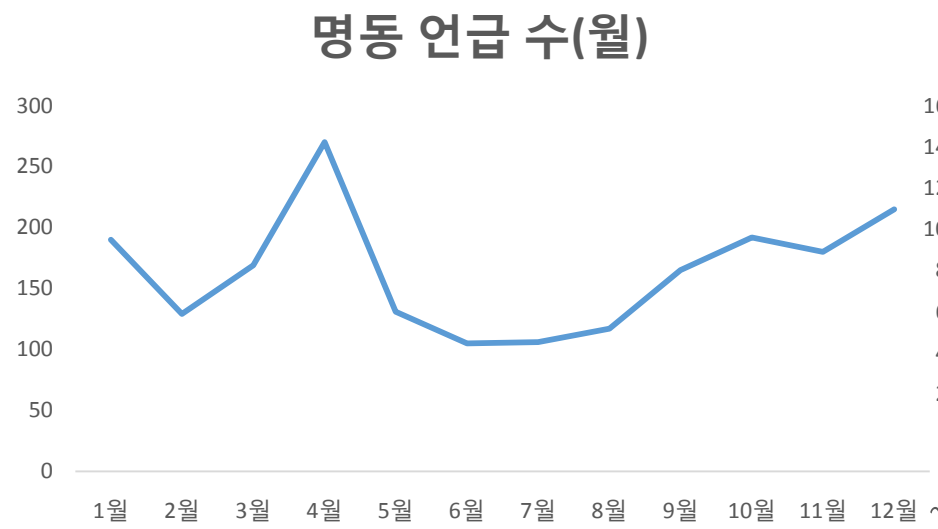
❖ 남산(남산타워+남산공원+남산 케이블카)

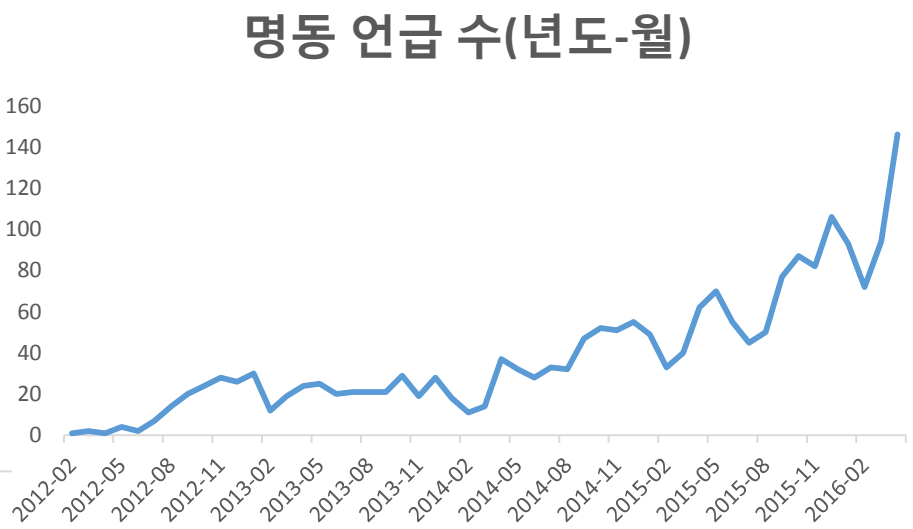




장소 방문 추세

❖ 명동


명동 언급 수(월)
300
250
200
150
100
50
0
1월
2월
3월
4월
5월
6월
7월
8월
9월
10월
11월
12월


명동 언급 수(년도-월)
160
140
120
100
80
60
40
20
0
2012-02
2012-05
2012-08
2012-11
2013-02
2013-05
2013-08
2013-11
2014-02
2014-05
2014-08
2014-11
2015-02
2015-05
2015-08
2015-11
2016-02

POS Tagging

Package 'qdap'

October 9, 2015

Type Package

Title Bridging the Gap Between Qualitative Data and Quantitative Analysis

Version 2.2.4

Date 2015-10-08

Maintainer Tyler Rinker <tyler.rinker@gmail.com>

Depends R (>= 3.1.0), qdapDictionaries (>= 1.0.2), qdapRegex (>= 0.1.2), qdapTools (>= 1.3.1), RColorBrewer

Imports chron, dplyr (>= 0.3), gdata, gender (>= 0.5.1), ggplot2 (>= 0.9.3.1), grid, gridExtra, igraph, methods, NLP, openNLP (>= 0.2-1), parallel, plotrix, RCurl, reports, reshape2, scales, stringdist, tidyr, tm (>= 0.6.2), tools, venneuler, wordcloud, xlsx, XML

Suggests koRpus, knitr, lda, proxy, stringi, SnowballC, testthat

LazyData TRUE

VignetteBuilder knitr

Description Automates many of the tasks associated with quantitative discourse analysis of transcripts containing discourse including frequency counts of sentence types, words, sentences, turns of talk, syllables and other assorted analysis tasks. The package provides parsing tools for preparing transcript data. Many functions enable the user to aggregate data by any number of grouping variables, providing analysis and seamless integration with other R packages that undertake higher level analysis and visualization of text. This affords the user a more efficient and targeted analysis. 'qdap' is designed for transcript analysis, however, many functions are applicable to other areas of Text Mining/Natural Language Processing.

License GPL-2

URL http://trinker.github.com/qdap/

BugReports http://github.com/trinker/qdap/issues

1	CC	Coordinating conjunction
2	CD	Cardinal number
3	DT	Determiner
4	EX	Existential there
5	FW	Foreign word
6	IN	Preposition or subordinating conjunction
7	JJ	Adjective
8	JJR	Adjective, comparative
9	JJS	Adjective, superlative
10	LS	List item marker
11	MD	Modal
12	NN	Noun, singular or mass
13	NNS	Noun, plural
14	NNP	Proper noun, singular
15	NNPS	Proper noun, plural
16	PDT	Predeterminer
17	POS	Possessive ending
18	PRP	Personal pronoun
19	PRP$	Possessive pronoun
20	RB	Adverb
21	RBR	Adverb, comparative
22	RBS	Adverb, superlative
23	RP	Particle
24	SYM	Symbol
25	TO	to
26	UH	Interjection
27	VB	Verb, base form
28	VBD	Verb, past tense
29	VBG	Verb, gerund or present participle
30	VBN	Verb, past participle
31	VBP	Verb, non-3rd person singular present
32	VBZ	Verb, 3rd person singular present
33	WDT	Wh-determiner
34	WP	Wh-pronoun
35	WP$	Possessive wh-pronoun
36	WRB	Wh-adverb

