제7회

KOSSDA

데이터 페어

사회과학자가 빅데이터를 만나면 뭘 할까?

2019. 6. 27.

(목) 14:00–17:00

서울대학교
박물관(70동) 강당

KOSSDA 데이터 페어는 서울대 사회과학대학
'미래 기초학문분야 기반조성사업'의 지원을 받아 개최되는 행사입니다.

제7회
KOSSDA
데이터 페어

사회과학자가 빅데이터를 만나면 뭘 할까?

2019. 6. 27. (목) 14:00–17:00
서울대학교 박물관(70동) 강당

프로그램

인사말 14:00–14:10
이봉주 학장 (서울대 사회과학대학)
박수진 소장 (서울대 아시아연구소)

1부: 빅데이터 활용 방법론 강의 14:10–14:50

사회과학 분야 빅데이터 연구방법론
김기훈 교수 (서울대 사회학과, 사이람 대표)

휴식 14:50–15:00

2부: 빅데이터 활용 사례

모빌리티 데이터로 바라보는 사회 15:00–15:30
김정민 연구원 (카카오 모빌리티)

정치연구와 빅데이터 15:30–16:00
한규섭 교수, 노선혜 연구원 (서울대 언론정보학과)

문학권력에 대한 세 가지 데이터 분석 16:00–16:30
이원재 교수 (카이스트 문화기술대학원)

질의응답 16:30–17:00

SNUAC
Seoul National University Asia Center
서울대학교 아시아연구소

제7회

KOSSDA

데이터 페어

목차

제7회

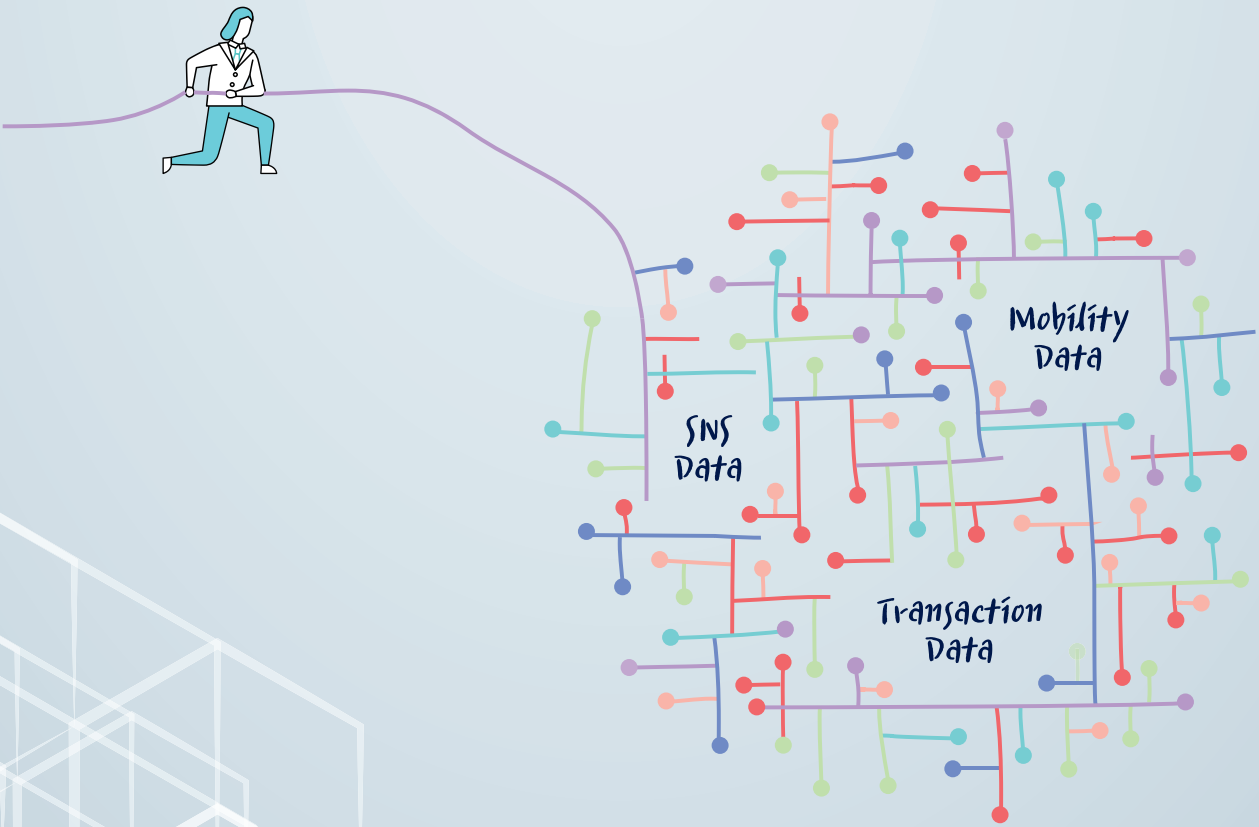
KOSSDA

데이터 페어

1부: 빅데이터 활용 방법론 강의

사회과학 분야 빅데이터 연구방법론

김기훈 교수 (서울대 사회학과, 사이람 대표)



제7회 KOSSDA 데이터 페어

사회과학 분야 빅데이터 연구방법론

2019-6-27

김 기훈
㈜사이람 대표이사
서울대 사회학과

Table of Contents

Big Data 의 본질

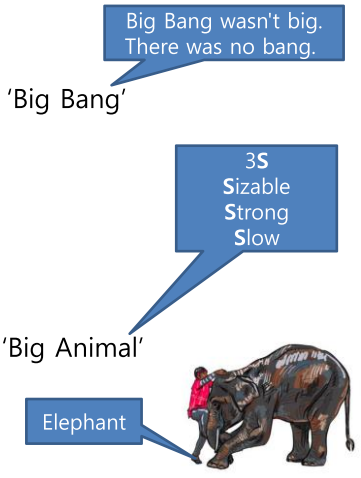
‘Big Data’ ?

‘Big Data’ as a misnomer(誤稱)?

IT Features of “Big Data” : 3V
1. Volume
2. Variety
3. Velocity

essence/identity(種別的 本質) of ‘Big Data ?

➔ “digital footprint(trace) data”



Big Data

highly detailed "digital footprint/trace data"
automatically captured/generated using IT systems / sensors

Maciej Beresewicz et. Al.(2018), An overview of methods for treating selectivity in big data sources

Digital Footprint

"A digital footprint is the data trail left by interactions in a digital environment; including the use of TV, mobile phone, the World Wide Web, the internet and other connected devices and sensors.

Digital Footprints provide data on what has been performed in the digital environment (e.g. what you clicked on, searched for, liked, where you went, your location, your IP address, what you said, what was said about you)"

- Wikipedia -

"cyber shadow"
"electronic footprint"
"digital shadow"

Big Data 의 과학적 성격: 이론의 종말?

“데이터로 하여금 말하게 하라”
-이론(Theory)의 종말?

WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

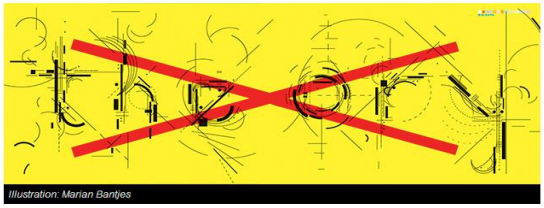


Illustration: Marian Bantjes

와이어드(Wired), 2008년 6월 23일자
커버스토리 ' 이론의 종말'
크리스 앤더슨'

“데이터로 하여금 말하게 하라” -이론(Theory)의 종말?

- 이제 세계는 엄청나게 많은 데이터와 응용 수학이 고려할 수 있는 다른 모든 도구들을 대체한다. 언어학에서부터 사회학에 이르기까지, 모든 인간 행동에 대한 이론은 버려라. 분류학, 존재론, 심리학도 잊어라. 어느 누가 사람들이 왜 무엇을 하는지 알까? 중요한 것은 사람들은 그것을(무엇인가를) 하고 있고, 우리는 사람들의 행위를 전례 없는 정확도로 추적하고 측정할 수 있다는 것이다. 충분한 양의 데이터와 함께라면, 숫자는 자기 생각을 말한다

“데이터로 하여금 말하게 하라” -이론(Theory)의 종말?

- 이렇게 큰 규모의 컴퓨터 사용을 배우는 것은 매우 도전적일지 모른다. 그러나 기회는 대단할 것이다. 엄청난 양의 데이터, 그리고 이러한 숫자를 다룰 수 있는 통계적 도구는 세계를 이해하기 위한 완전히 새로운 방식을 제공한다. 상관성(Correlation)이 인과성(Causation)을 대체하고 과학은 일관적인 모델, 통합된 이론, 혹은 설명 기제 없이도 진보할 것이다.

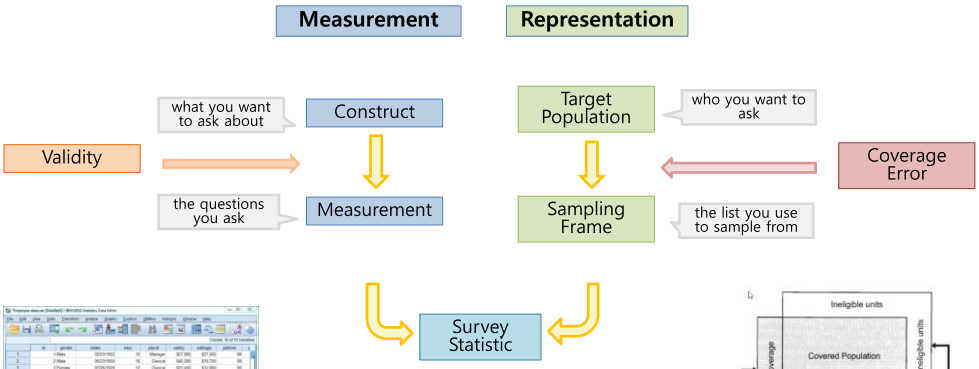
Big Data 의 Survey Error

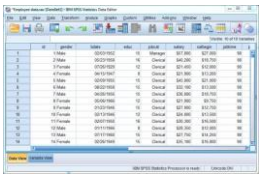
Analytic features of Big Data

- not designed for (statistical) analysis ("organic data", "administrative data")
- non-probabilistic character
- mediation of digital platform
- unstructured / highly detailed

Maciej Beresewicz et. Al.(2018), An overview of methods for treating selectivity in big data sources

Survey Error





Groves et. al., Survey Methodology

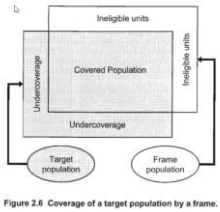


Figure 2.6 Coverage of a target population by a frame.

Big Data 의 Survey Error

	cons	pros
representation /coverage of population	non-probabilistic sample data → coverage error	coverage of 'entire' population
validity /measurement	mediation of digital platform → measurement error	unobtrusive / nonreactive (cf. mediation of questionnaire) "설문조사는 그 자체가 하나의 외부적 요인으로서, 우리가 묘사하고자 하는 사회적 상황을 침범(intrude)함으로써, 어떤 태도를 측정하고 동시에 형성"
unit & variable of interest	• unit = individual → unit error • variable = attribute → proxy variable ➔ "unstructured"	• unit = event (→ social action/interaction) • variable = (structural) position ➔ "detailed"

Big Data 의 Coverage :

연구의 Target population(A)과 해당 Big Data 를 생성한 특정 digital platform(B)과의 관계

관계의 종류	A의 예 (B: Twitter)	Big data 분석의 의미	비고
A = B	"Twitter"	complete enumeration survey	단, 해당 플랫폼의 데이터 수집 방법에 의존적 기술통계, 기계학습에 의거한 예측모델링. (한정된 경계 내의)
A ⊃ B	"online social network"	case study	소규모(개인, 소집단) → 초대규모(전체 디지털 플랫폼) (참여)관찰 → (객관적)기록(log) 수작업 → 자동화 심층(in-depth) → 상세(detail) 질적 분석 → 양적 분석 기술적 → 기술적(+탐색적/확증적)
A ? B	"한국 국민"	non-probabilistic (sample) survey	with proper method of ***calibration***/adjustment (using auxiliary variable) → "big data survey"(?)

Big Data 의 사회과학적 의의

MIT Technology Review

이제 우리는 과거에는 절대 불가능했던 정도의 초정밀현미경으로 인간의 사회적 행위를 들여다 볼 수 있게 되었다. 나아가 수백만 명을 상대로 실험을 하는 것조차 가능하다.

Cameron Marlow

I am a research scientist and "in-house sociologist" at Facebook. My research focuses on various aspects of online communities including the diffusion of information across online social networks, access to information and social capital, and the incentives that impact social media production.

I received my Ph.D. at the MIT Media Laboratory on the topic of media contagion and weblogs. You might also be familiar with my project Blogdex which tracks (used to track?) diffusion in the weblog community. The results of th work and the MIT Weblog Survey can be found in my Ph.D. Thesis.

After finishing my Ph.D. I joined Yahoo Research Berkeley where I ran the Social Motives group studying the social incentives in emerging applications such as Flickr, del.icio.us, and Last.fm. I joined Yahoo! Research and continued my work on various topics related to social search, community evolution, and network topology.

In addition to my research, I currently manage the Data Science team at Facebook, and am engaged in trying to understand how the data platforms of the future will change the roles of researchers in the enterprise and academia.

