

텍스트 마이닝을 이용한 물류분야 논문의 기술동향 분석



2018 -1학기 비정형데이터분석 Team project

4조 권구포 유지원 전은석 김웅 음수민

목차

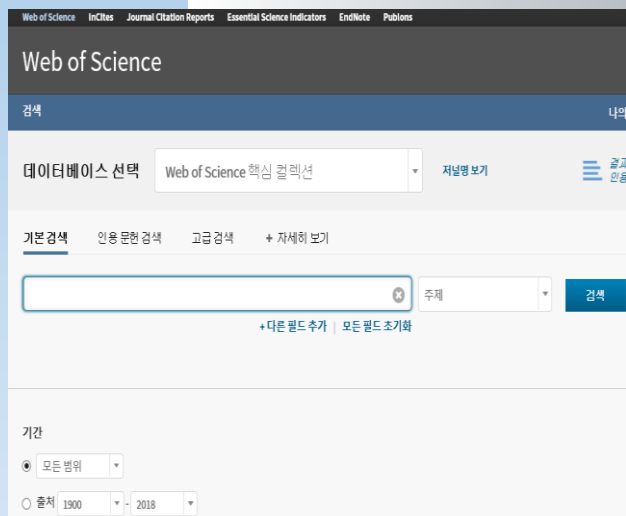
1. 개요
 - 1.1. 연구배경 및 목적
 - 1.2. 프레임워크
 - 1.3. 데이터 설명
2. 연구방법
 - 2.1. 빈도분석
 - 2.2. 네트워크 분석
 - 2.3. 토픽모델링
3. 결론
 - 3.1. 기술 트렌드 분석
 - 3.1. 연도별 특징 키워드 분석

✓ 물류 관련 논문의 Abstract 분석을 통해 연구 트렌드 파악

- 최근 물류 트렌드를 과거 논문의 Abstract의 Keyword 통해 찾아보고자 함
- 과거의 논문자료를 Review하여 앞으로의 물류 트렌드를 예측해보고, 물류 연구분야자들의 연구논문주제 선정 및 발전 가능한 분야를 알아보고자 함

✓ KEYWORD

- Logistics + (shipping, technology, status) **총 31,040건**



Web of Science
(물류관련 keyword 검색)

Data Sources

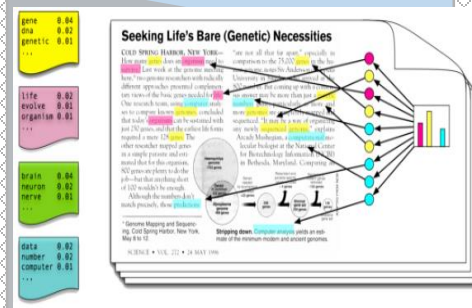
Data Crawling

Preprocessing

Document Summarization

Keywords

Analyze



Topic Modeling

**Abstract 분석을 통해
연구트렌드 파악**

Output



- 검색 전체를 연결하고 다음을 통해 프로세스를 탐색 함 (수집기간 : 2008년 ~ 2018년)





데이터 필터링

- Logistics + (shipping, technology, status) **총 31,040건**
- logistics regression과 같은 통계관련 논문과 의학관련 논문이 대다수를 차지한다는 문제점 발생.

[illegible]

1차 정리

- 4,827건의 논문
- 데이터를 거의 80%이상 제거하였음에도 불구하고, 단어들을 보았을 때 한눈에 보아도 비관련 단어들이 다수 포함.

2차 정리

- 2,767건의 논문
- 최종 데이터로 선정.

연도별 물류관련 논문 빈도



연도별 물류관련 논문 빈도



❖ 빈도수 TOP 20

✓ TF

```
> head(d1, 20)
```

word	freq
use	3899
logist	3770
model	3046
technolog	2843
studi	2758
develop	2006
result	1832
system	1821
product	1775
research	1744
paper	1701
suppli	1616
chain	1575
data	1524
manag	1519
process	1466
can	1362
factor	1328
analysi	1316
perform	1290

✓ TF-IDF

```
> head(d2, 20)
```

word	freq
model	28.05116
port	27.43181
suppli	27.39090
chain	27.33605
product	26.36454
ship	26.12399
system	25.61474
transport	25.15561
cost	23.21806
servic	23.21749
process	22.98514
manag	22.53593
perform	21.98245
research	21.36531
oper	21.26894
adopt	21.26211
technolog	21.13065
contain	20.88974
factor	20.78465
industri	20.73998

❖ 500번 이상 나타난 단어

```
> #빈도수별 상위 난500번이상 나타나 단어 확인
> findFreqTerms(abs_dtm, lowfreq=500, highfreq=Inf)
```