Qdap package로 POS Tagging실행

POS Tagging

```
[1] "My family enjoyed going here. It is huge and all free! Lots of military memorabilia to see.
 Loved seeing the B52 up close!"
```

POS Tagging

```
[1] "huge/JJ museum/NN my/PRP$ family/NN enjoyed/VBD going/VBG here/RB it/PRP is/VBZ huge/JJ and
/CC all/DT free/JJ lots/NNS of/IN military/JJ memorabilia/NNS to/TO see/VB loved/VBN seeing/VBG
the/DT b/NN up/RP close/RB"
```

명사, 형용사, 부사, 동사 추출

```
 [1] "huge/JJ"        "museum/NN"       "family/NN"     "enjoyed/VBD"    "going/VBG"
 [6] "here/RB"        "is/VBZ"          "huge/JJ"       "free/JJ"        "lots/NNS"
[11] "military/JJ"    "memorabilia/NNS" "see/VB"        "loved/VBN"      "seeing/VBG"
[16] "b/NN"           "close/RB"
```

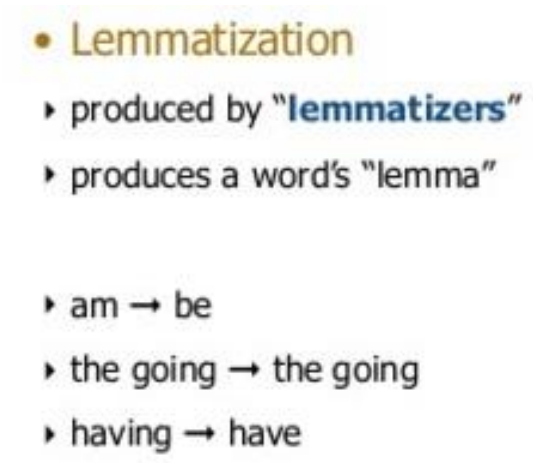
감정 분석을 위한 단어 추출

```
 [1] "huge"     "museum"     "family"     "enjoyed"    "going"      "here"
 [7] "is"       "huge"       "free"       "lots"       "military"   "memorabilia"
[13] "see"      "loved"      "seeing"     "b"          "close"
```

Lexical Analysis

❖ Lemmatization

- 댓글의 제목(title)과 POS Tagging 처리가 된 댓글 내용 데이터를 Lemmatization 실시
- 모든 텍스트를 소문자로 변경
- 불필요한 구두점(Punctuation), 숫자(Digits), 공백(Space) 제거



	raw_text	lemmatize
9	bits of tears in my eyes worth a visit join the free gui...	bit tear eye worth visit join free guide tour show sou...
10	very interesting and relatively extensive collection ...	interest extensive collection aircraft nice opportunit...
11	gracious memorial to the korean war and interesting ...	gracious memorial korean war interest place view thi...
12	breathtaking display of hope and horror overwhelmin...	breathtaking display hope horror overwhelm size int...
13	good historical museum very sad to see the names o...	good historical museum sad name dead country help...
14	pleasantly surprised we visited this museum becaus...	pleasant surprise visit museum father law interest ko...
15	great lesson on the history of korea this was my first ...	great lesson history korea time impress visit full und...
16	airplanes and tanks galore if you are a history buff thi...	airplane tank galore history buff great place visit mus...
17	must visit loved it i visited war memorial with my fami...	visit love visit war memorial family month june make ...
18	honor and respect our family trip to seoul meant that ...	honour respect family trip seoul mean need pay resp...
19	a must see during your trip to s korea history impress...	trip korea history impressive capture beautiful sacre...
20	war memorial a museum not a monument the name m...	war memorial museum monument slight put wrong dir...

장소 분류

❖ 장소의 특성에 따라 크게 3가지 유형으로 분류

- Case 1 : 명소 & 랜드마크 (Sights & Landmarks)

  ex) 남산, 경복궁, 인사동, 홍대 거리, 한강공원, 63빌딩, 명동성당

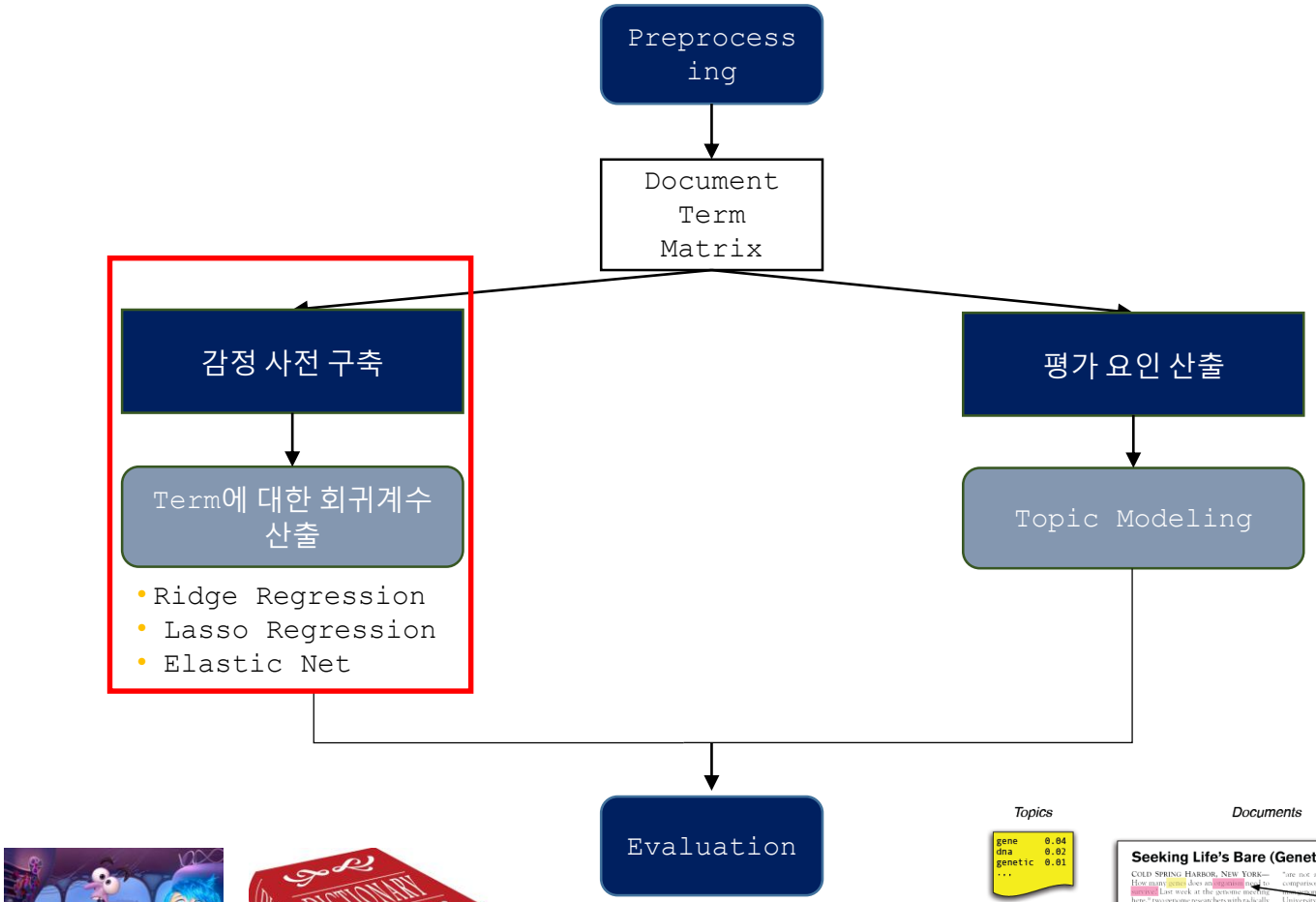
- Case 2 : 쇼핑거리 & 쇼핑몰 & 시장 (Shopping district, Shopping mall, Market)