Facebook이 풀어야만 하는 최대의 과제는 사회과학이 풀어야 하는 과제와 똑같은 것이다.

email: cam@alum.mit.edu
aim: cameronfactor
flickr
facebook
linkedin

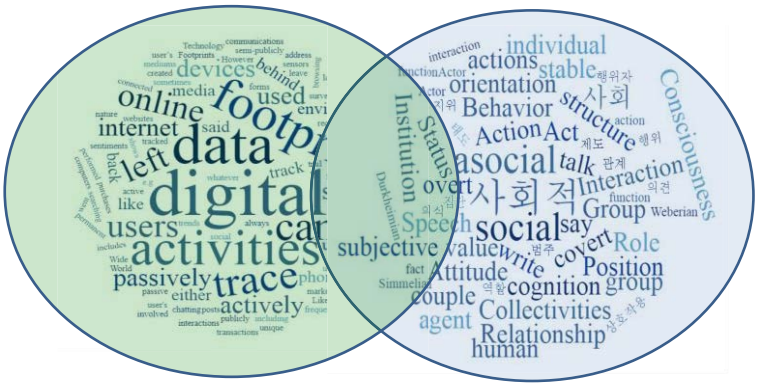
Facebook의 Data Science Team은 사회학자가 진두지휘 하고 있음.

Recent Papers

- Structural diversity in social contagion
- The Role of Social Networks in Information Diffusion
- Why Most Facebook Users Get More Than They Give: The Effect of Facebook "Power Users" on Everybody Else
- A 61-million-person experiment in social influence and political mobilization
- Social capital on Facebook: Differentiating uses and users

A spectre is haunting social sciences – the spectre of big data.

social big data



big data

digital footprint data

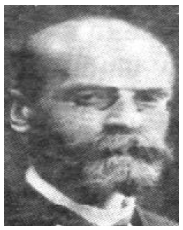
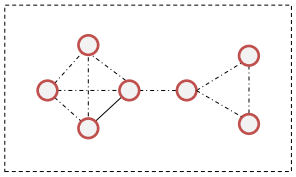
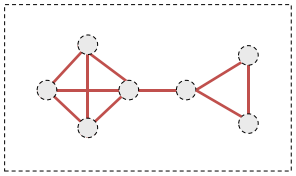
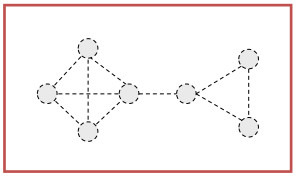
something ***social***

- Weberian social action
- Simmelian sociation
- Durkheimian social facts

소셜 빅데이터의 사회과학 방법론적 성격

	전통적 사회조사	소셜 빅데이터
Social	Proxy social (attributes of actors)	***N***on-proxy social • social action/interaction • speech act
Obtrusiveness	Obtrusive (questionnaire/interview)	***N***on-obtrusive (observation)
Sampling	Sampling	***N***on-sampling (complete enumeration)

the "Social"

Max Weber	Georg Simmel	Emile Durkheim
(1864-1920)	(1858-1918)	(1858-1917)
		
Social Action	Sociation	Social Fact
		
consideration/orientation meaning	Geometry position	sui generis, emergent properties

Rediscovery of hidden dimension in Big Data : "social action" as data

"The devil is in the detail"

Big Data = detailed transaction data left in a digital environment

"an action is '**social**' if the acting individual takes account of the behavior of others and is thereby oriented in its course".

– Max Weber

'sleeping beauties' in detail fields of big data
: social action, *Verstehen* (Weber)
: sociation (Simmel)

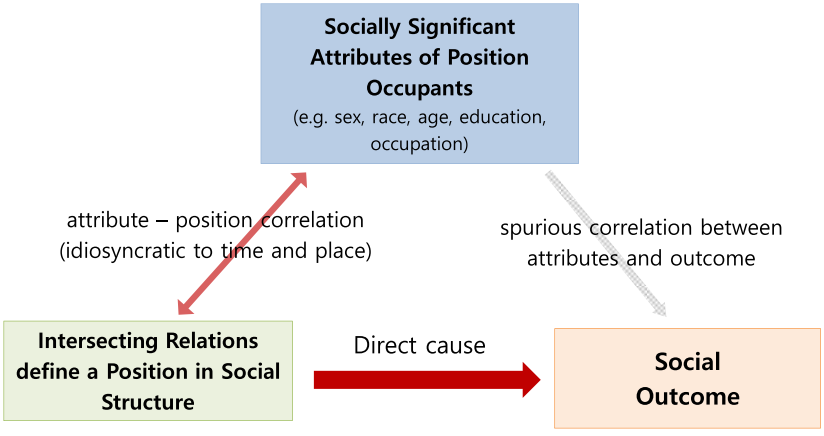
Max Weber
(1864-1920)

Social action

Georg Simmel
(1858-1918)

Sociation

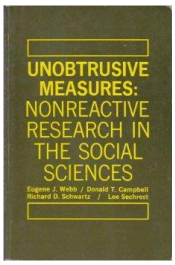
Rediscovery of hidden dimension in Big Data : "social action" as data

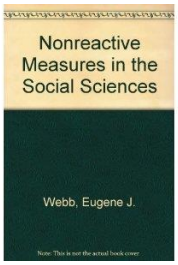


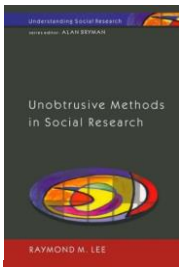
Ronald S. Burt. (1991). Structure : Reference Manual, p. 13

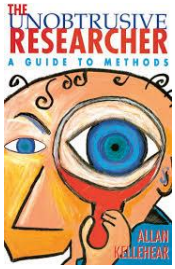
unobtrusive/nonreactive data (무반응/불개입 데이터)

"오늘날 사회과학 연구조사는 인터뷰 혹은 설문조사를 기초로 한다. 우리는 이런 오류가 있을 수 있는 단일한 방법에 지나치게 의존하는 것을 아쉽게 생각한다.
인터뷰와 설문조사는 그 자체가 하나의 외부적 요인으로서, 우리가 묘사하고자 하는 사회적 상황을 침범(intrude)함으로써, 어떤 태도를 측정하고 동시에 형성한다. 또한 그러한 방법들은 응답에 협조적이고, 접근가능한 사람에게만 사용될 수 있을 뿐만 아니라, 조사 토픽과 관계없는 개별적 차이의 차원에 의해 형성된 응답일 수 있다는 한계점을 가진다.
그러나, 가장 주된 반대(objection)이유는 이런 방법들이 단독으로 사용되어서는 안 된다는 것이다."



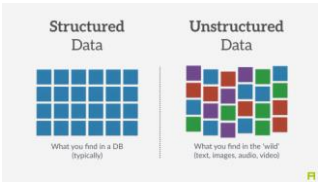


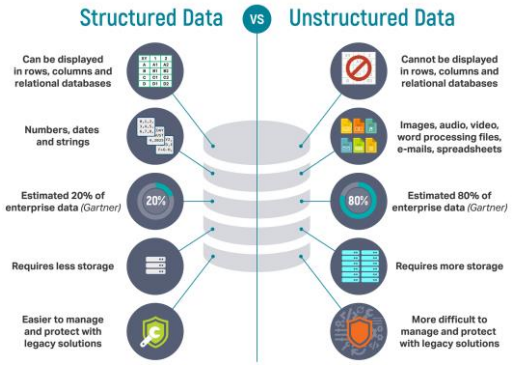




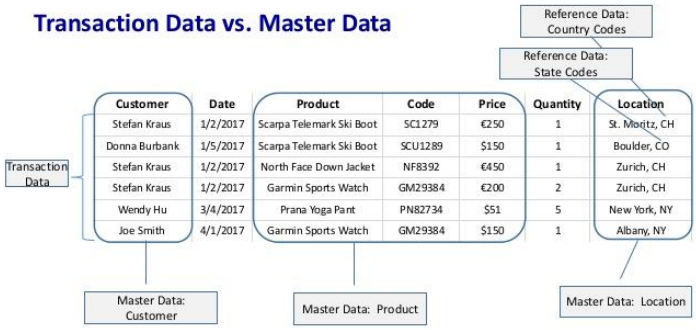
Big Data 의 종류

Structured data vs. Unstructured data





Master Data vs. Transaction Data



Master data represents the business objects that contain the most valuable, agreed upon information shared across an organization.[1] It can cover relatively static reference data, transactional, unstructured, analytical, hierarchical and metadata.[2] It is the primary focus of the information technology (IT) discipline of master data management (MDM).

Master data is usually non-transactional in nature, but in some cases gray areas exist where transactional processes and operations may be considered master data by an organization. For example, master data may contain information about customers, products, employees, materials, suppliers, and vendors. Though rare, if that information is only contained within transactional data such as orders and receipts and is not housed separately, it may be considered master data.[3]

Transaction data is data describing an event (the change as a result of a transaction) and is usually described with verbs. Transaction data always has a time dimension, a numerical value and refers to one or more objects (i.e. the reference data).

Typical transactions are:

- Financial: orders, invoices, payments
- Work: plans, activity records
- Logistics: deliveries, storage records, travel records, etc.

Big Data 의 3 종류 (3T)

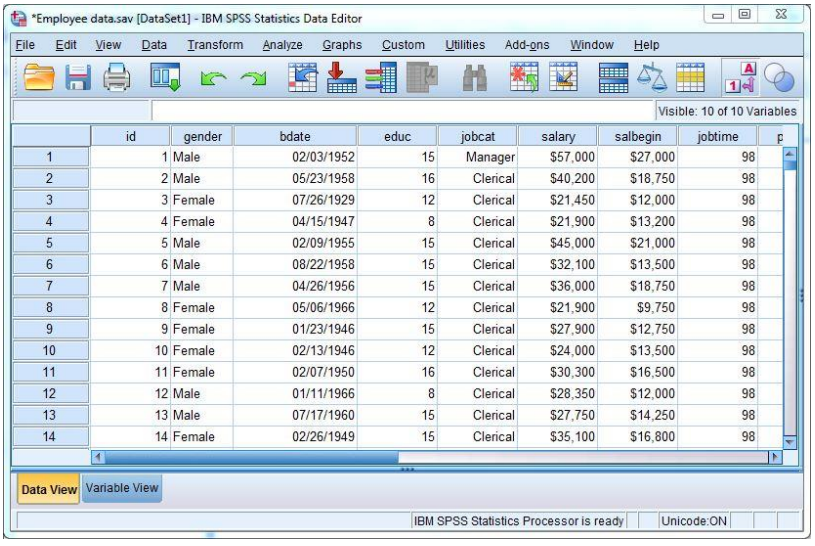
빅데이터 중 팔할은 **텍스트** 데이터 or **트랜젝션** 데이터
(텍스트 데이터 반, 트랜젝션 데이터 반)

1. **T**raditional Data
   (개체-속성 데이터)

2. **T**ext Data

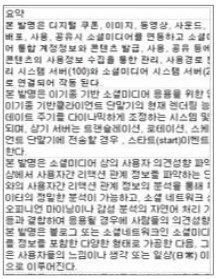
3. **T**ransaction Data

(1) Traditional Data: 개체-속성 데이터 (master data)



	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime
1	1	Male	02/03/1952	15	Manager	$57,000	$27,000	98
2	2	Male	05/23/1958	16	Clerical	$40,200	$18,750	98
3	3	Female	07/26/1929	12	Clerical	$21,450	$12,000	98
4	4	Female	04/15/1947	8	Clerical	$21,900	$13,200	98
5	5	Male	02/09/1955	15	Clerical	$45,000	$21,000	98
6	6	Male	08/22/1958	15	Clerical	$32,100	$13,500	98
7	7	Male	04/26/1956	15	Clerical	$36,000	$18,750	98
8	8	Female	05/06/1966	12	Clerical	$21,900	$9,750	98
9	9	Female	01/23/1946	15	Clerical	$27,900	$12,750	98
10	10	Female	02/13/1946	12	Clerical	$24,000	$13,500	98
11	11	Female	02/07/1950	16	Clerical	$30,300	$16,500	98
12	12	Male	01/11/1966	8	Clerical	$28,350	$12,000	98
13	13	Male	07/17/1960	15	Clerical	$27,750	$14,250	98
14	14	Female	02/26/1949	15	Clerical	$35,100	$16,800	98

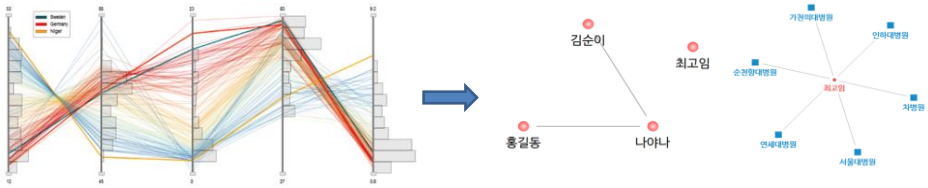
(2) Text 데이터


요약
본 발명은 디지털 쿠폰, 이미지, 동영상, 사운드, 버튼, 사용, 공유서 소셜미디어를 연동하고 소셜어 통합 계정정보와 콘텐츠 발급, 사용, 공유 등에 콘텐츠의 사용정보 수집을 통한 관리, 사용경로 리 시스템 서버(100)와 소셜미디어 시스템 서버(로 연결되어 작동 된다.
본 발명은 이기종 기반 소셜미디어 응용을 위한 이기종 기반플라이언트 단말기의 형재 컨더팀 데이터등 주기를 다이나믹하게 조정하는 시스템 및 되며, 상기 서버는 트랜슬레이션, 코테이션, 스케언트 단말기에 전송할 경우 , 스타트(start)이펜트한다.
본 발명은 소셜미디어 상의 사용자 의견성향 파악 상에서 사용자간 라먹산 관계 정보를 파악하는 다와의 사용자간 라먹산 관계 정보의 분석을 통해 이타의 정량한 분석이 가능하고, 소셜 네트워크 오피니언 마이닝이나 감성 분석의 자연어 처리 기등과 결합하여 응용될 경우에 사람들의 의견성향본 발명은 불로그 또는 소셜네트워크의 소셜미디을 정보를 포함한 다양한 형태로 가공한 다음, 그은 사용자들의 느낌이나 생각 또는 일상(日常) 이으로 이루어진다.