[1] "adoption"	"also"	"among"
[4] "analysis"	"approach"	"based"
[7] "can"	"case"	"chain"
[10] "cost"	"costs"	"data"
[13] "design"	"developed"	"development"
[16] "different"	"elsevier"	"factors"
[19] "findings"	"high"	"however"
[22] "impact"	"important"	"industry"
[25] "information"	"level"	"logistic"
[28] "logistics"	"management"	"may"
[31] "method"	"model"	"models"
[34] "network"	"new"	"one"
[37] "order"	"paper"	"performance"
[40] "port"	"potential"	"problem"
[43] "process"	"production"	"proposed"
[46] "purpose"	"regression"	"research"
[49] "reserved"	"results"	"rights"
[52] "risk"	"service"	"shipping"
[55] "significant"	"status"	"study"
[58] "supply"	"system"	"systems"
[61] "technologies"	"technology"	"time"
[64] "transport"	"transportation"	"two"
[67] "use"	"used"	"using"
[70] "well"	"will"	

❖ 빈도수 TOP 20

✓ TF

```
> head(d1, 20)
```

	word	freq
use	use	3899
logist	logist	3770
model	model	3046
technolog	technolog	2843
studi	studi	2758
develop	develop	2006
result	result	1832
system	system	1821
product	product	1775
research	research	1744
paper	paper	1701
suppli	suppli	1616
chain	chain	1575
data	data	1524
manag	manag	1519
process	process	1466
can	can	1362
factor	factor	1328
analysi	analysi	1316
perform	perform	1290

✓ TF-IDF

```
> head(d2, 20)
```

	word	freq
model	model	28.05116
port	port	27.43181
suppli	suppli	27.39090
chain	chain	27.33605
product	product	26.36454
ship	ship	26.12399
system	system	25.61474
transport	transport	25.15561
cost	cost	23.21806
servic	servic	23.21749
process	process	22.98514
manag	manag	22.53593
perform	perform	21.98245
research	research	21.36531
oper	oper	21.26894
adopt	adopt	21.26211
technolog	technolog	21.13065
contain	contain	20.88974
factor	factor	20.78465
industri	industri	20.73998

❖ 500번 이상 나타난 단어

```
> #빈도수별 상위 난500번이상 나타나 단어 확인
> findFreqTerms(abs_dtm, lowfreq=500, highfreq=Inf)
```

[1]	"adoption"	"also"	"among"
[4]	"analysis"	"approach"	"based"
[7]	"can"	"case"	"chain"
[10]	"cost"	"costs"	"data"
[13]	"design"	"developed"	"development"
[16]	"different"	"elsevier"	"factors"
[19]	"findings"	"high"	"however"
[22]	"impact"	"important"	"industry"
[25]	"information"	"level"	"logistic"
[28]	"logistics"	"management"	"may"
[31]	"method"	"model"	"models"
[34]	"network"	"new"	"one"
[37]	"order"	"paper"	"performance"
[40]	"port"	"potential"	"problem"
[43]	"process"	"production"	"proposed"
[46]	"purpose"	"regression"	"research"
[49]	"reserved"	"results"	"rights"
[52]	"risk"	"service"	"shipping"
[55]	"significant"	"status"	"study"
[58]	"supply"	"system"	"systems"
[61]	"technologies"	"technology"	"time"
[64]	"transport"	"transportation"	"two"
[67]	"use"	"used"	"using"
[70]	"well"	"will"	

❖ 빈도수 TOP 20

✓ TF

```
> head(d1, 20)
```

word	freq
use	3899
logist	3770
model	3046
technolog	2843
studi	2758
develop	2006
result	1832
system	1821
product	1775
research	1744
paper	1701
suppli	1616
chain	1575
data	1524
manag	1519
process	1466
can	1362
factor	1328
analysi	1316
perform	1290

✓ TF-IDF

```
> head(d2, 20)
```

word	freq
model	28.05116
port	27.43181
suppli	27.39090
chain	27.33605
product	26.36454
ship	26.12399
system	25.61474
transport	25.15561
cost	23.21806
servic	23.21749
process	22.98514
manag	22.53593
perform	21.98245
research	21.36531
oper	21.26894
adopt	21.26211
technolog	21.13065
contain	20.88974
factor	20.78465
industri	20.73998

- TF-IDF를 고려하였을 때,
Port, Suppli, Ship, Transport, Cost 등
물류관련 용어들이 상위권에 속한 것을 확인.