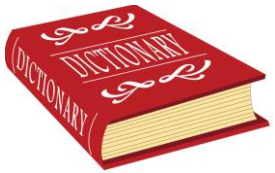
  ex) 명동 쇼핑거리, 코엑스 몰, 롯데 백화점, 타임스퀘어, 광장시장, 노량진 수산시장

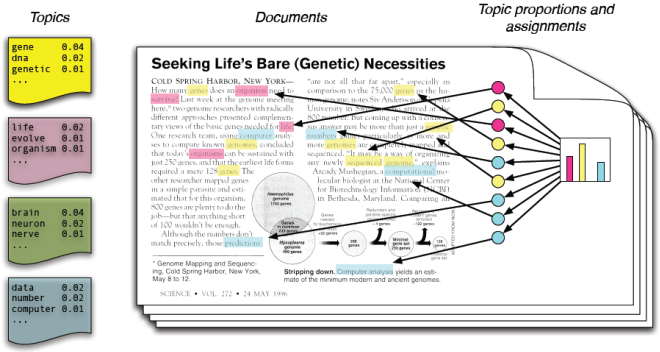
- Case 3 : 박물관 & 전시관 & 미술관 (Museum & Exhibition & Art museum)

  ex) 박물관이 살아있다, 트릭아이 박물관, 전쟁기념관, 국립민속박물관,
  서대문형무소, 국립현대미술관, 코엑스 아쿠아리움

Flow chart


Preprocess ing
Document Term Matrix
감정 사전 구축
Term에 대한 회귀계수 산출
• Ridge Regression
• Lasso Regression
• Elastic Net
평가 요인 산출
Topic Modeling
Evaluation


DICTIONARY


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
life 0.02
evolve 0.01
organism 0.01
brain 0.04
neuron 0.02
nerve 0.01
data 0.02
number 0.02
computer 0.01
Seeking Life's Bare (Genetic) Necessities

감정 분석

❖ 감정 분석

- 텍스트에서 감정 단어를 추출하여 점수화 문장에 사용된 단어로 감정을 예측
- 단어 사전 기반
  - ✓ 장점 : 사용하기 간편
  - ✓ 단점 : 주제에 따라 사전이 달라 짐, 동음이의어 처리 힘듦
    ex)bank, arms, lie, fall
- 기계 학습(Machine learning), 통계학(Statistics) 기반
  - ✓ 장점 : 높은 정확도
  - ✓ 단점 : Over-fitting(과적합) 해결이 필요함
    많은 데이터가 필요함

감정 분석

❖ 회귀분석 기반의 감정 분석

- 사용자의 리뷰에 대한 평점을 회귀분석을 통해 예측가능
- 사용된 단어의 회귀계수가 0보다 크면 긍정, 0보다 작으면 부정, 0이면 중립

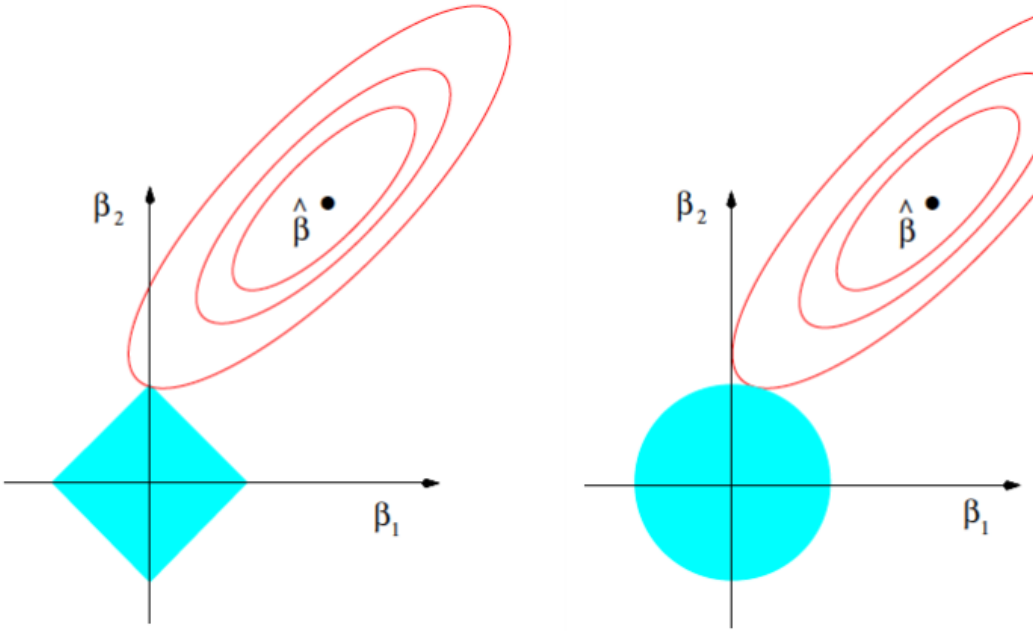
❖ 회귀분석의 문제

- 변수가 많아지면 과적합(overfitting)이 발생
- 회귀계수가 극단적으로 커지거나 작아짐
- 과적합이 발생하면 unseen data에 대한 예측력이 떨어짐
- 과적합을 막아주는 방법이 필요함