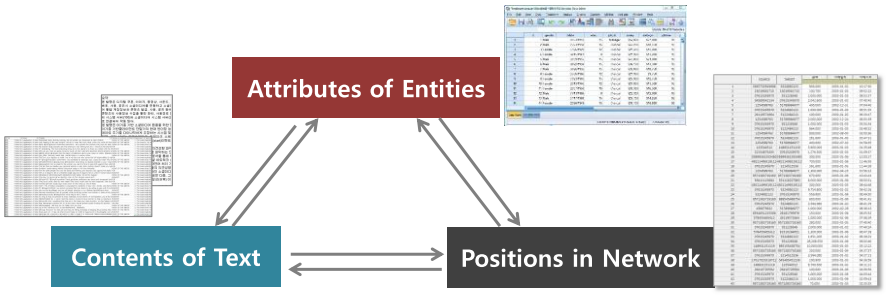
Created Time	Writer	APP	HashTags	RT Coun	Full Text	Type
2018-12-07 19:00:	sophiakaito	Twitter for i	None	0	@hi_romi974 절다... 언니는 추워서.. 절대 스커트 이 날씨에 못입어요....	Reply
2018-12-07 19:00:	seojohn_nctzen	Twitter for i	NCT	2302	RT @limeharu: 181207 SMile 에스엠엔터테인먼트 사회공헌 추운 날씨에도 #NCT 봉사활동을 가	RT
2018-12-07 19:00:	yuiri19_tao	Twitter for i	채연Dream	4040	RT @official_izone: FNS가요제에 조청해주셔서 감사합니다!!! 정말 많이 맬렸던 무대였습니다...	RT
2018-12-07 19:00:	HS_OB_	Twitter for i	박형식,park	1	RT @sik_shine: 181207 뮤지컬 <엘리자벳> 퇴근길 날씨가 너무 추워지니까 놀란 쥬금님 쥬금님	RT
2018-12-07 19:00:	wsj053161	Twitter for i	None	495	RT @ASeoyeon777: 광주터미널에서 '이재명 제명' 첫 시위광주시당을 거쳐 광주고속버스터미널	RT
2018-12-07 19:00:	sturdy_feather	Twitter for i	None	7311	RT @BTS_twt: 안녕하세요 진인데요 오늘 날씨가 다시 추워진듯해요 다들 핫팩 챙겨가지고 다니	RT
2018-12-07 19:00:	wjlover_	Twitter for i	None	528	RT @Babybird_lo: 영하의 추운 날씨에 오전부터 열일한 우리원을 위해 고기 반찬 가득가득 맛있	RT
2018-12-07 19:00:	mongmongxg	Twitter for i	None	9464	RT @iseulgiyou: 우연히 산 중고책. 세월이 흘러 테두리가 누렇게 바랜 첫 장에 23년 전 누군가의	RT
2018-12-07 19:00:	lssgg	Twitter for i	None	242	RT @hjyang0404: <따뜻한 국가인재원> 내 눈에서 눈물나게 한 사람들. 취임 100일이라고 직원	RT
2018-12-07 19:00:	bamjswz	Twitter for i	None	50	RT @daystar_990529: 지푼이 추운 날씨에 이른 아침부터 수고했어 체리딸기왕자님 https://t.co	RT
2018-12-07 19:00:	Mabi_Seanan	twittbot.net	None	0	내일 날씨는 어떨까?	Tweet
2018-12-07 19:00:	jumrye_0621	Twitter for i	None	9464	RT @iseulgiyou: 우연히 산 중고책. 세월이 흘러 테두리가 누렇게 바랜 첫 장에 23년 전 누군가의	RT
2018-12-07 19:00:	dvracham	Twitter for i	None	528	RT @Babybird_lo: 영하의 추운 날씨에 오전부터 열일한 우리원을 위해 고기 반찬 가득가득 맛있	RT
2018-12-07 19:00:	red_btsglasses	Twitter for i	9일째,낙태	29	RT @safe_abortion_: 헌법재판소의 낙태죄 위헌 판결을 축구하는 헌법재판소 앞 1인시위 #9일째	RT
2018-12-07 19:00:	movement_0529	Twitter for i	None	50	RT @daystar_990529: 지푼이 추운 날씨에 이른 아침부터 수고했어 체리딸기왕자님 https://t.co	RT
2018-12-07 19:00:	Bboing___39	Twitter for i	None	0	진짜 오늘은 그냥 나가면 죽을 수도 있겟구나를 생각하게 하는 날씨	Tweet
2018-12-07 19:00:	tahongthuongg	Twitter for i	재현,JAEHY	445	RT @mydarlinghyun: 181123 날씨가 많이 춥네요 옷 따뜻하게 입으세요 #재현 #JAEHYUN #N	RT

(3) Transaction 데이터

순번	보험종목	회사	사고발생일자	지급일자	병원이름	사유	피보험자주민번호	피보험자	계약자주민번호	계약자	수익자	설계사
1	생명보험	삼성생명	1998-12-30	1998-12-31	서울대병원	OO	1985xxxx-2xxxxxx	홍길동	1946xxxx-1xxxxxx	나야나	나야나	최고임
2	생명보험	현대해상	1998-12-31	1999-01-01	연세대병원	XX	1985xxxx-2xxxxxx	홍길동	1946xxxx-1xxxxxx	나야나	나야나	최고임
3	자동차보험	MG	1999-01-01	1999-01-02	순천향대병원	OO	1946xxxx-1xxxxxx	김순이	1985xxxx-2xxxxxx	나야나	나야나	최고임
4	상해보험	메리츠	1999-01-02	1999-01-02	서울대병원	DD	1946xxxx-1xxxxxx	김순이	1985xxxx-2xxxxxx	나야나	나야나	최고임
5	생명보험	삼성생명	2000-10-20	2000-10-20	연세대병원	AA	1985xxxx-2xxxxxx	홍길동	1946xxxx-1xxxxxx	김순이	나야나	최고임
6	생명보험	현대해상	2000-10-20	2000-10-20	가천의대병원	BB	1946xxxx-1xxxxxx	김순이	1985xxxx-2xxxxxx	홍길동	홍길동	최고임
7	자동차보험	MG	2000-10-20	2000-10-20	서울대병원	CC	1946xxxx-1xxxxxx	김순이	1985xxxx-2xxxxxx	홍길동	홍길동	최고임
8	생명보험	삼성생명	2003-04-05	2003-04-05	연세대병원	EE	1985xxxx-2xxxxxx	홍길동	1946xxxx-1xxxxxx	김순이	김순이	최고임
9	생명보험	현대해상	2003-04-05	2003-04-05	인하대병원	FF	1985xxxx-2xxxxxx	홍길동	1946xxxx-1xxxxxx	김순이	김순이	최고임
10	자동차보험	MG	2003-04-05	2003-04-05	차병원	OO	1946xxxx-1xxxxxx	김순이	1985xxxx-2xxxxxx	홍길동	홍길동	최고임
11	상해보험	메리츠	2003-04-05	2003-04-05	연세대병원	OO	1946xxxx-1xxxxxx	김순이	1985xxxx-2xxxxxx	홍길동	나야나	최고임
12	생명보험	삼성생명	2010-02-03	2010-02-03	순천향대병원	XX	1985xxxx-2xxxxxx	홍길동	1946xxxx-1xxxxxx	김순이	나야나	최고임
13	생명보험	삼성생명	2019-02-19	2019-02-19	서울대병원	XX	1985xxxx-2xxxxxx	홍길동	1946xxxx-1xxxxxx	김순이	김순이	최고임
14	상해보험	메리츠	2019-02-19	2019-02-19	연세대병원	XX	1946xxxx-1xxxxxx	김순이	1985xxxx-2xxxxxx	홍길동	홍길동	최고임



	Traditional	Text	Transaction
Object	Entity	Contents	Relationship
Focus	Association among Attributes of Entities	Meanings contained in the Text	Structural Positions in Network
Analytic Methodology	Multivariate Statistics Machine Learning	NLP Text Mining	Network Analysis Graph Mining



Big Data 의 수집

빅데이터의 일반적 획득 방법

데이터 획득 방법	장점	단점
데이터 제공로부터 Data File 획득 (TXT, CSV 또는 엑셀 파일 형태)	직접 수집할 필요가 없음	• 정보가 변경되었을 때 실시간으로 전달할 방법이 없음 • 시간 비용 소모가 큼
Database 접근권한을 얻어 데이터 수집	정형화된 데이터 수집이 용이함	보안의 위험성(현관문 비밀번호를 알려준 것과 동일)
데이터 제공 업체의 API를 통한 데이터 수집	보안과 실시간 전달의 문제점을 해결	데이터 수집 제한이 있음
Web Crawling 통한 데이터 수집	거의 대부분의 웹 상의 데이터를 수집 가능	• 웹 상의 소스코드를 분석하고 parsing 하는 것이 어려움 • 저작권 위반의 위험

WEB crawling

news.naver.com

HTML source

문자열 데이터를 변수에 넣고

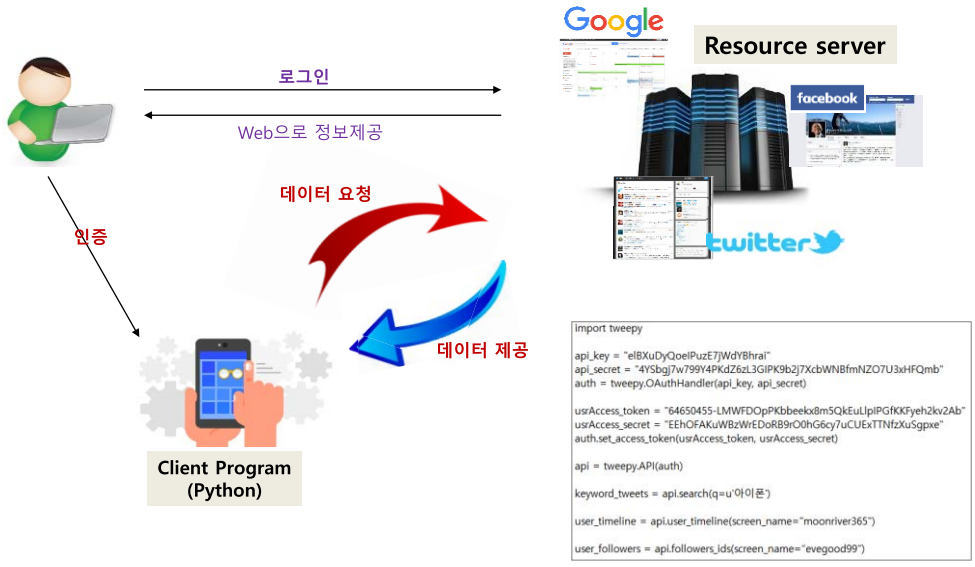
HTML 페이지를 가져옴

python

Parsing(특정 문자 추출)을 통해 원하는 문자 정보를 얻는 것

```
urlOpen = urllib2.urlopen("https://www.clien.net/service/board/sold")
html = urlOpen.read()

fIndex = 0
titleFirst = '<span data-role="list-title-text  title="'
titleLast = '">'

while True:
    try:
        titleFirstIndex = html.index(titleFirst,  fIndex)+len(titleFirst)
    except:
        break
    titleLastIndex = html.index(titleLast,  titleFirstIndex)
    print html[titleFirstIndex : titleLastIndex]
    fIndex = titleLastIndex
```

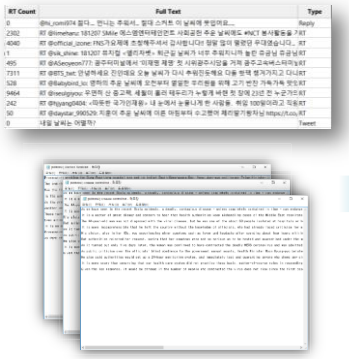
API 를 이용한 데이터 획득 (cf. Web crawling)



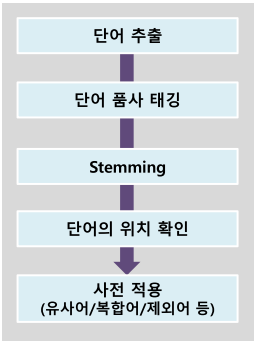

```
import tweepy

api_key = "elBXuDyQoelPuzE7jWdYBhrai"
api_secret = "4YSbgj7w799Y4PKdZ6zL3GIPK9b2j7XcbWNBfmNZO7U3xHFQmb"
auth = tweepy.OAuthHandler(api_key, api_secret)

usrAccess_token = "64650455-LMWFDOpPKbbeekx8m5QkEuLlpIPGfKKFyeh2kv2Ab"
usrAccess_secret = "EEhOFAKuWBzWrEDoRB9rO0hG6cy7uCUExTTNfzXuSgpxe"
auth.set_access_token(usrAccess_token, usrAccess_secret)

api = tweepy.API(auth)

keyword_tweets = api.search(q=u'아이폰')

user_timeline = api.user_timeline(screen_name="moonriver365")

user_followers = api.followers_ids(screen_name="evegood99")
```