(wordcloud)

Non TF-IDF



- # Study Logistic research

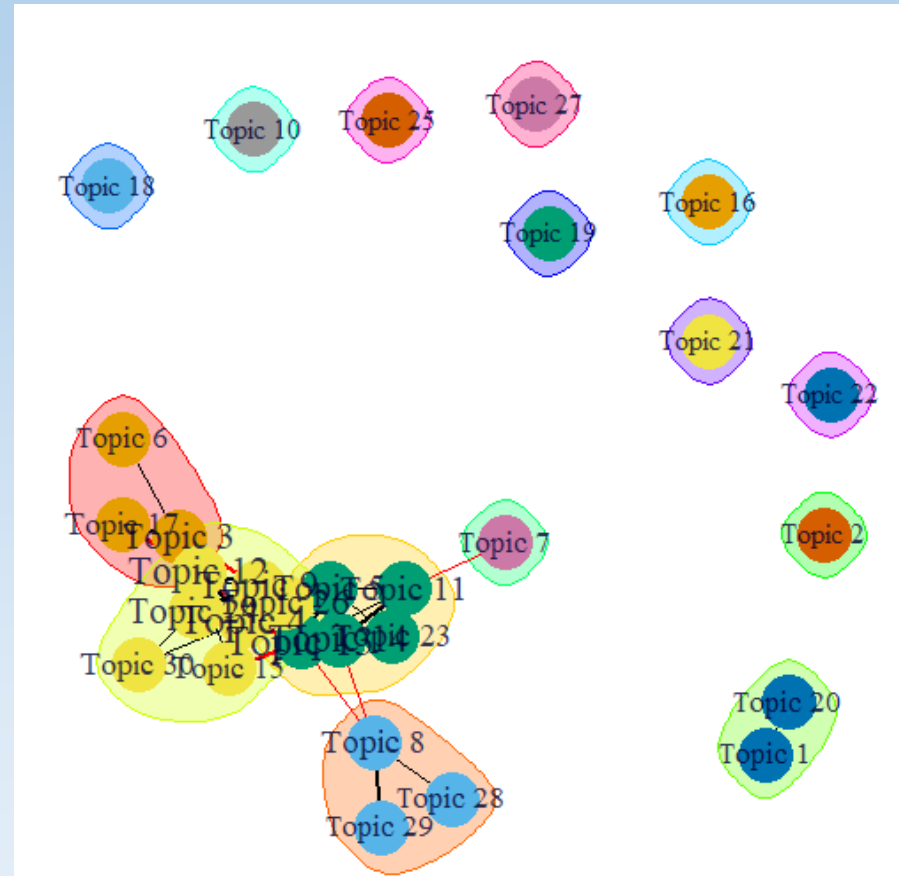
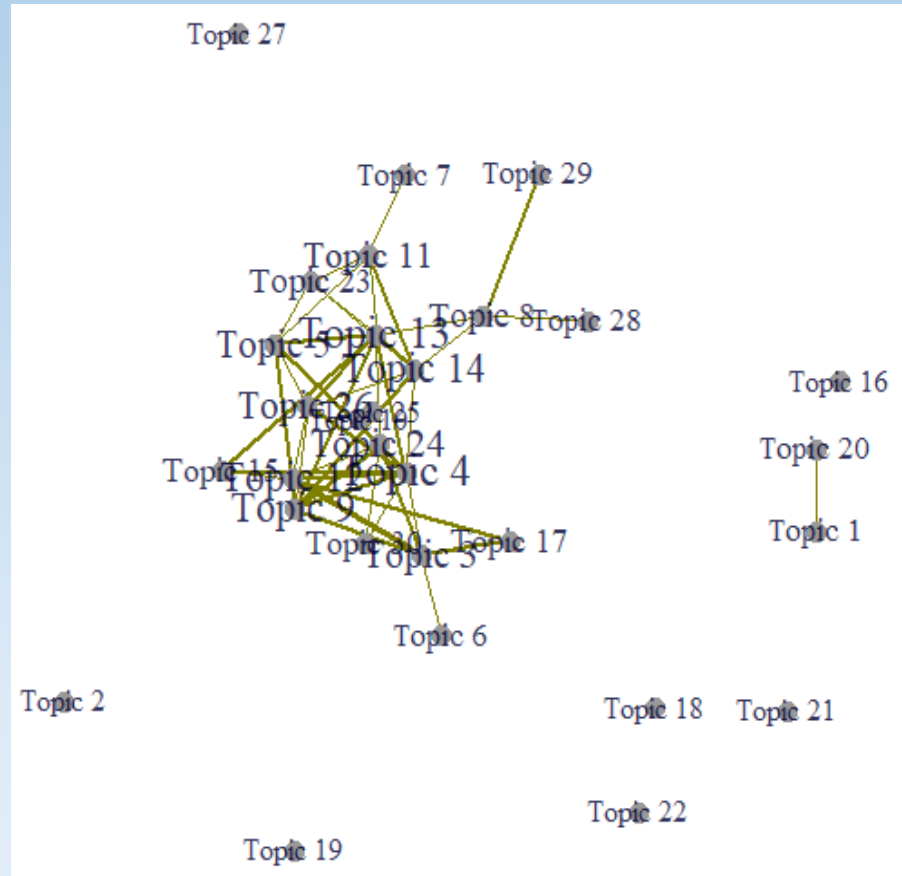
TF-IDF