감정 분석

❖ 과적합 방지하는 방법론

- Lasso : 작은 회귀계수를 0으로 만듦(Alpha = 1)
- Ridge : 전반적으로 회귀계수를 줄여줌(Alpha = 0)
- Elastic net : lasso + ridge (0 < Alpha < 1)
- 감정과 상관없는 단어들을 0이나 작은 값으로 만들어서 긍/부정 사전을 만들 수 있음



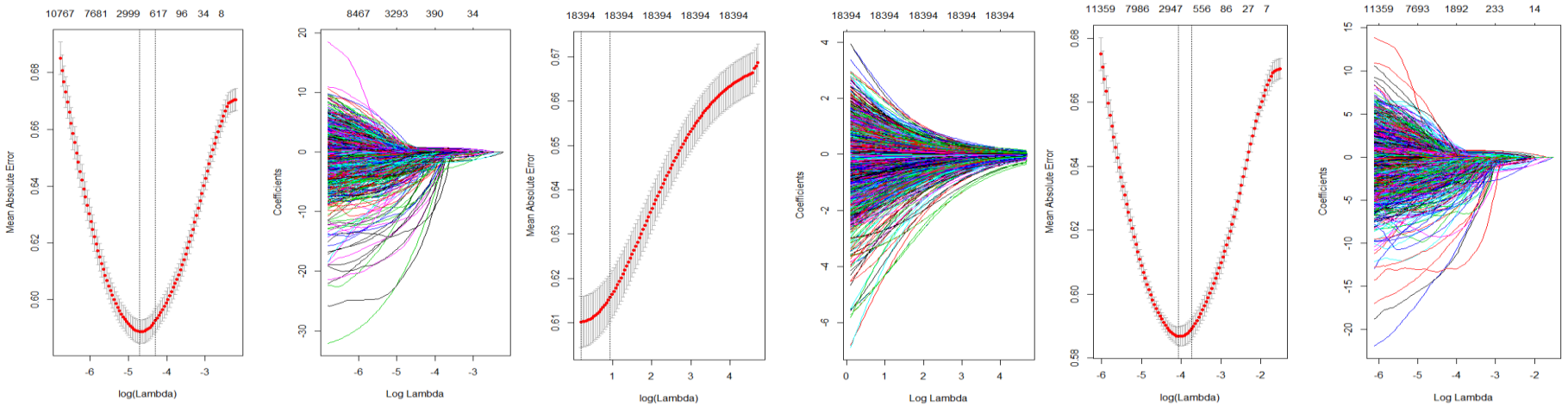
[Lasso] [Ridge]

http://www.stat.cmu.edu/~ryantibs/datamining/lectures/17-modr2.pdf

감정 분석

❖ Lasso vs Ridge vs Elastic net

- 10 - cross validation with MAE (Mean Absolute Error)
- 여러가지 지표를 종합한 결과 Elastic net을 통해서 얻어진 단어들에 대한 회귀계수를 감정점수로 활용



Lasso Ridge Elastic

	lasso	ridge	elastic net
accuracy	0.8383	0.822	0.8358
sensitivity	0.9029	0.8794	0.898
specificity	0.4871	0.5099	0.498
correlation	0.4755	0.4062	0.474

✓ Accuracy, Seneitivity, Specificity의 경우는 Elastic net 모형의 결과 4점이상인 경우는 class 1로 4점미만인 경우는 class 0으로 confusion matrix를 구성하여 얻은 결과물

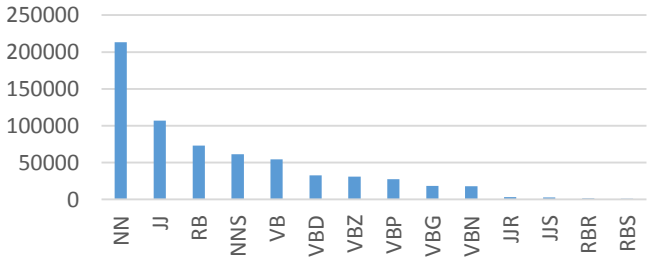
✓ 위의 경우는 case 1(명소 & 랜드마크)에 관한 결과물로 case2, case3의 경우에도 같은 Framework를 적용

감정 분석 : 품사별 감소 비율

❖ Case 1 : 명소 & 랜드마크

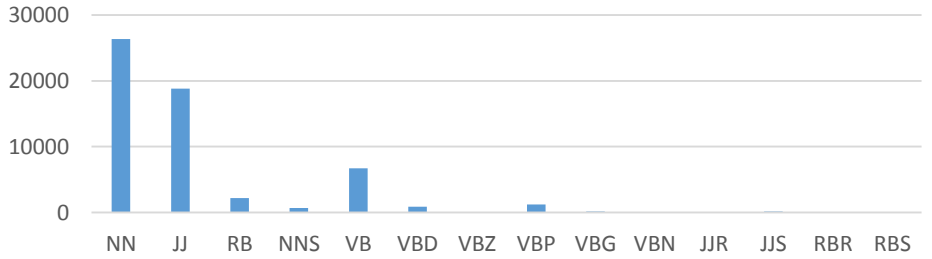
기존

품사	빈도
NN	213,166
JJ	107,133
RB	73,114
NNS	61,249
VB	54,266
VBD	32,905
VBZ	31,198
VBP	27,459
VBG	18,530
VBN	17,926
JJR	3,245
JJS	2,820
RBR	1,688
RBS	894



Elastic net 적용

품사	빈도	감소율
NN	26,340	87.6434
JJ	18,816	82.4368
RB	2,194	96.9992
NNS	676	98.8963
VB	6,699	87.6552
VBD	867	97.3651
VBZ	8	99.9744
VBP	1,227	95.5315
VBG	148	99.2013
VBN	0	100
JJR	8	99.7535
JJS	146	94.8227
RBR	0	100
RBS	0	100