Big Data 의 전처리

1) 텍스트 데이터의 전처리: 자연어 처리



비정형 텍스트 데이터
(뉴스기사, SNS, 논문, 책 등)



데이터 전처리

1) 텍스트 데이터의 전처리: 자연어 처리

Word Table

	Part of Speech(POS)	Word length
국민	"Common Noun"	2
국방력	"Common Noun"	3
국정	"Common Noun"	2
군림	"Common Noun"	2
권력	"Common Noun"	2
권위	"Common Noun"	2
기관	"Common Noun"	2
기록	"Common Noun"	2
기회	"Common Noun"	2
긴장	"Common Noun"	2
길	"Common Noun"	1
끝	"Common Noun"	1

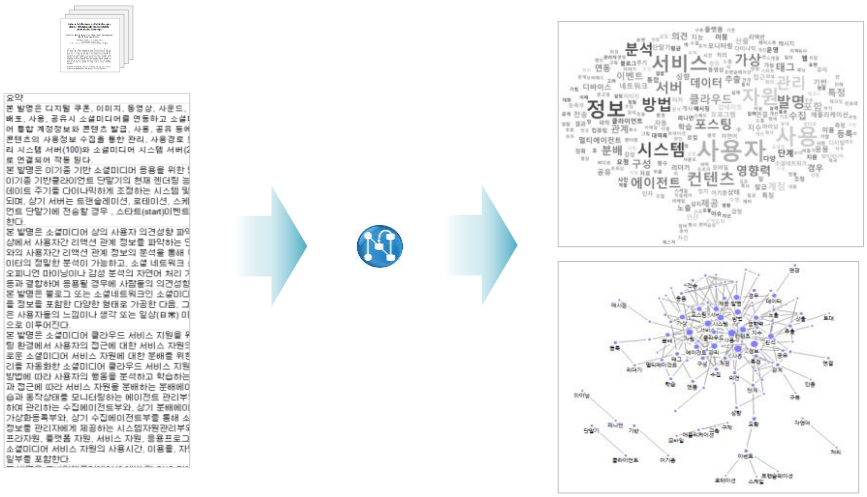
추출된 단어, 품사, 글자수

Word Occurrence Table

단어	등장한 문장	단어 등장 순서
해결	41	315
추구	6	62
정부	69	467
평양	45	329
기회	16	146
경쟁	12	115
어깨	31	257
계층	66	448
이웃	90	589
볼행	74	492
일	85	566
기관	38	296
마련	52	379

단어가 문장/문단/문서에서
등장한 위치

1) 텍스트 데이터의 전처리: Semantic network modeling



1) 텍스트 데이터의 전처리: Semantic network modeling

키워드 네트워크 구성

- 텍스트에서 추출한 단어들로 네트워크를 구성하는 방법은 두 가지로 구분할 수 있음
- 등장순서에 기반한 방법과 동시출현에 기반한 방법이 있음

Window Size vs Proximity

<예시>

Text

동대문 길가에서 방망이를 깎아 파는 노인이 있었다. 값을 굉장히 비싸게 부르는 것 같았다. 대단히 무뚝뚝한 노인이었다.

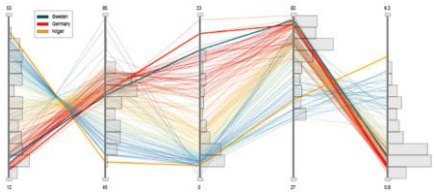
명사만 추출

동대문 | 길가 | 방망이 | 노인 | 값 | 노인

	등장순서(Sequence/Distance)	동시출현 정도(Proximity)
네트워크	동대문 — 길가 — 방망이 노인 — 값 — 노인	동대문, 길가, 값, 노인, 방망이
특징	단어 간 인접성(거리)에 기반	동시출현에 기반
	가깝게 등장한 단어끼리만 링크가 형성됨	문장/문단/문서 내 등장한 모든 단어 사이에 링크가 형성됨
	단어 간 거리가 네트워크에 반영됨	단어 간 거리를 의미하지 않음
	단어 간 거리를 조정하여 네트워크를 형성할 수 있음	동시출현 유사성 측정 방법을 선택할 수 있음

2) Transaction data 의 전처리: network modeling

순번	보험종목	회사	사고발생일자	지급일자	병원이름	사유	피보험자주민번호	피보험자	계약자주민번호	계약자	수익자	설계사
1	생명보험	삼성생명	1998-12-30	1998-12-31	서울대병원	OO	1985xxxx-2xxxxxx	홍길동	1946xxxx-1xxxxxx	나야나	나야나	최고임
2	생명보험	현대해상	1998-12-31	1999-01-01	연세대병원	XX	1985xxxx-2xxxxxx	홍길동	1946xxxx-1xxxxxx	나야나	나야나	최고임
3	자동차보험	MG	1999-01-01	1999-01-02	순천향대병원	OO	1946xxxx-1xxxxxx	김순이	1985xxxx-2xxxxxx	나야나	나야나	최고임
4	상해보험	메리츠	1999-01-02	1999-01-02	서울대병원	DD	1946xxxx-1xxxxxx	김순이	1985xxxx-2xxxxxx	나야나	나야나	최고임
5	생명보험	삼성생명	2000-10-20	2000-10-20	연세대병원	AA	1985xxxx-2xxxxxx	홍길동	1946xxxx-1xxxxxx	김순이	나야나	최고임
6	생명보험	현대해상	2000-10-20	2000-10-20	가천의대병원	BB	1946xxxx-1xxxxxx	김순이	1985xxxx-2xxxxxx	홍길동	홍길동	최고임
7	자동차보험	MG	2000-10-20	2000-10-20	서울대병원	CC	1946xxxx-1xxxxxx	김순이	1985xxxx-2xxxxxx	홍길동	홍길동	최고임
8	생명보험	삼성생명	2003-04-05	2003-04-05	연세대병원	EE	1985xxxx-2xxxxxx	홍길동	1946xxxx-1xxxxxx	김순이	김순이	최고임
9	생명보험	현대해상	2003-04-05	2003-04-05	인하대병원	FF	1985xxxx-2xxxxxx	홍길동	1946xxxx-1xxxxxx	김순이	김순이	최고임
10	자동차보험	MG	2003-04-05	2003-04-05	차병원	OO	1946xxxx-1xxxxxx	김순이	1985xxxx-2xxxxxx	홍길동	홍길동	최고임
11	상해보험	메리츠	2003-04-05	2003-04-05	연세대병원	OO	1946xxxx-1xxxxxx	김순이	1985xxxx-2xxxxxx	홍길동	나야나	최고임
12	생명보험	삼성생명	2010-02-03	2010-02-03	순천향대병원	XX	1985xxxx-2xxxxxx	홍길동	1946xxxx-1xxxxxx	김순이	나야나	최고임
13	생명보험	삼성생명	2019-02-19	2019-02-19	서울대병원	XX	1985xxxx-2xxxxxx	홍길동	1946xxxx-1xxxxxx	김순이	김순이	최고임
14	상해보험	메리츠	2019-02-19	2019-02-19	연세대병원	XX	1946xxxx-1xxxxxx	김순이	1985xxxx-2xxxxxx	홍길동	홍길동	최고임



2) Transaction data 의 전처리: network modeling

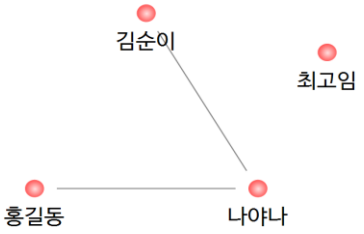
순번	보험종목	회사	사고발생일자	지급일자	병원이름	사유	피보험자주민번호	피보험자	계약자주민번호	계약자	수익자	설계사
1	생명보험	삼성생명	1998-12-30	1998-12-31	서울대병원	OO	1985xxxx-2xxxxxx	홍길동	1946xxxx-1xxxxxx	나야나	나야나	최고임
2	생명보험	현대해상	1998-12-31	1999-01-01	연세대병원	XX	1985xxxx-2xxxxxx	홍길동	1946xxxx-1xxxxxx	나야나	나야나	최고임
3	자동차보험	MG	1999-01-01	1999-01-02	순천향대병원	OO	1946xxxx-1xxxxxx	김순이	1985xxxx-2xxxxxx	나야나	나야나	최고임
4	상해보험	메리츠	1999-01-02	1999-01-02	서울대병원	DD	1946xxxx-1xxxxxx	김순이	1985xxxx-2xxxxxx	나야나	나야나	최고임
5	생명보험	삼성생명	2000-10-20	2000-10-20	연세대병원	AA	1985xxxx-2xxxxxx	홍길동	1946xxxx-1xxxxxx	김순이	나야나	최고임
6	생명보험	현대해상	2000-10-20	2000-10-20	가천의대병원	BB	1946xxxx-1xxxxxx	김순이	1985xxxx-2xxxxxx	홍길동	홍길동	최고임
7	자동차보험	MG	2000-10-20	2000-10-20	서울대병원	CC	1946xxxx-1xxxxxx	김순이	1985xxxx-2xxxxxx	홍길동	홍길동	최고임
8	생명보험	삼성생명	2003-04-05	2003-04-05	연세대병원	EE	1985xxxx-2xxxxxx	홍길동	1946xxxx-1xxxxxx	김순이	김순이	최고임
9	생명보험	현대해상	2003-04-05	2003-04-05	인하대병원	FF	1985xxxx-2xxxxxx	홍길동	1946xxxx-1xxxxxx	김순이	김순이	최고임
10	자동차보험	MG	2003-04-05	2003-04-05	차병원	OO	1946xxxx-1xxxxxx	김순이	1985xxxx-2xxxxxx	홍길동	홍길동	최고임
11	상해보험	메리츠	2003-04-05	2003-04-05	연세대병원	OO	1946xxxx-1xxxxxx	김순이	1985xxxx-2xxxxxx	홍길동	나야나	최고임
12	생명보험	삼성생명	2010-02-03	2010-02-03	순천향대병원	XX	1985xxxx-2xxxxxx	홍길동	1946xxxx-1xxxxxx	김순이	나야나	최고임
13	생명보험	삼성생명	2019-02-19	2019-02-19	서울대병원	XX	1985xxxx-2xxxxxx	홍길동	1946xxxx-1xxxxxx	김순이	김순이	최고임
14	상해보험	메리츠	2019-02-19	2019-02-19	연세대병원	XX	1946xxxx-1xxxxxx	김순이	1985xxxx-2xxxxxx	홍길동	홍길동	최고임

1 mode network

2) Transaction data 의 전처리: network modeling

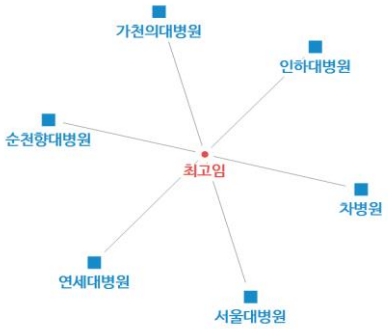
1 mode network

1-mode network Table(피보험자-계약자)



2 mode network

2-mode network Table(설계사-병원)



Big Data 의 분석

	데이터 수집	전처리	측정
	SNS에서 이용자 활동, 작성글, 정보 등을 수집	수집 데이터 가공	분석 방법 적용
Network	**이용자 상호작용** • 페이스북: 좋아요, 댓글-답글 등 • 트위터: RT, Reply • 유튜브: 좋아요, 댓글-답글 등	**네트워크 구성** • 페이스북: 좋아요 관계, 댓글-답글 관계, 같은 글에 좋아요한 관계,.. • 트위터: RT 관계, Reply 관계, 같은 글에 RT한 관계,... • 유튜브: 좋아요 관계, 댓글-답글 관계, 같은 동영상에 댓글을 작성한 관계,...	**네트워크 분석** • 중요도: Centrality • 클러스터링: Community
Text	**비정형 텍스트 데이터** • 페이스북: 포스트, 댓글, 답글.. • 트위터: 트윗글, 자기소개.. • 유튜브: 동영상 제목 및 설명, 댓글, 답글..	**자연어 처리** • 형태소 추출 및 품사 태깅 • 빈도 및 중요도 측정 • 단어의 위치 정보 추출 • 문장/문단/문서 추출 • 사전(유의어, 고유명사 등) 처리	**텍스트 마이닝** • 중요도: TF-IDF, 빈도 • 클러스터링: LDA(토픽모델링)

1) Text Mining

텍스트 데이터에 대한 체계적/정량적 분석의 필요성

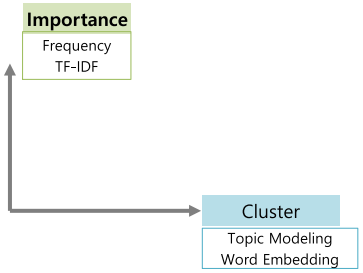
	Structured Data	Unstructured Data
Quantitative Analysis	Statistics Social Network Analysis Machine Learning	???
Qualitative Analysis		Contents Analysis