- # Logistics Container Port



Topic Network & Communities





토픽모델링

A	B	C	D	E
Author	Title	Source	Abstract	Year
1	Ma, S.C.	China's on	China's on	2018
2	Tran, T.K.V.	Gender di	ASIAPAI in Vietnam	2018
3	Liu, M. Y.	Quantum-PROCEED	Process pl	2018
4	Macheka, Identificat	FOOD CO	Postharvest	2018
5	Macheka, Identificat	FOOD CO	Postharvest	2018
6	Macheka, Identificat	FOOD CO	Postharvest	2018
7	Faradinas, Electronic	HABRI RE	Background	2018
8	Kim, S.H.	A Case stud	BUILDING in this stu	2018
9	Thiede, B.C.	Working JOURNAL	A secure j	2018
10	Bogaert, A.	Evaluating ANNALS	C The purp	2018
11	Zimmerm	The releva	ZETTSCHR	The Wisco
12	Laak	Journal	JOURNAL	The Califo
13	Liu, M. Y.	Quantum-PROCEED	Process pl	2018
14	Macheka, Identificat	FOOD CO	Postharvest	2018
15	Macheka, Identificat	FOOD CO	Postharvest	2018
16	Macheka, Identificat	FOOD CO	Postharvest	2018
17	Faradinas, Electronic	HABRI RE	Background	2018
18	Kim, S.H.	A Case stud	BUILDING in this stu	2018
19	Thiede, B.C.	Working JOURNAL	A secure j	2018
20	Bogaert, A.	Evaluating ANNALS	C The purp	2018
21	Zimmerm	The releva	ZETTSCHR	The Wisco
22	Laak	Journal	JOURNAL	The Califo
23	Liu, M. Y.	Quantum-PROCEED	Process pl	2018
24	Macheka, Identificat	FOOD CO	Postharvest	2018
25	Macheka, Identificat	FOOD CO	Postharvest	2018
26	Macheka, Identificat	FOOD CO	Postharvest	2018
27	Faradinas, Electronic	HABRI RE	Background	2018
28	Kim, S.H.	A Case stud	BUILDING in this stu	2018
29	Thiede, B.C.	Working JOURNAL	A secure j	2018
30	Bogaert, A.	Evaluating ANNALS	C The purp	2018
31	Zimmerm	The releva	ZETTSCHR	The Wisco

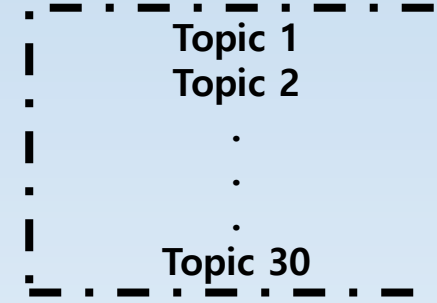
Tokenizing

A	B	C	D	E
Author	Title	Source	Abstract	Year
1	Ma, S.C.	China's on	China's on	2018
2	Tran, T.K.V.	Gender di	ASIAPAI in Vietnam	2018
3	Liu, M. Y.	Quantum-PROCEED	Process pl	2018
4	Macheka, Identificat	FOOD CO	Postharvest	2018
5	Macheka, Identificat	FOOD CO	Postharvest	2018
6	Macheka, Identificat	FOOD CO	Postharvest	2018
7	Faradinas, Electronic	HABRI RE	Background	2018
8	Kim, S.H.	A Case stud	BUILDING in this stu	2018
9	Thiede, B.C.	Working JOURNAL	A secure j	2018
10	Bogaert, A.	Evaluating ANNALS	C The purp	2018
11	Zimmerm	The releva	ZETTSCHR	The Wisco
12	Laak	Journal	JOURNAL	The Califo
13	Liu, M. Y.	Quantum-PROCEED	Process pl	2018
14	Macheka, Identificat	FOOD CO	Postharvest	2018
15	Macheka, Identificat	FOOD CO	Postharvest	2018
16	Macheka, Identificat	FOOD CO	Postharvest	2018
17	Faradinas, Electronic	HABRI RE	Background	2018
18	Kim, S.H.	A Case stud	BUILDING in this stu	2018
19	Thiede, B.C.	Working JOURNAL	A secure j	2018
20	Bogaert, A.	Evaluating ANNALS	C The purp	2018
21	Zimmerm	The releva	ZETTSCHR	The Wisco
22	Laak	Journal	JOURNAL	The Califo
23	Liu, M. Y.	Quantum-PROCEED	Process pl	2018
24	Macheka, Identificat	FOOD CO	Postharvest	2018
25	Macheka, Identificat	FOOD CO	Postharvest	2018
26	Macheka, Identificat	FOOD CO	Postharvest	2018
27	Faradinas, Electronic	HABRI RE	Background	2018
28	Kim, S.H.	A Case stud	BUILDING in this stu	2018
29	Thiede, B.C.	Working JOURNAL	A secure j	2018
30	Bogaert, A.	Evaluating ANNALS	C The purp	2018
31	Zimmerm	The releva	ZETTSCHR	The Wisco