감정 분석 : 감정 단어

❖ Case 1 : 명소 & 랜드마크

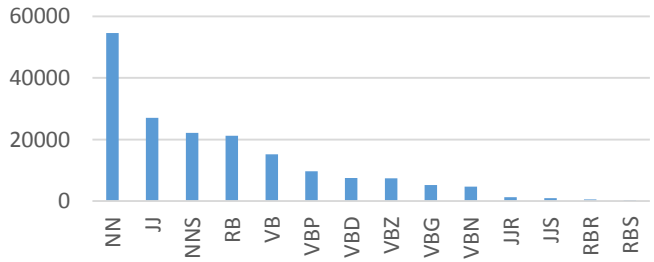
긍정	부정

감정 분석 : 품사별 감소 비율

❖ Case 2 : 쇼핑 거리& 시장

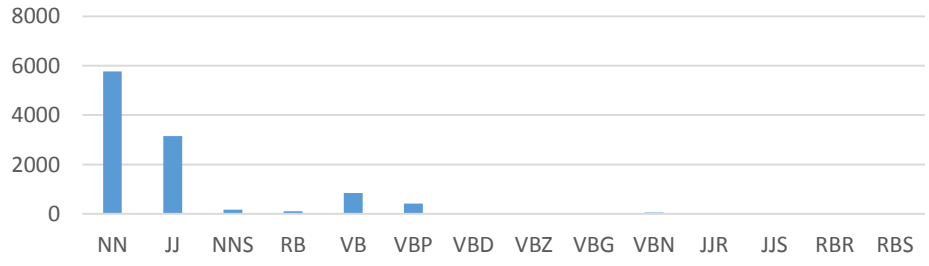
기존

품사	빈도
NN	54,575
JJ	27,009
RB	22,138
NNS	21,215
VB	15,175
VBD	9,714
VBZ	7,456
VBP	7,426
VBG	5,193
VBN	4,627
JJR	1,210
JJS	974
RBR	480
RBS	312



Elastic net 적용

품사	빈도	감소율
NN	5,768	89.4310
JJ	3,153	88.3261
RB	173	99.2185
NNS	112	99.4721
VB	844	94.4382
VBD	423	95.6455
VBZ	31	99.5842
VBP	0	100
VBG	7	99.8652
VBN	57	98.7681
JJR	4	99.6694
JJS	8	99.1786
RBR	0	100
RBS	0	100



감정 분석 : 감정 단어

❖ Case 2 : 쇼핑 거리& 시장

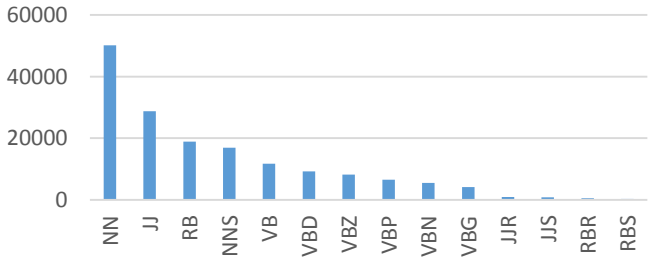
긍정	부정
	

감정 분석 : 품사별 감소 비율

❖ Case 3 : 박물관 & 전시관 & 미술관

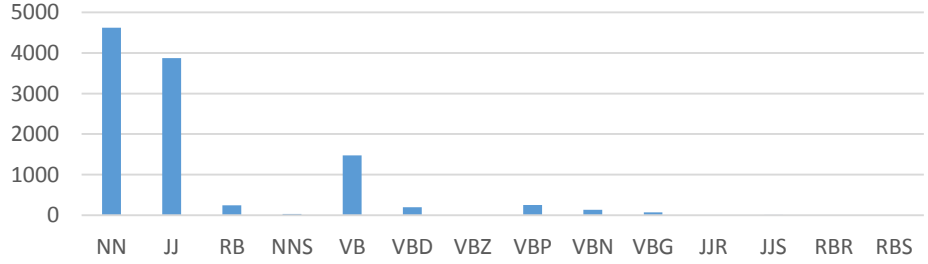
기존

품사	빈도
NN	50187
JJ	28796
RB	18866
NNS	16884
VB	11728
VBD	9182
VBZ	8128
VBP	6500
VBG	5477
VBN	4083
JJR	868
JJS	766
RBR	425
RBS	245



Elastic net 적용

품사	빈도	감소율
NN	4619	90.79642
JJ	3872	86.55369
RB	241	98.72257
NNS	29	99.82824
VB	1471	87.45737
VBD	199	97.83272
VBZ	0	100
VBP	249	96.16923
VBG	131	97.60818
VBN	68	98.33456
JJR	0	100
JJS	10	98.69452
RBR	0	100
RBS	0	100