⇧

수작업과 분석자의 직관, 경험적 해석

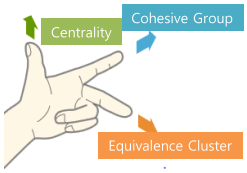
1) Text Mining

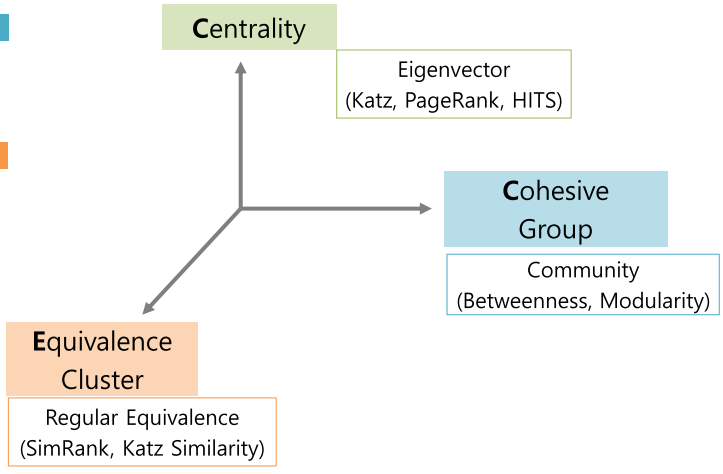
Text Mining



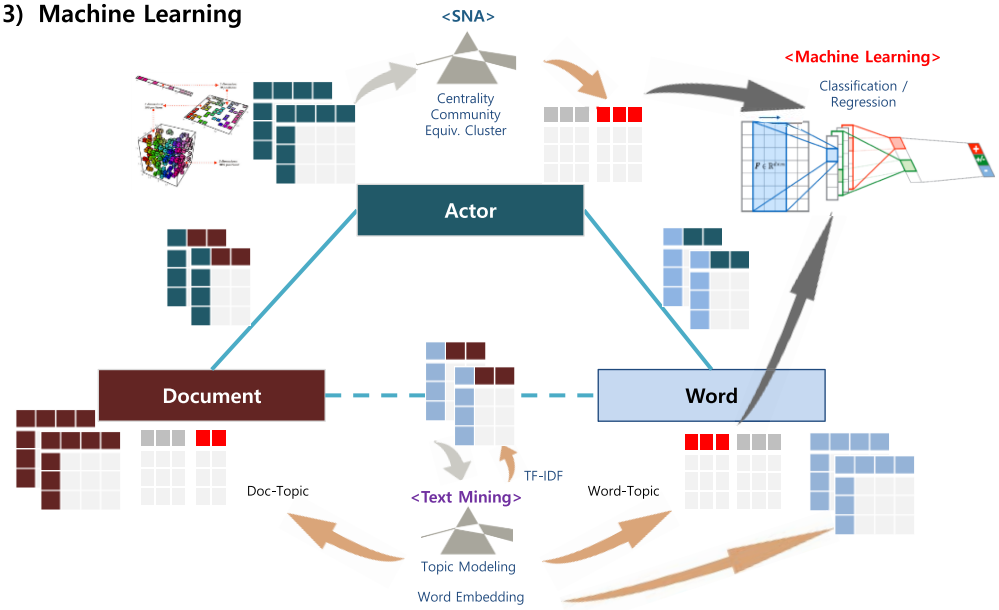
2) Social Network Analysis

Measure positions in social structure defined by intersecting relations





3) Machine Learning



사회과학 연구자를 위한 Big Data 교육과정

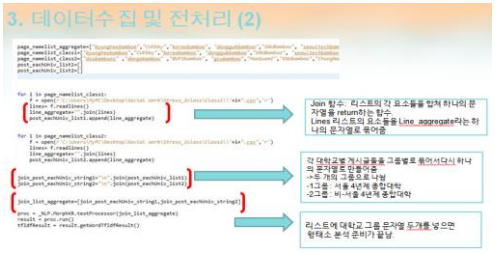
2019/1학기/정규	학사	M1304.001200	001	소셜 빅데이터 조사분석	전선	3-3-0	한국어

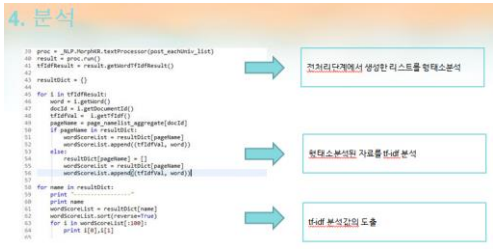
I. Python 문법

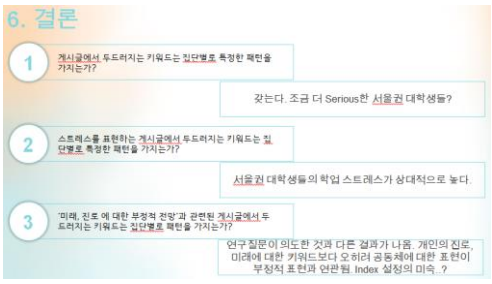
목차	세부 목차
1. 프로그래밍 입문(개요)	프로그래밍 언어
	Why Python?
	파이썬 설치 및 실행
	파이썬 개발환경(IDLE)
2. 데이터	변수
	Python의 자료형 (Data Type)
	Python의 자료구조 (Data Structure)
3. 흐름제어	조건문
	반복문
	예외처리
4. 함수	개요
	Input
	Output
	Return
	함수 안에서의 변수
5. 파일 입출력	파일 입출력의 필요성
	파일 읽기
	파일 쓰기
6. 클래스	클래스의 필요성
	클래스의 정의
	클래스의 사용
7. 모듈	모듈의 생성과 사용
	내장 모듈
	외부 모듈의 사용

II. 빅데이터 수집 분석 연습

목차	세부 목차
1. 빅데이터 분석 개요	빅데이터 입문
	빅데이터 분석 절차
2. 수집	개발환경
1) 데이터 획득 방법	크롤링, API, 인증 등
2) Web 크롤링	개요
	게시판 제목 수집
	네이버 실시간 검색어 수집
3) Twitter	Tweepy 개요
	트위터 가입 및 앱 생성
	Tweepy 설치
	인증
	데이터 수집 API
3. 전처리	형태소 분석
	데이터 구조화
	NetMiner 입력
4. 분석	통계 분석: 빈도 분석
	핵심단어분석(TF-IDF)
	소셜네트워크영향력 분석
	단어-문서 연관성 분석 - word2vec



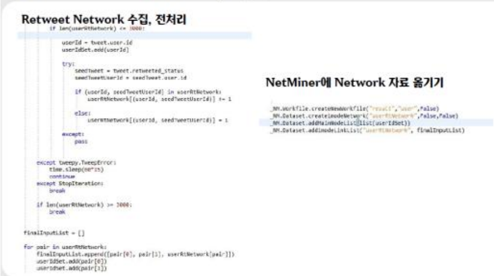




개념어 검색 및 수집

개념어 리스트 간 Tf-Idf 분석

Katz Status Centrality 분석

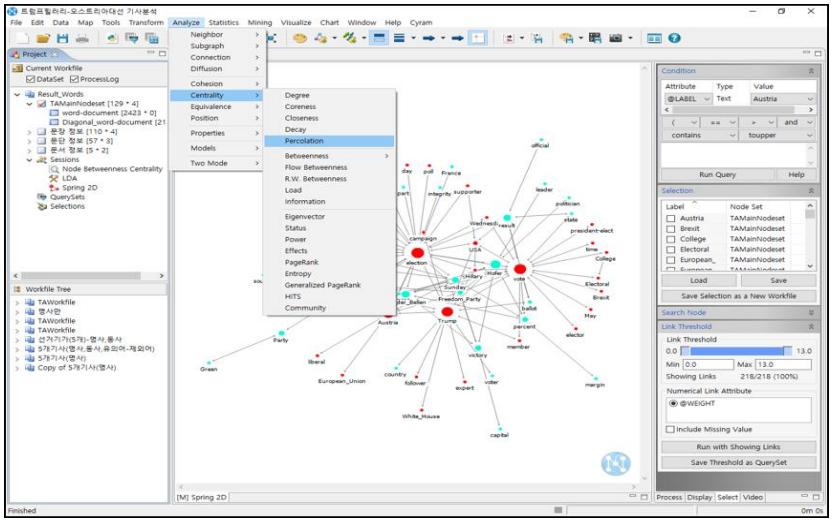


상위 영향력자 트윗글 리스트 간 Tf-Idf 분석

사회과학 연구자를 위한 Big Data 소프트웨어

CYRAM NetMiner

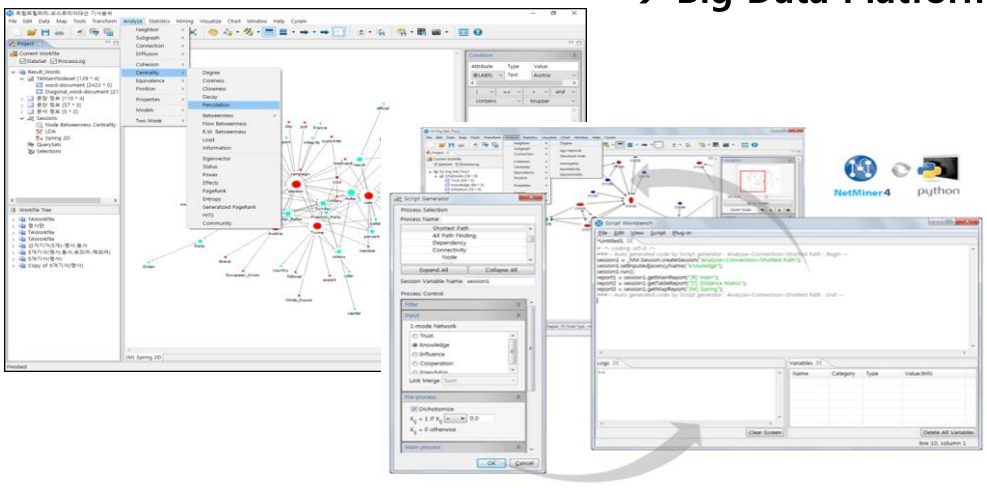
Since 2001



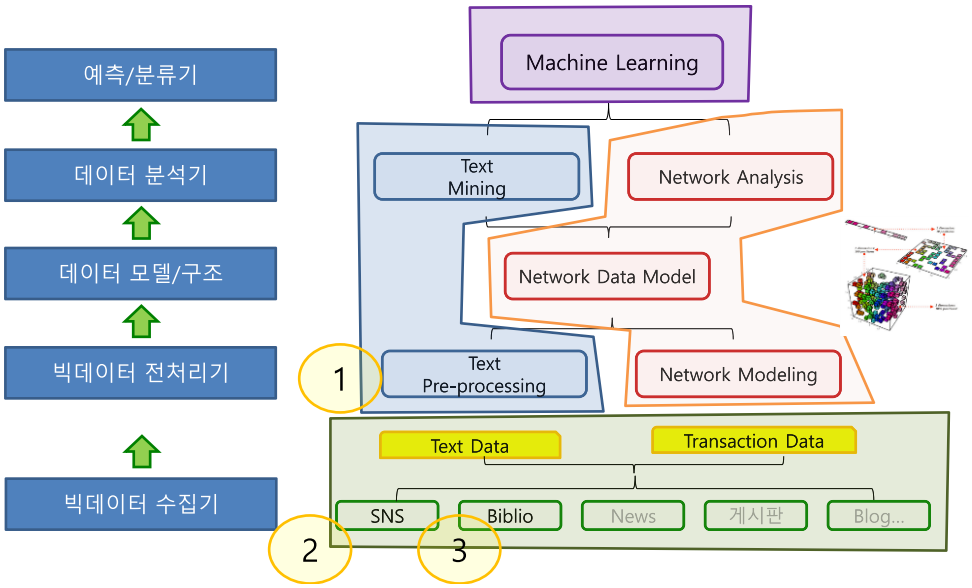
NetMiner + Python

→ Big Data Platform

NetMiner 4.0 (2011. 5)

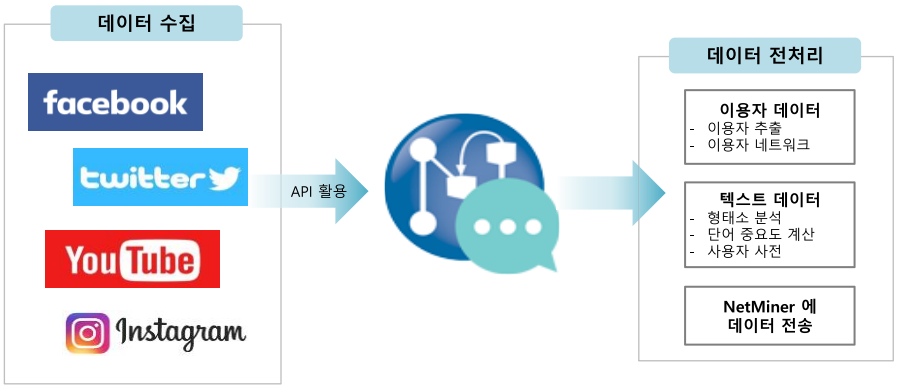


NetMiner 빅데이터 분석 Architecture



2. SNS Data Collector

주요 소셜미디어 의 공개 API(Application Programming Interface)를 이용하여 데이터를 수집하고,
수집한 데이터에 대한 전처리, 네트워크 모델링을 자동으로 수행하는
NetMiner의 확장 프로그램