Corpus Data

Term-Document Matrix

Topic modeling



Trend

HOT

OR

COLD

단어제거

*LDA Iteration : 5000

✓ 각 토픽의 연도별 추이의 상승과 하락 판단

✓ 선형회귀분석 수행 후, 선형회귀계수 값을 통해 판별

*중속변수=키워드의 연도별 비중 평균값

*독립변수=연도

❖ Topic Extraction (30 Topics)

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic10	Topic11	Topic12	Topic13	Topic14	Topic15
concentr	select	associ	logist	develop	forest	vessel	biomass	model	food	ship	use	system	model	internet
water	evalu	risk	firm	technolog	popul	wind	product	use	sensor	port	regress	logist	problem	mobil
treatment	decis	age	perform	logist	mortal	offshor	cost	data	risk	transport	status	technolog	optim	user
growth	fuzzi	status	servic	new	speci	speed	suppli	logist	safeti	contain	factor	process	cost	digit
process	criteria	year	factor	chang	habitat	ship	plant	analysi	vehicl	freight	survey	use	solut	servic
cell	analyt	factor	use	global	season	oper	use	predict	driver	termin	social	inform	time	communic
pressur	process	use	adopt	research	surviv	forecast	energi	differ	cold	maritim	group	manag	ship	iot
temperatur	hierarchi	signific	technolog	econom	fish	north	chain	estim	temperatur	servic	particip	develop	logist	diffus
function	prefer	regress	relationship	countri	chang	wave	feedstock	regress	traffic	cargo	signific	can	decis	thing
resist	decisionmak	logist	research	polici	site	condit	produc	perform	accid	compani	examin	product	oper	ict
Topic16	Topic17	Topic18	Topic19	Topic20	Topic21	Topic22	Topic23	Topic24	Topic25	Topic26	Topic27	Topic28	Topic29	Topic30
student	use	rfid	associ	plant	construct	chines	industri	adopt	imag	suppli	bank	product	energi	area
learn	smoke	identif	protein	machin	project	china	logist	farmer	map	chain	milk	market	emiss	map
educ	smoker	tag	genotyp	tree	materi	ecommerc	product	household	space	research	anim	consum	environmen t	landslid
academ	drug	frequenc	diseas	fire	design	crossbord	korean	agricultur	secur	manag	vaccin	retail	fuel	suscept
technolog	alcohol	radio	sequenc	stand	mainten	team	network	use	chaotic	purpos	embryo	revers	carbon	use
robot	among	chain	gene	crop	tunnel	block	countri	factor	encrypt	literatur	semen	recycl	technolog	factor
school	associ	track	genet	seedl	inspect	cluster	inc	farm	surfac	collabor	infect	wast	reduct	high
univers	current	technolog	clinic	seed	naval	militari	associ	food	mission	review	group	return	green	field
teach	year	suppli	fractur	yield	bim	intellectu	china	adapt	print	practic	cell	manufactur	impact	spatial
scienc	gambl	traceabl	allel	plot	safeti	member	host	technolog	sequenc	find	treatment	chain	consumpt	soil

❖ Topic Extraction (30 Topics)

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic10	Topic11	Topic12	Topic13	Topic14	Topic15	
Sensing concentr water treatment growth process cell pressur temperatur function resist	select evalu	Customer risk	logist firm	Port technolog	forest popul	Marine shipping	biomass product	Supply Chain	food sensor	Cross border	use regress	Communication Technology	model problem	Integration internet mobil	
	age status year		perform service factor		logist new chang		offshor speed ship		cost logist analysi		risk safeti freight		transport contain survey		status social use
	analyt	factor	use	global	season	oper	use	predict	driver	termin	social	inform	time communic		
	process	hierarchi	signific	technolog	econom	fish	north	chain	estim	temperatur	servic	particip	develop	logist diffus	
	process	prefer	regress	relationship	countri	chang	wave	feedstock	regress	traffic	cargo	signific	can	decis thing	
	decisionmak	logist	research	polici	site	condit	produc	perform	accid	compani	examin	product	oper	ict	
	Topic16	Topic17	Topic18	Topic19	Topic20	Topic21	Topic22	Topic23	Topic24	Topic25	Topic26	Topic27	Topic28	Topic29	Topic30
	Education learn educ academ technolog robot school univers teach scienc	use smoke	RFID	associ protein	Environment	construct project	Ship building	industri logist	Rural	imag map	Green logistics	bank milk	Reverse logistics	energi emiss	Hazard logistics
tag frequenc		typ class		tree life		ecommerc design		household agricultur		space secur		research manag		consum retail	
drug		radio	sequenc	stand	mainten	team	network	use	chaotic	purpos	embryo	revers	carbon use		
alcohol		chain	gene	crop	tunnel	block	countri	factor	encrypt	literatur	semen	recycl	technolog factor		
among		track	genet	seedl	inspect	cluster	inc	farm	surfac	collabor	infect	wast	reduct high		
associ		technolog	clinic	seed	naval	militari	associ	food	mission	review	group	return	green field		
current		suppli	fractur	yield	bim	intellectu	china	adapt	print	practic	cell	manufactur	impact spatial		
year		traceabl	allel	plot	safeti	member	host	technolog	sequenc	find	treatment	chain	consumpt soil		
gambl															