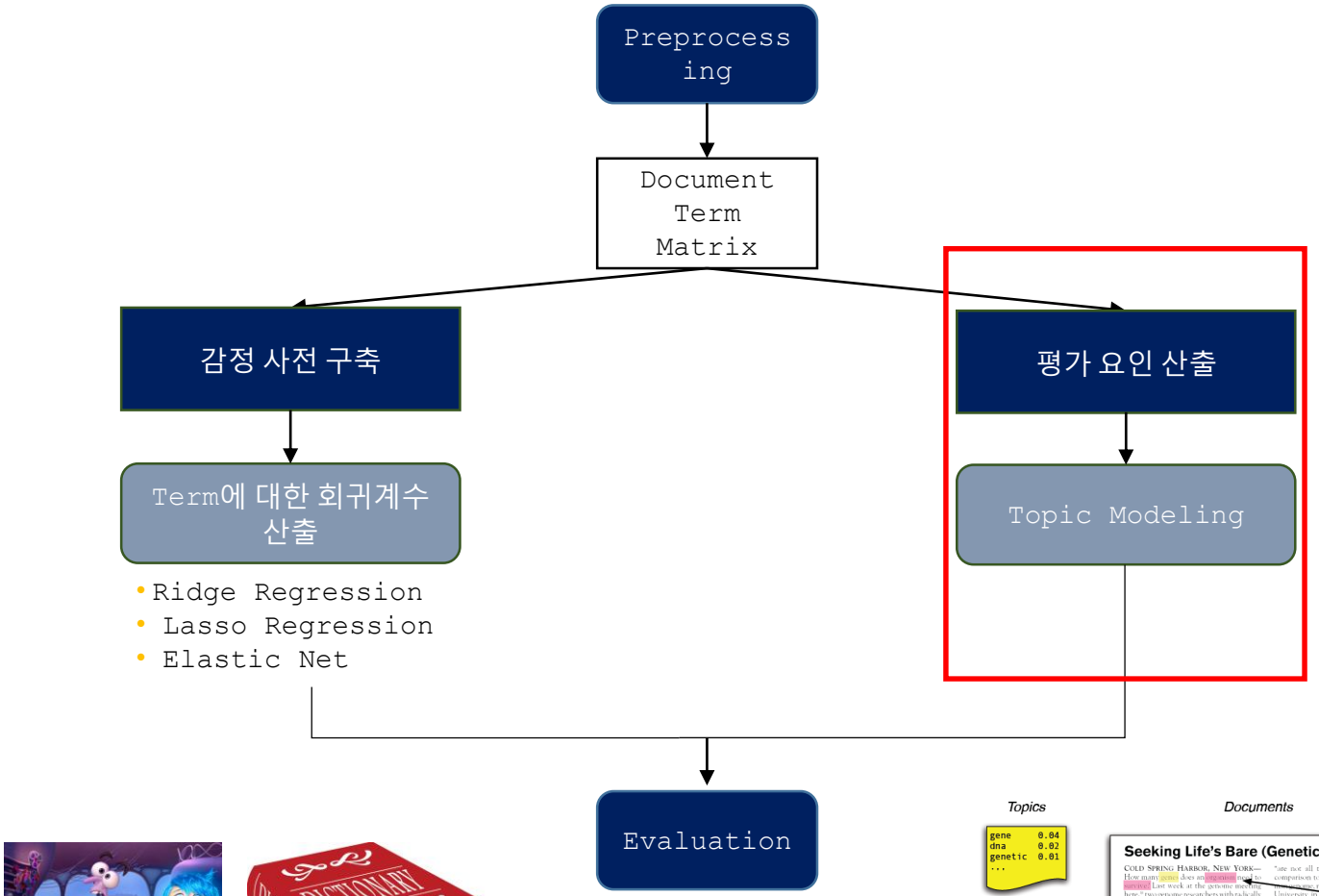


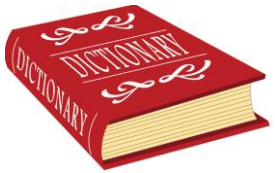
감정 분석 : 단어

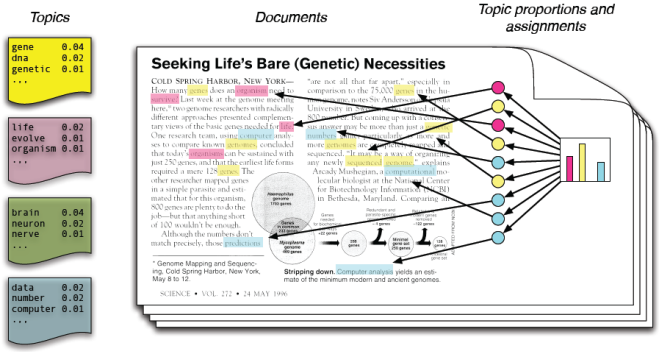
❖ Case 3 : 박물관 & 전시관 & 미술관

긍정	부정

Flow chart


Preprocess ing
Document Term Matrix
감정 사전 구축
Term에 대한 회귀계수 산출
• Ridge Regression
• Lasso Regression
• Elastic Net
평가 요인 산출
Topic Modeling
Evaluation


DICTIONARY
DICTIONARY


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
life 0.02
evolve 0.01
organism 0.01
brain 0.04
neuron 0.02
nerve 0.01
data 0.02
number 0.02
computer 0.01
Seeking Life's Bare (Genetic) Necessities

Topic Modeling

❖ Case 1 : 명소 & 랜드마크

복잡함	전통	경치	음식&카페	장소의 분위기
ride	tour	top	food	time
time	garden	night	restaurant	relax
kid	guide	hike	souvenir	night
crowd	guard	time	find	river
ticket	time	lock	art	building
long	change	bus	tourist	stroll
korea	Secret	station	cafe	middle
game	Building	hill	tea	lovely
theme	Korea	climb	sell	peaceful
wait	Ground	couple	local	blossom
indoor	museum	long	store	cherry
attraction	english	high	small	architecture
food	free	hour	price	water
hour	hour	crowd	time	busy
line	ceremony	ticket	cheap	quiet
queue	architecture	restaurant	coffee	bike
child	historical	reach	crowd	sit
outdoor	entrance	bring	korea	festival
inside	site	museum	young	lantern
family	miss	trip	item	centre

Topic Modeling

❖ Case 2 : 쇼핑 거리& 시장

화장품 쇼핑 용이	의사 소통	저녁 여행	합리적인 쇼핑	먹거리
cosmetic	speak	night	find	seafood
product	english	restaurant	cheap	local
brand	chinese	time	clothes	stall
find	find	busy	bargain	restaurant
beauty	staff	find	item	fresh
paradise	tax	hotel	time	cook
restaurant	time	brand	shoe	pancake
cafe	sale	stay	stuff	live
night	purchase	local	souvenir	vendor
skin	back	open	clothe	crab
cheap	item	subway	quality	fry
till	refund	close	stall	taste
drop	duty	evening	local	octopus
clothes	language	high	hour	delicious
shopper	cosmetic	end	bag	sashimi
skincare	mandarin	stall	accessory	find
care	airport	stop	open	cheap
mask	pay	atmosphere	low	win
face	service	vendor	high	time
sample	card	court	variety	serve

Topic Modeling

❖ Case 3 : 박물관 & 전시관 & 미술관

박물관 전시	미술품	자녀 동반	재미	유익한 정보
free	collection	kid	photo	display
time	building	fish	picture	plane
hour	exhibit	child	ice	exhibit
tour	floor	small	time	time
spend	free	animal	friend	learn
english	display	mall	hour	move
exhibit	exhibition	family	find	informative
learn	time	traditional	trick	inside
display	architecture	people	eye	people
guide	work	food	people	free
culture	gallery	walk	spend	ship
easy	artifact	inside	pose	north
walk	build	adult	crowd	understand
information	culture	area	small	hour
recommend	shop	sea	funny	walk
inside	piece	shop	camera	spend
informative	space	price	kid	conflict
trip	ground	time	family	educational
exhibition	easy	entrance	recommend	find
city	paint	expect	ticket	feel

토픽 별 감정점수

❖ 계산 방법

DTM	Word1	Word2
doc1	3	1
doc2	1	5

•

감정 사전	감정 점수
word1	5
word2	3

×

평가 요인	topic1	topic2
word1	0.7	0.2
word2	0.3	0.8

=

토픽별 감정 점수	topic1	topic2
document1	3*5*0.7+1*3*0.3 =11.4	3*5*0.2+1*3*0.8 =5.4
document2	1*5*0.7+5*3*0.3 =8	1*5*0.2+5*3*0.8 =13