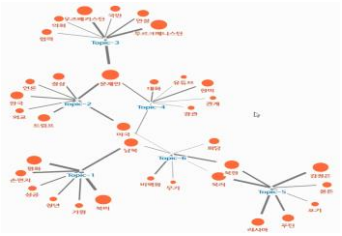


* Instagram 은 5월 출시 예정

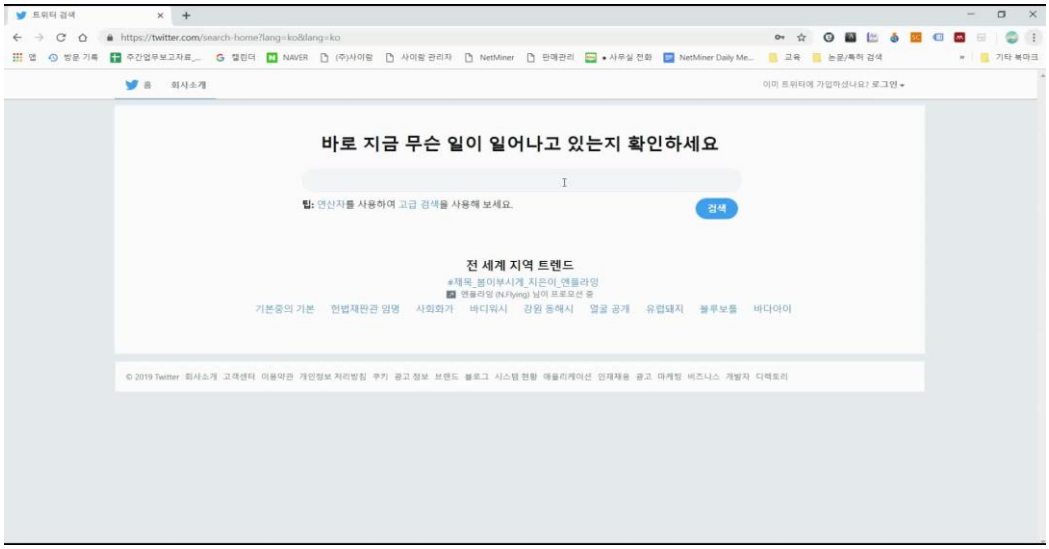
2. SNS Data Collector

- **수집데이터: '정상회담' 관련 트윗**
- **수집기간: 19년 4월 3째주(2019.4.13~4.19)**

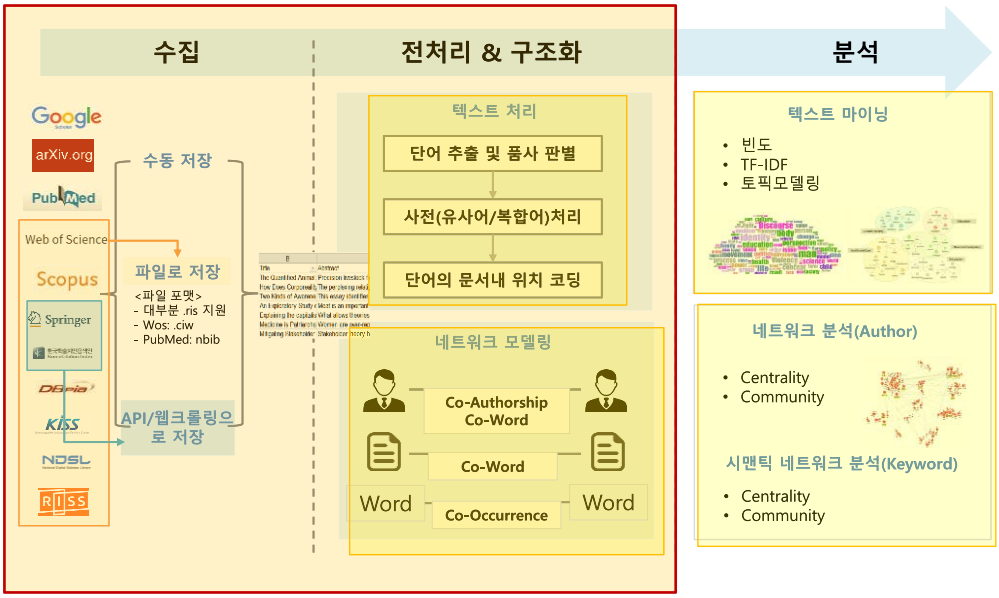
데이터 수집	데이터 전처리	관계망 구성	관계망 분석
• 수집 대상: 트위터 • 수집 내용 : '정상회담'을 포함한 트윗글 • 2000건 트윗글 수집	• 트윗글에서 단어 추출 - 명사 단어만 추출 - 사전 처리(유사어/복합어) • 트윗 이용자 정보 추출 - 이름, 사용언어, 팔로워 수..	• 시맨틱 네트워크 구성 : 단어 위치 정보를 토대로 동시 출현 단어 관계망 • 이용자 네트워크 구성 : RT 관계, 유사 관심 그룹	**① 시맨틱 네트워크 분석** • 세부 토픽 추출 ~~② 이용자 네트워크 분석~~ ~~- 인플루엔서 분석~~ ~~- 이용자 그룹 추출~~



NetMiner SNS Data Collector



3. Biblio Data Collector



3. Biblio Data Collector

- **수집데이터: 범죄자 심리를 연구한 국내 논문 서지 정보(문헌 정보)**
- **수집기간: 2000.1~2019.3**

데이터 수집

- 수집 대상: KCI(학술연구재단)
- 수집 내용
  - 논문 제목, 초록, 키워드
  - 연도, 출처 등
- 1000건 논문 서지 정보 수집

데이터 전처리

- 논문 초록에서 단어 추출
  - 명사 단어만 추출
  - 사전 처리(유사어/복합어)
- 논문 저자 정보 추출

관계망 구성

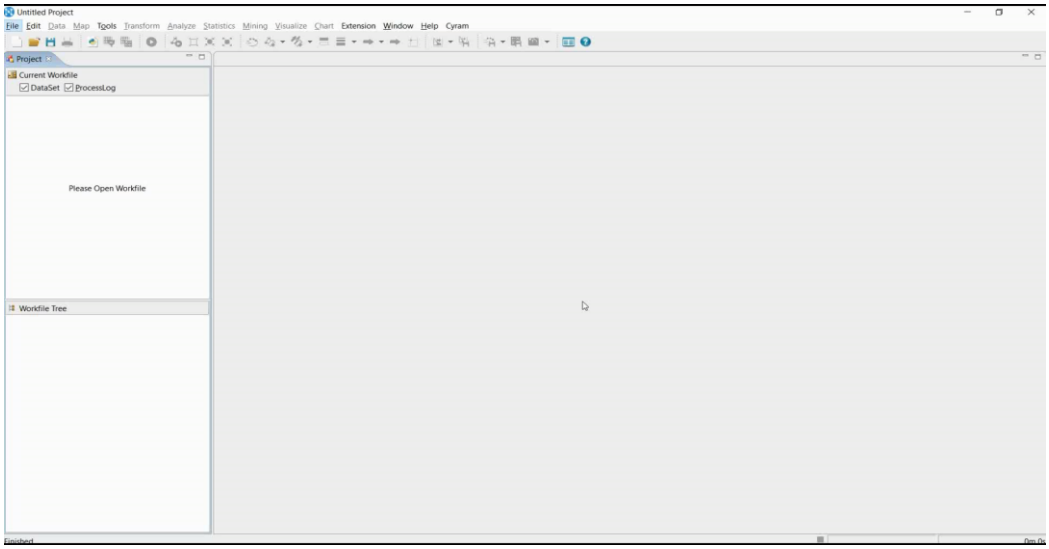
- 시맨틱 네트워크 구성
  : 동시 출현 단어 간 관계망
- 저자 네트워크 구성
  : 공동 저술 저자 간 관계망

관계망 분석

① **시맨틱 네트워크 분석**
- 세부 토픽 추출
- 관심 단어 빈도 추이

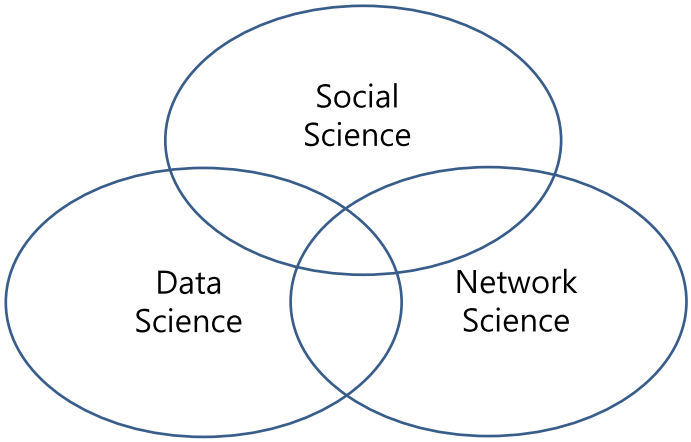
② **저자 네트워크 분석**
- 전문가 탐색

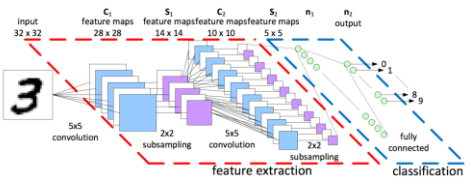
NetMiner Biblio Data Collector

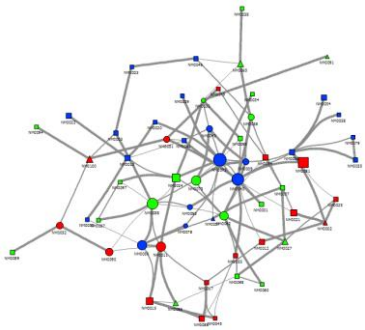


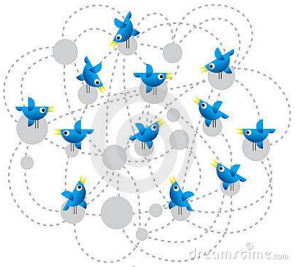
맺음말

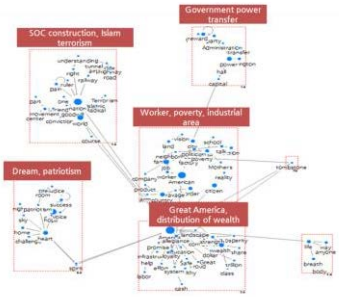
Future direction of social sciences ?











제7회

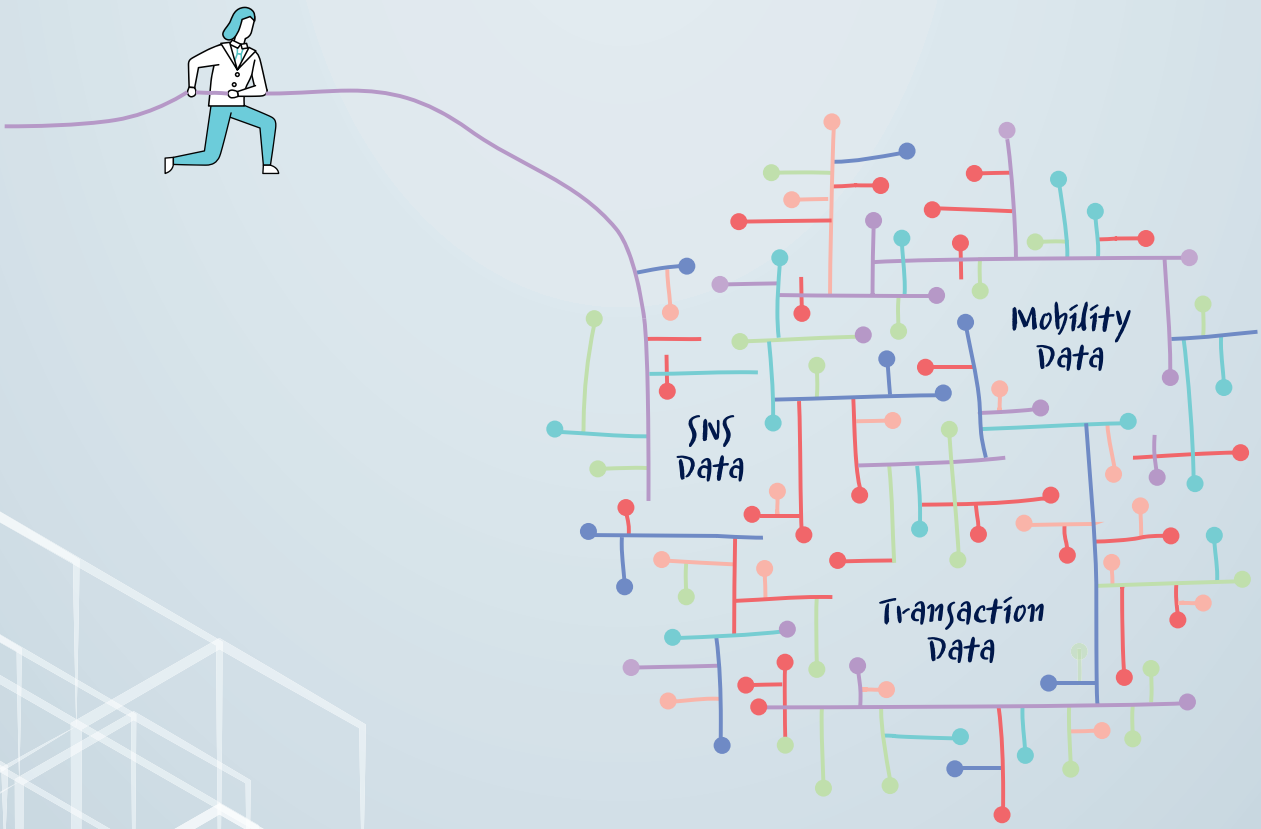
KOSSDA

데이터 페어

2부: 빅데이터 활용 사례

모빌리티 데이터로 바라보는 사회

김정민 연구원 (카카오 모빌리티)



모빌리티 데이터로 바라보는 사회

김정민 / 데이터랩 / 카카오모빌리티
dominic.jmkim@kakaomobility.com
KOSSDA 데이터페어
2019. 06. 27.