❖ Topic Extraction (30 Topics)

TOPIC 1	Sensing
TOPIC 2	Decision making
TOPIC 3	Customer
TOPIC 4	Technology adoption
TOPIC 5	Port
TOPIC 6	Logistics regulation
TOPIC 7	Marine shipping
TOPIC 8	Biomass
TOPIC 9	Supply Chain
TOPIC 10	Cold chain
TOPIC 11	Cross border
TOPIC 12	Social
TOPIC 13	Communication Technology
TOPIC 14	Optimization
TOPIC 15	Integration

TOPIC 16	Education
TOPIC 17	Bad Behavior
TOPIC 18	RFID
TOPIC 19	DNA experiment
TOPIC 20	Environment
TOPIC 21	Container Terminal
TOPIC 22	ship building
TOPIC 23	Marine Terminal
TOPIC 24	Rural
TOPIC 25	Network
TOPIC 26	Green logistics
TOPIC 27	Farm SCM
TOPIC 28	Reverse logistics
TOPIC 29	CO2 emission
TOPIC 30	Hazard logistics



결과 : 물류 주요 토픽

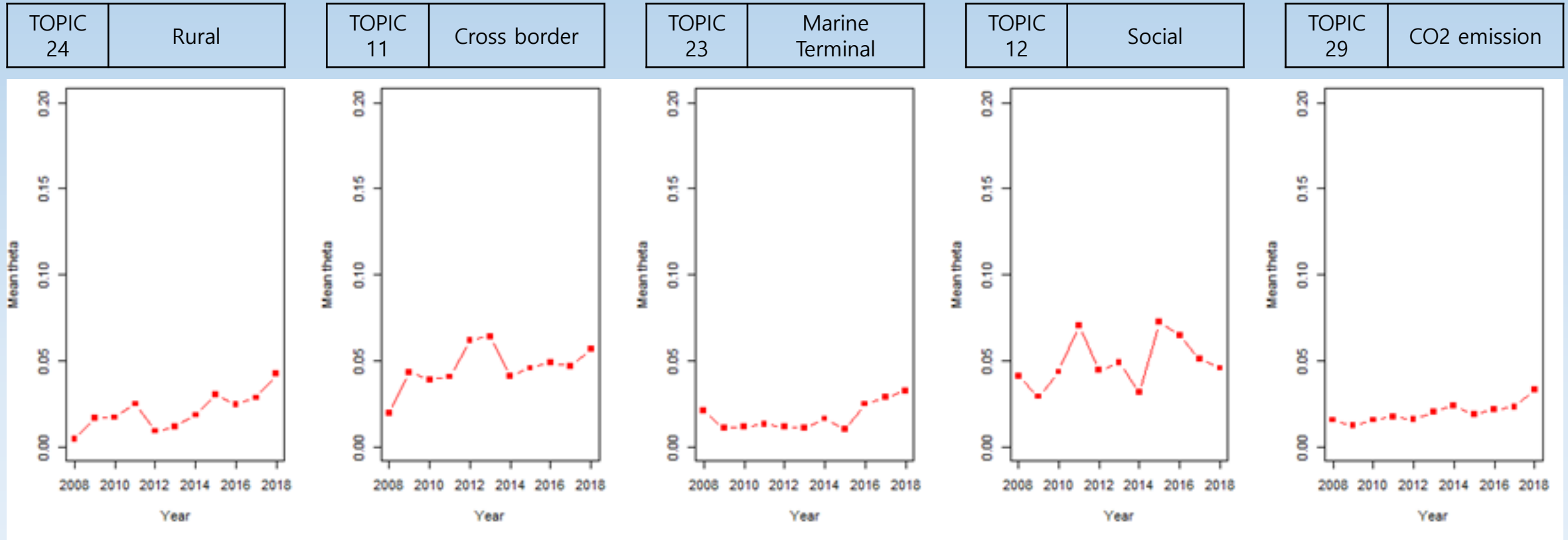
Topic	Name	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Topic 1	Sensing		1		1			1	5	3	4	
Topic 2	Decision Making	3			2	1	1		1	4	2	
Topic 3	Customer	1		2	1	1	1		5		3	
Topic 4	Technology Adaption	3	1	1	3	1	1			1	2	1
Topic 5	Port		1		2	2	1	1	3	2	1	2
Topic 6	Logistic Regulation		1	1	2					4	5	2
Topic 7	marine shipping			1	1	1	1	2	4		3	2
Topic 8	Biomess			2		1	2	3	1	3	2	1
Topics9	Supply Chain	1	1	1	3		1	3	1	1	3	
Topics10	Cold Chain	2					3	3	1		2	3
Topics11	Cross border		1	2	1	2	1		3		4	1
Topics12	social	1	1	2	1	1	1		3	1	4	
Topics13	Communication Technology	1	1	2	1				2	4	3	1
Topics14	Optimization	1	1	1	1			1	7	1	2	
Topics15	integration	1	1		1	3		1	1	3		2