토픽별 감정점수에서

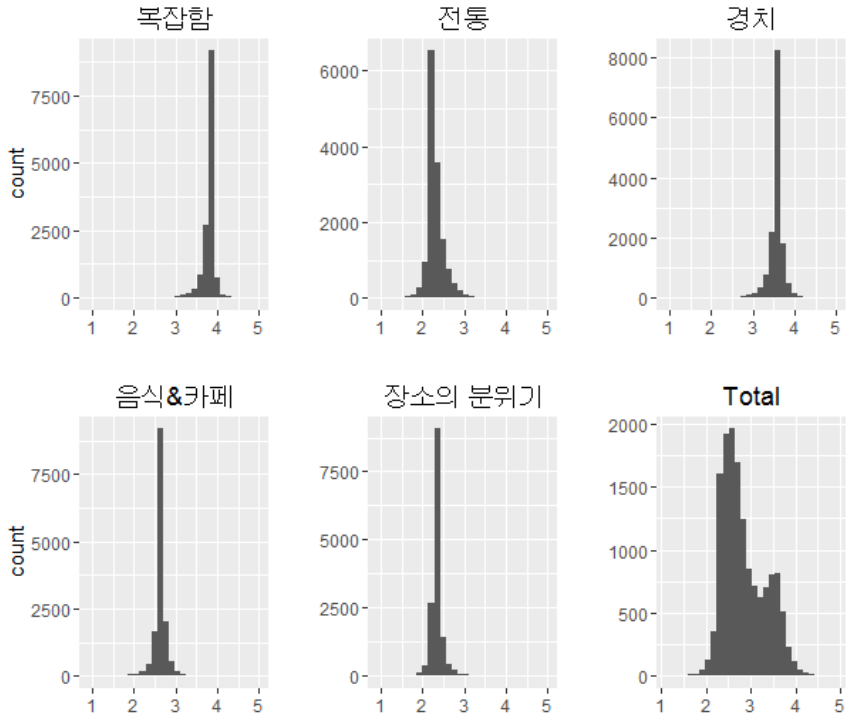
$\left(\frac{x-\min(x)}{\max(x)-\min(x)}\text{x4}\right)+1$ **[min-max 정규화]** 를 수행하면

5점 척도로 감정 점수를 구할 수 있음

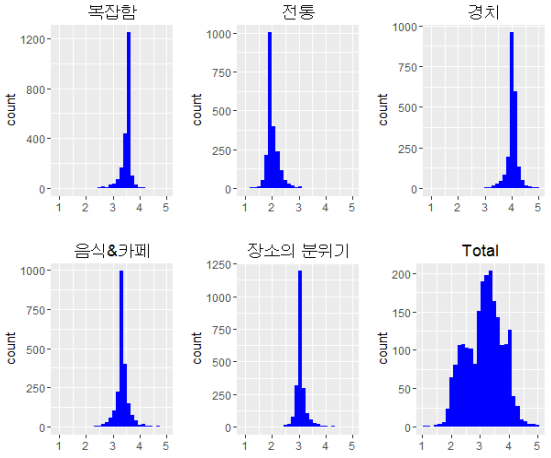
토픽 별 감정점수

❖ Case 1 : 명소 & 랜드마크

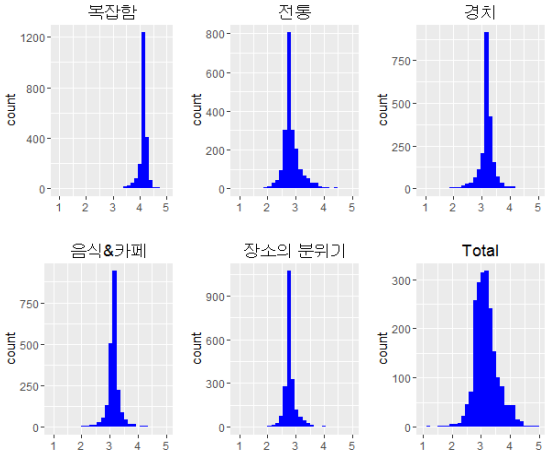
[전체 감정점수]



[경복궁]



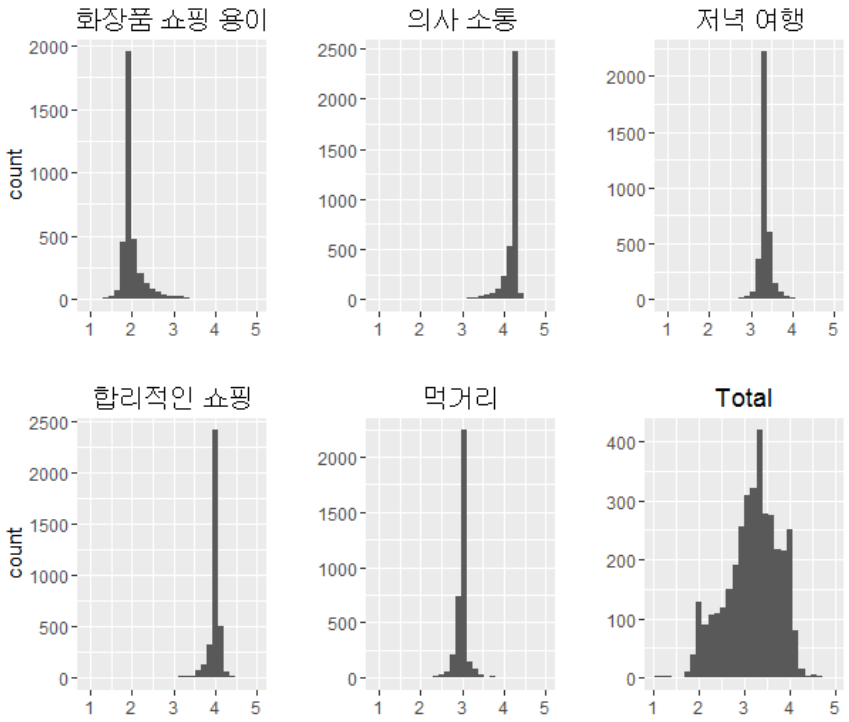
[남산 타워]