1

Contents

- 모빌리티 데이터
- 데이터 분석 방법론
- 모빌리티 데이터로 바라보는 사회 (연구사례)

2

모빌리티 데이터

3

모빌리티 데이터

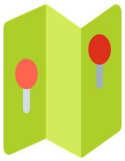

승객


기사


호출


운행


위치


지도


결제


프로모션

수요공급 불일치 문제를 완화하여 승객과 기사 모두에서 '빠르고 편리한 운행'을 제공하는 것이 목표

4

모빌리티 데이터

장소(Place) = 위치(Location) + 맥락정보(Context)
위치정보 변환: 단말 신호 (GPS, Wifi, LTE 등)
맥락정보: 카카오맵에서 장소와 카테고리를 지속적으로 관리
** 주소, 상호, 카테고리 등

데이터	장점	단점
단순 휴대폰 위치 데이터	지속적으로 정확한 위치 파악 가능	사용자 방문지, 혹은 행위 유추가 어려움
금융 결제 데이터	정확한 행위 파악 가능	이동 경로, 결제가 없는 행위 유추 불가
대중교통 승하차 데이터	이동 경로 및 시간 정확히 파악 가능	행위 유추 및 최종 목적지 확인 불가
검색 데이터 [다음, 네이버 등]	관심사 파악 가능, 행위 예측 가능	실제 행위 여부 파악 불가
모빌리티 이동 데이터	이동 경로 및 시간, 비결제를 포함한 행위 파악 가능	사용하지 않는 시간대가 존재

표: 사용자 위치 및 행동 파악을 위한 데이터 비교

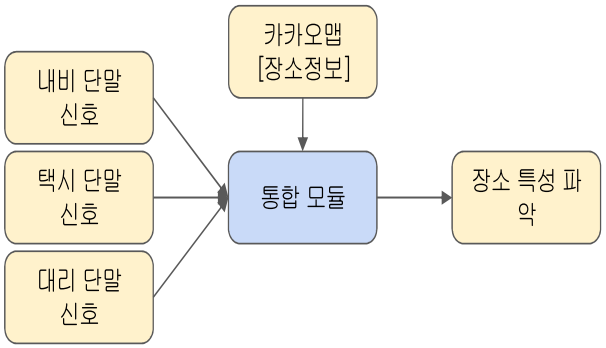
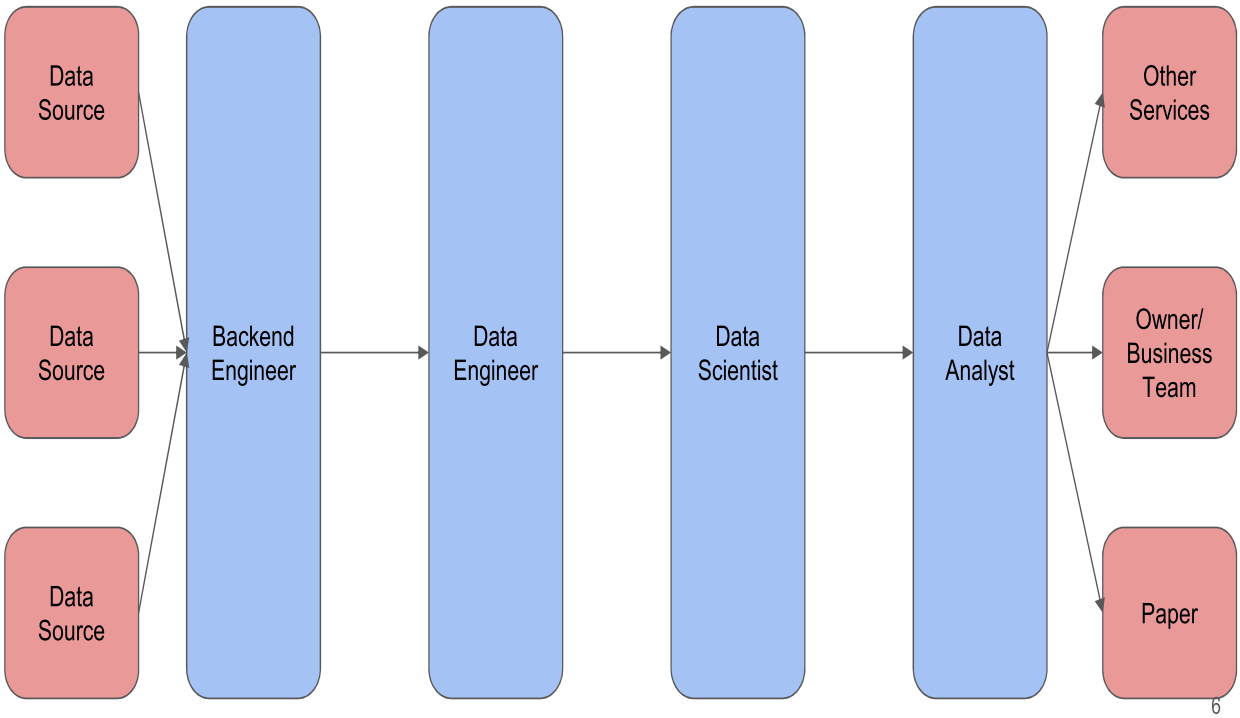


그림: 카카오모빌리티 이동 데이터 수집 및 분석 과정

5

데이터 분석 흐름



6

ETL (Extract, Transform, Load)

데이터는 그 의미를 추출(Extract)하고 분석 가능한 형태로 변환(Transform)하여 적재(Load)해야 한다.

추출: 텍스트(자연어처리), 이미지(물체 인식), 사운드 (주파수 분석) 등..
변환: 분석 하고자 하는 내용을 Byte/Number/String, 정형/비정형 등 적절한 형태로 변환
** 추출, 변환 및 저장의 효율을 높이기 위해서 다양한 분산처리 시스템 (Distributed System)이 사용된다. [Hadoop 등]

적재 방식:
Batch → Micro Batch → Streaming
Batch: 주기적으로 ETL 작업을 실행 (매달, 매주, 매일, 매시 ..)
Micro Batch: Batch의 주기를 매우 짧게하여 Streaming과 유사하게 처리
Streaming: 데이터를 하나의 흐름처럼 실시간으로 처리

Spark Streaming, kafka

Streaming Framework

저장소: DB(SQL/NoSQL), File System(Local/HDFS)

MySQL, ORACLE, cloudera IMPALA, HIVE

SQL DB

mongoDB, cassandra, APACHE HBASE

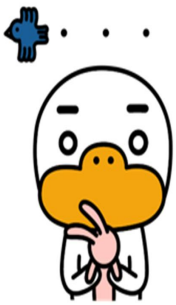
NoSQL DB

hadoop

분산처리 프레임워크
HDFS (Hadoop File System)
MapReduce: 병렬처리 연산 프레임워크

7

대한민국 모빌리티 연구의 현실?



학교, 연구소
연구할 사람은 많은데
좋은 데이터가 없다

VS

회사
데이터는 많은데
연구할 사람이 없다

8

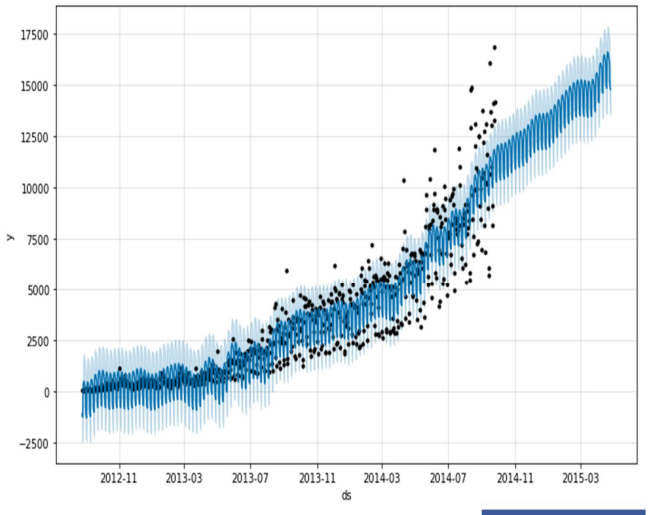
데이터 분석 방법론

9

시계열 분석

시계열 분석은 시간대에 따른 값을 통해
패턴이나 특이점을 찾아내는 분석

경향(trend): 장기적 변화
계절성(seasonality): 시간대별 유사 패턴
주기(cycle): 일정한 진폭마다 유사 변동
불규칙성(irregular): 오차, 잡음, 이벤트



PROPHET

10

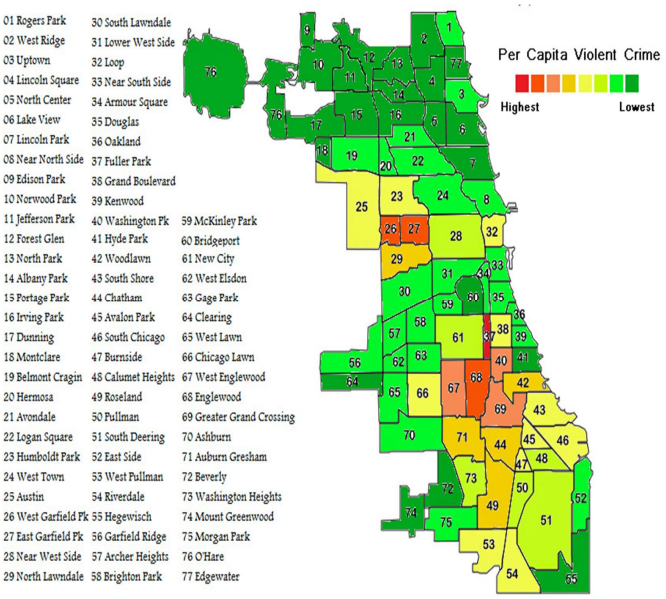
공간 통계 분석

공간 통계 분석은 위상, 기하, 지리적 특성을 사용하여 개체를 설명하는 방법이다.

지리적으로 가까울수록 유사한 특성을 나타낸다.
(First Law of Geography, Tobler, 1970)

→ 공간 자기 상관 **(Spatial Autocorrelation)**

일반적으로 통계에서는 각 사건의 독립성을 전제한다. 그러나 공간 통계 분석에서는 지리적 유사성, 즉 공간 자기상관을 고려하여 분석한다.



http://www.city-data.com/forum/chicago/2411609-crime-rate-chicago-exaggerated.html

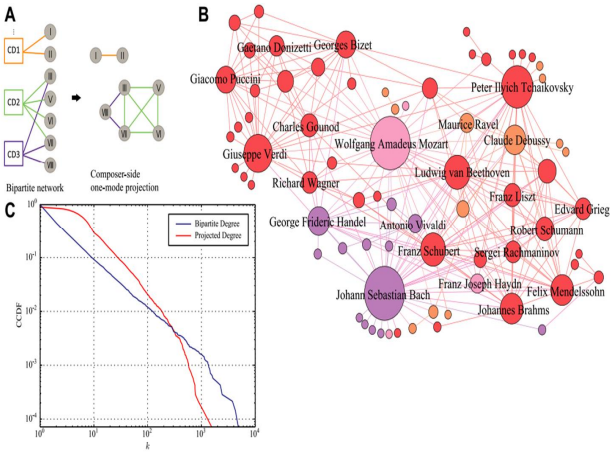
11

네트워크 분석

Point/Node: 하나의 개체
Link/Edge: 개체 사이의 관계

분석 방법:
중심도(centrality), 밀도(density), 반경(diameter) ..

주 분석 분야:
사회관계망, 인프라망, 문서분석 ..