결과 : 물류 주요 토픽

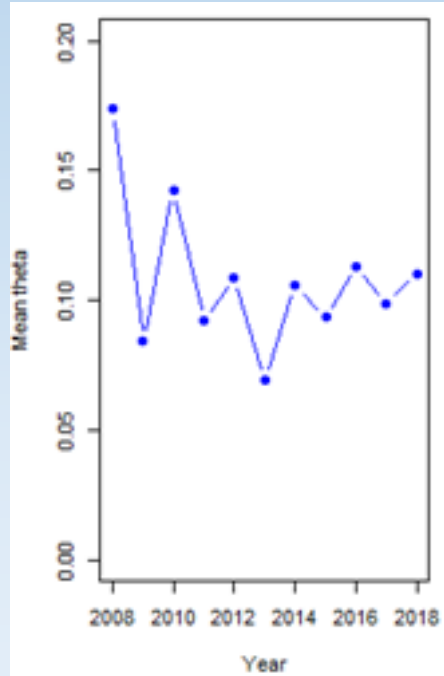
Topic	Name	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Topic 16	Education			1	1			2	1	4	3	1
Topic 17	Bad Behavior			2					2	4	4	3
Topic 18	RFID	1			2	2		2	4	1	2	
Topic 19	DNA Experiment	1		1	2	1	1	2		1	5	
Topic 20	Environment	1		1	2	1	1		4	2	3	
Topic 21	Container Terminal				3	3	2	1	3		3	
Topic 22	Ship Building		1	1		2		1	1	2	2	
Topic 23	Marine Terminal	2						1		2	6	3
Topics24	Rural		1	1	2		1		2	1	5	
Topics25	Network		1	1		1	2	1		2	3	2
Topics26	Green Logistics	1	1		1			1	3		7	
Topics27	Farm SCM	1	3	1	1		2	1		2	1	2
Topics28	Reverse Logistics	1	1	1	2	2	1			3	4	
Topics29	CO2 Emission			1				2	2	5	2	3
Topics30	Harzrd			2	4	1	1	3	1		3	

<Hot Topic>

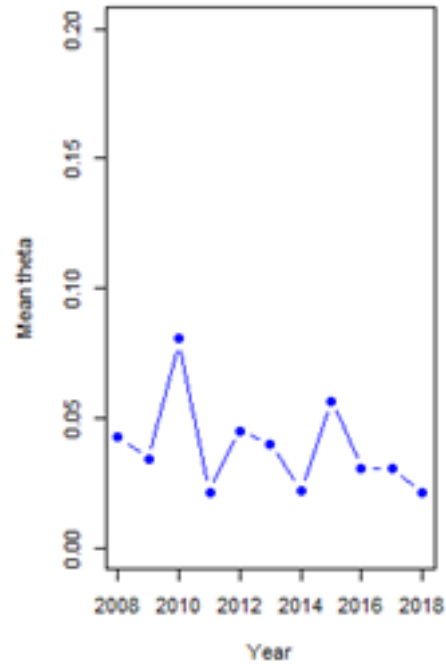


<Cold Topic>

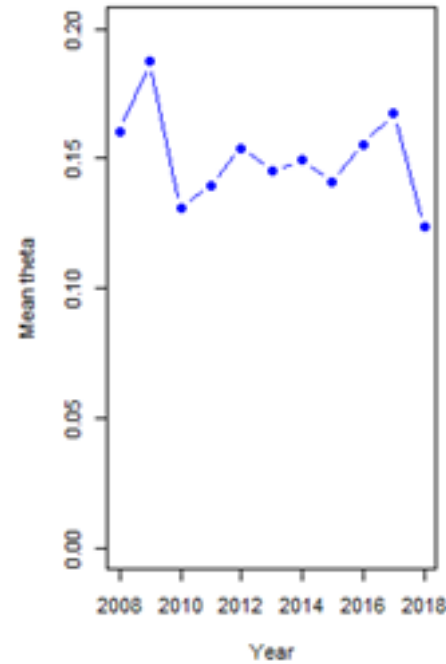
TOPIC 4	Technology adoption
------------	------------------------