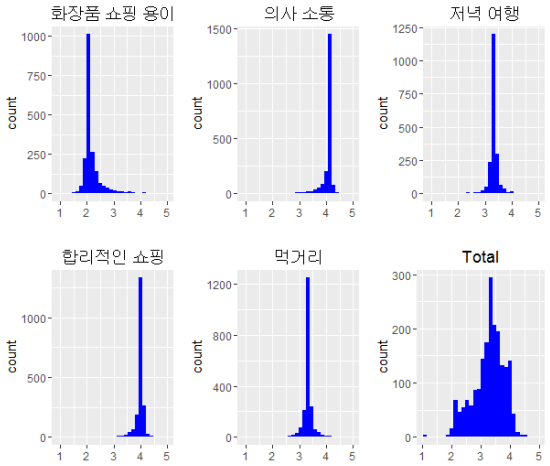
토픽 별 감정점수

❖ Case 2 : 쇼핑거리 & 쇼핑몰 & 시장

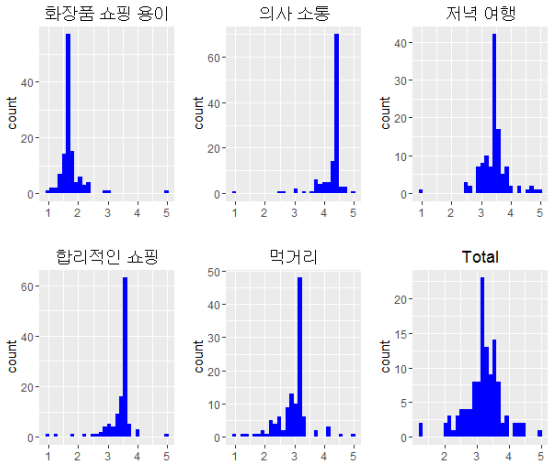
[전체 감정점수]



[명동]



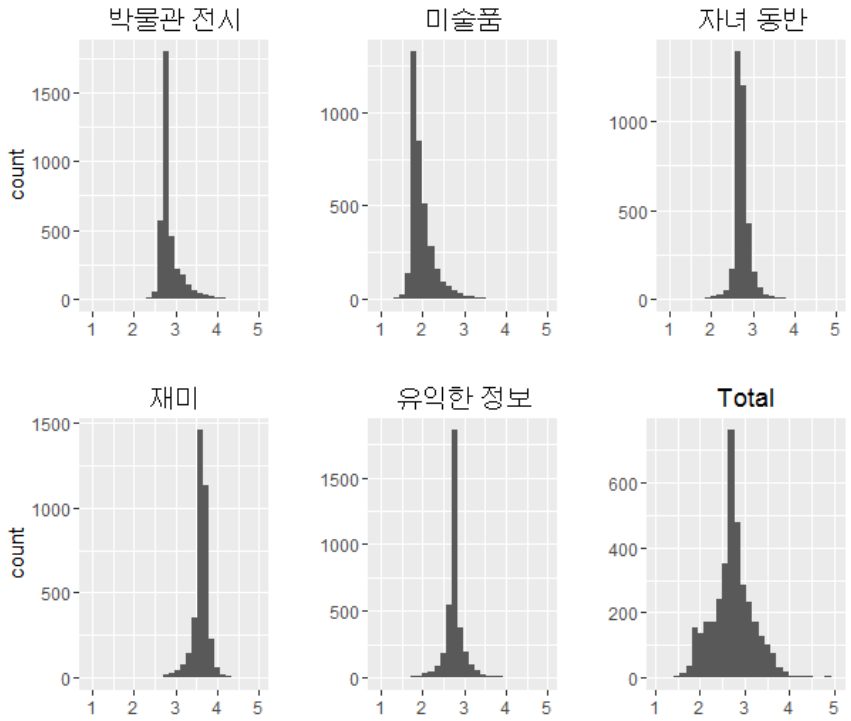
[동대문 종합시장]



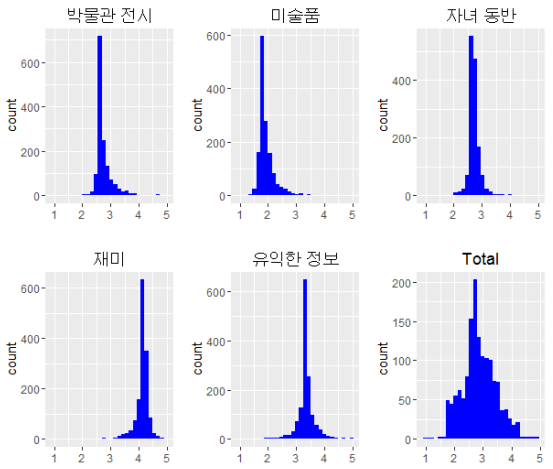
토픽 별 감정점수

❖ Case 3 : 박물관 & 전시관 & 미술관

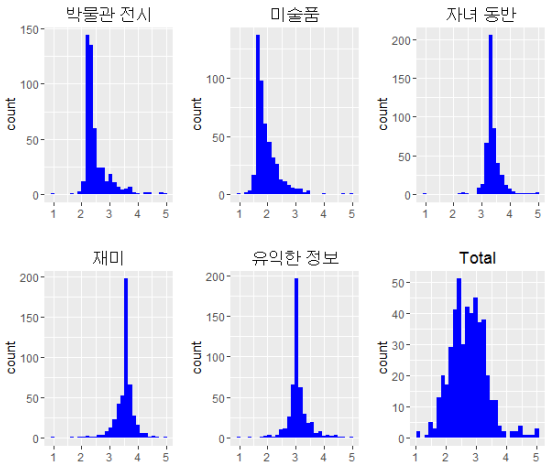
[전체 감정점수]



[전쟁 기념관]



[트릭아이 미술관]



결론

❖ Contribution

- 회귀분석(ridge, lasso, elastic net)을 통해 의견(Comment)별 점수를 예측하고, 회귀계수를 통해 감정사전을 구축할 수 있음
- 다수의 사람들의 의견(comment)를 Topic modeling에 적용하여 여행지의 평가 요인을 얻어내는 방법론을 제안
- 감정 사전과 Topic modeling의 결과를 이용하여 평가 요인에 대해 점수를 산출하는 방법론을 제안

❖ Limitation

- 평가사이트의 경우 별점의 분포가 치우쳐져 있어 회귀분석을 적용하기 어려움(정규성 가정)
- Topic modeling에서 Topic naming을 하기 어려움

질의 응답