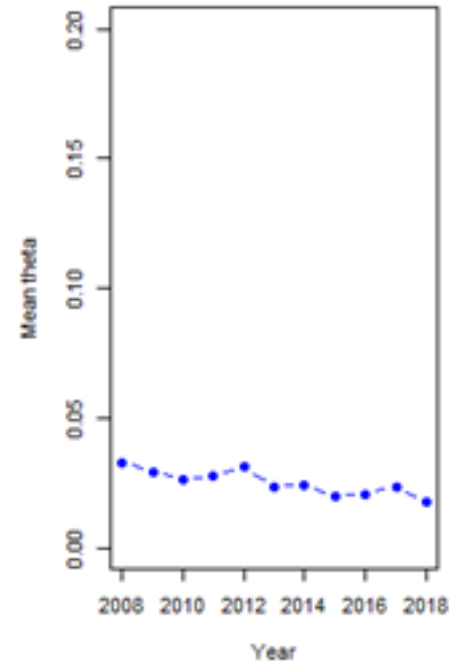
TOPIC 3	Customer
------------	----------



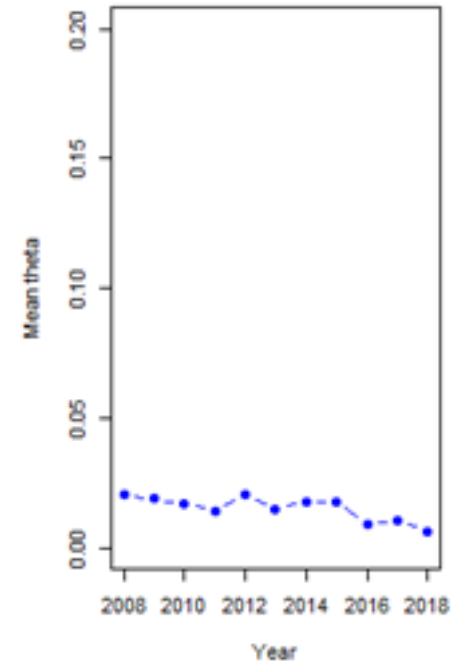
TOPIC 13	Communicatio n Technology
-------------	------------------------------



TOPIC 28	Reverse logistics
-------------	----------------------



TOPIC 18	RFID
-------------	------





트렌드분석

Hot Topic

TOPIC 11 Cross border
TOPIC 12 Social
TOPIC 23 Marine Terminal
TOPIC 24 Rural
TOPIC 29 CO2 emission



환경 문제에 초점을 맞춘 물류 기술

Cold Topic

TOPIC 3 Customer
TOPIC 4 Technology adoption
TOPIC 13 Communication Technology
TOPIC 18 RFID
TOPIC 28 Reverse logistics



통신을 기반으로 한 고객 위주의 물류 기술



연도별 주요 연구주제

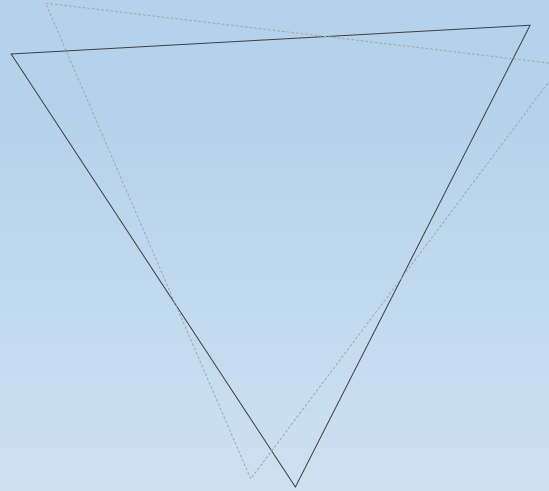
2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
<ul style="list-style-type: none"> • Decision Making • Technology Adaption 	<ul style="list-style-type: none"> • Farm SCM 		<ul style="list-style-type: none"> • Technology Adaption • Supply Chain • Container Terminal 	<ul style="list-style-type: none"> • Integration • Container Terminal 	<ul style="list-style-type: none"> • Cold Chain 	<ul style="list-style-type: none"> • Biomass • Supply Chain • Cold Chain • Hazard 	<ul style="list-style-type: none"> • Customer • Marine Shipping • Optimization 	<ul style="list-style-type: none"> • Decision Making • Logistic Regulation • Communication Technology • Education • Bad Behavior • CO2 Emission 	<ul style="list-style-type: none"> • Sensing • Logistic Regulation • Cross Border • Social • Bad Behavior • DNA Experiment • Marine Terminal • Green Logistics • Reverse Logistics

* (2008~2014 : 3개 논문, 2014~ : 4개 이상 논문)



Conclusion

- ✓ 의학과 관련 논문이 너무 많이 써치 되었다는 문제점.
- ✓ 키워드중심이 아니라 저널 또는 전문잡지 기준으로 데이터 마이닝을 하면 더 좋은 결과가 나왔을 듯 함.



THANKS