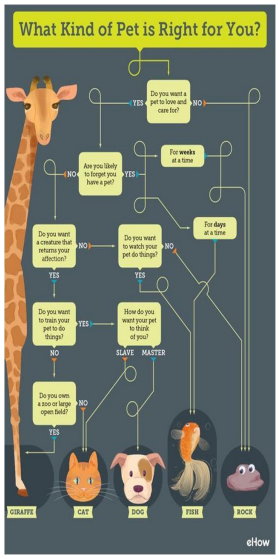


Park, D., Bae, A., Schich, M. et al. EPJ Data Sci. (2015) 4: 2. https://doi.org/10.1140/epjds/s13688-015-0039-z
[고전 음악가 네트워크 분석 - 발표자 주]

12

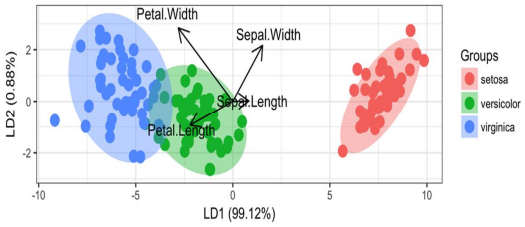
분류 분석 (classification)

특정 데이터들을 몇 개의 정해진 기준으로 분류하는 분석



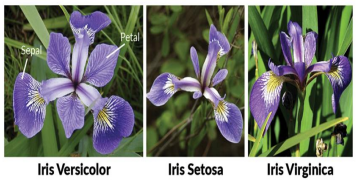
https://www.pinterest.com/pin/764205249061635 43/a

미리 정의된 그룹으로 데이터를 분류
: Decision Tree, Neural Networks,
SVM(Support Vector Machine), Ensemble
Method, kNN(k Nearest Neighborhood),
Logistic Regression ..



https://lchblogs.netlify.com/post/2017-12-22-r-addconfellipselda/

데이터 내부 특성을 사용하여 자동 분류
: Clustering



http://www.lac.inpe.br/~rafael.santos/Docs/R/CAP394/WholeStory-Iris.html

13

자연어 분석

인간의 언어를 기계를 이용해서 분석
기계학습의 하위 분야로 인식

분석 순서(한국어): 형태소 분석 → 품사 부착
→ 구절단위분석 → 구문분석

정보 검색, 정보 추출, 음성 인식, 단어분류,
문장/문서 분류, 감정분석, 기계번역

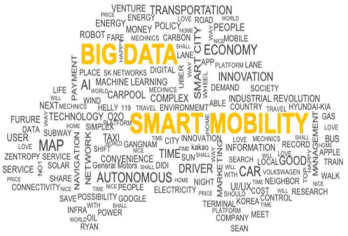


Figure 4: Anger. Figure 5: Fear. Figure 6: Disgust.

Figure 7: Happiness. Figure 8: Sadness. Figure 9: Surprise.

소셜 미디어 기반(geo tagged tweets) 의 감정 지도 (워싱턴) [발표자 -주]

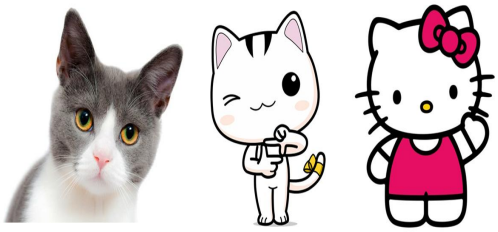
Yerach Doytsher, Ben Galon, and Yaron Kanza. 2017. Emotion Maps based on Geotagged Posts in the Social Media. In Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities (GeoHumanities'17). ACM, New York, NY, USA, 39-46. DOI: https://doi.org/10.1145/3149858.3149862

14

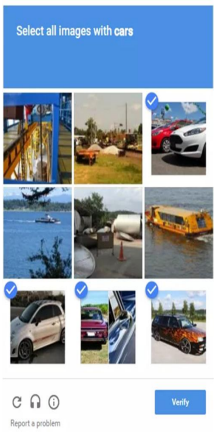
이미지 분석

이미지 분석은 인간이 컴퓨터보다 더 잘하는 분야이다.
이미지의 맥락을 이해하는 것은 오랜 기간 어려운 문제였다.

Google Cloud Vision API:
Label Detection(사물 인식)
Logo Detection(회사 로고 인식)
Landmark Detection (랜드마크 인식)
Face Detection (사람 얼굴 찾기 및 감정 분석)
Optical Character Recognition (문자인식) 등



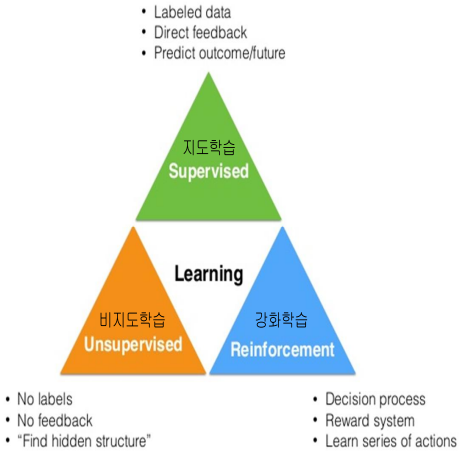
위의 이미지들은 고양이인가요?



reCAPTCHA
매크로 방지 및 이미지 분석 태그 획득

고양이 사진: https://www.wadiz.kr/web/campaign/detail/20802
고양고양이: https://www.facebook.com/goyangcity
헬로키티: https://www.altpress.com/news/hello_kitty_creators_reveal_character_isnt_actually_a_cat/

머신 러닝



지도 학습(Supervised Learning): 데이터에 대한 명시적인 정답이 주어진 상태에서 컴퓨터를 학습 / classification, regression
비지도 학습(Unsupervised Learning): 데이터에 대한 명시적인 정답이 주어지지 상태에서 컴퓨터를 학습 / clustering
강화 학습(Reinforcement Learning): 에이전트가 주어진 환경(state)에 대해 어떤 행동(action)을 취하고 이로부터 어떤 보상(reward)을 얻으면서 학습,
에이전트는 보상(reward)을 최대화(maximize)하도록 학습 / alphago

많은 경우에 머신러닝은 기존에 있던 분석 방법들을 더 빠르고 정확하게 수행해 주는 역할을 한다.

모빌리티 데이터로 바라보는 사회

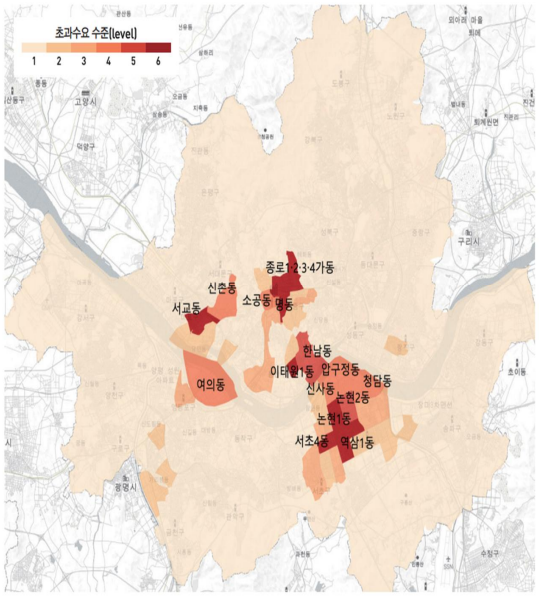
17

시공간 분석

18

시야 교통 불균형 현황(택시) 빅데이터 분석

<그림 4-3> 서울의 심야시간 택시 초과수요 현황



<표 4-1> 심야시간 택시의 초과수요가 많은 지역

초과수요 수준	초과수요가 많이 발생한 행정동
Level 6-2	역삼1동(강남구), 종로1·2·3·4가동(종로구)
Level 6-1	서교동(마포구), 논현1동(강남구), 명동(중구), 서초4동(서초구), 이태원1동(용산구)
Level 5	논현2동(강남구), 한남동(용산구), 여의동(영등포구), 압구정동(강남구), 청담동(강남구), 신사동(강남구), 신촌동(서대문구), 소공동(중구)

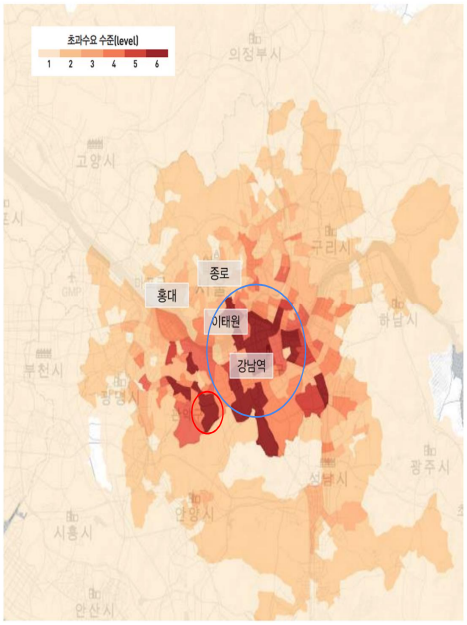
초과수요:
특정 시간대에 택시를 이용하고자 했으나 이용하지 못한 승객(혹은 콜) 수

시민 이동성 증진을 위한 심야 교통 현황 분석 [카카오모빌리티, 서울디지털재단, 2018. 12.]

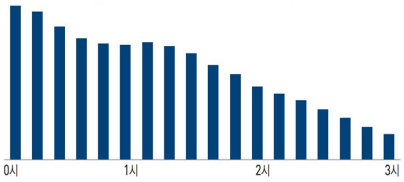
19

시야 교통 불균형 현황(택시) 빅데이터 분석

<그림 4-11> 강남역 심야시간 택시 초과수요의 목적지



<그림 4-10> 강남역 심야시간 시간대별 택시 초과수요 분포



<표 4-3> 강남역 심야시간 택시 초과수요의 목적지

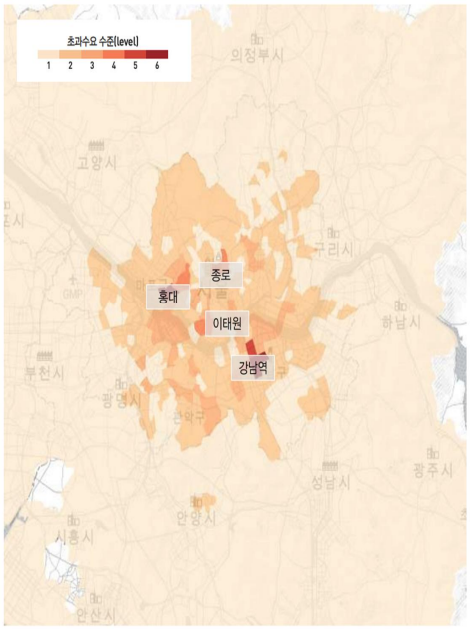
초과수요 수준	초과수요의 목적지(행정동)
Level 6-2	논현1동(강남구), 역삼1동(강남구), 청담동(강남구), 이태원1동(용산구), 잠원동(서초구), 서원동(관악구), 행운동(관악구), 논현2동(강남구), 압구정동(강남구)
Level 6-1	삼성2동(강남구), 신림동(관악구), 인헌동(관악구), 역삼2동(강남구), 자양4동(광진구), 청룡동(관악구), 남현동(관악구), 문정2동(송파구), 서초3동(서초구), 상도1동(동작구), 양재1동(서초구), 잠실2동(송파구), 한남동(용산구), 잠실본동(송파구), 양재2동(서초구)
Level 5	대치4동(강남구), 화양동(광진구), 잠실3동(송파구), 세곡동(강남구), 대방동(동작구), 삼전동(송파구), 반포4동(서초구), 석촌동(송파구), 서초4동(서초구)
Level 4	사당1동(동작구), 방이2동(송파구), 신사동(강남구), 삼성1동(강남구), 흑석동(동작구), 대치2동(강남구), 사당2동(동작구), 행당1동(성동구), 자양2동(광진구), 구로3동(구로구), 노량진1동(동작구), 장안1동(동대문구), 구의3동(광진구), 대학동(관악구), 신사동(관악구), 옥수동(성동구)

시민 이동성 증진을 위한 심야 교통 현황 분석 [카카오모빌리티, 서울디지털재단, 2018. 12.]

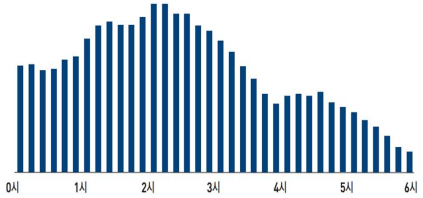
20

시야 교통 불균형 현황(택시) 빅데이터 분석

<그림 4-17> 이태원 심야시간 택시 초과수요의 목적지



<그림 4-16> 이태원 심야시간 시간대별 택시 초과수요 분포



<표 4-6> 이태원 심야시간 택시 초과수요의 목적지

초과수요 수준	초과수요의 목적지(행정동)
Level 6-1	서교동(마포구), 역삼1동(강남구)
Level 5	논현1동(강남구)
Level 4	없음

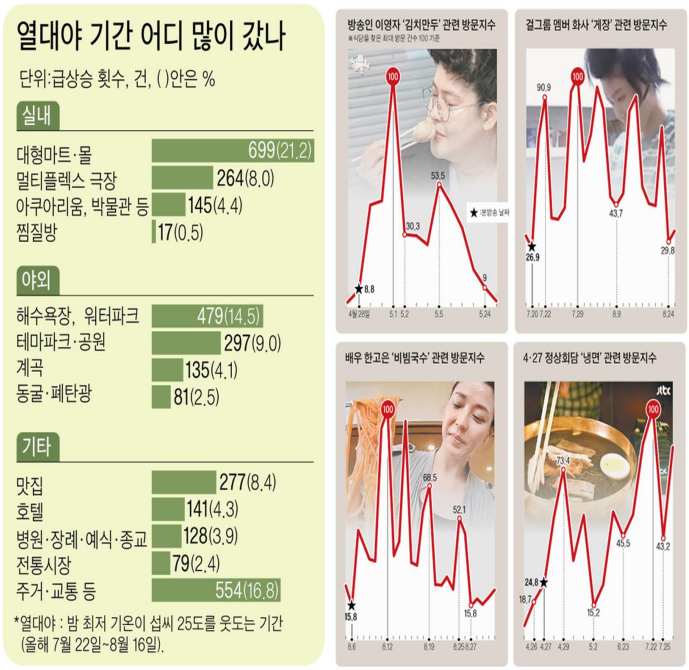
시민 이동성 증진을 위한 심야 교통 현황 분석 [카카오모빌리티, 서울디지털재단, 2018. 12.]

21

Abnormal(Event) Detection

22

이동 패턴 분석 [핫스팟 분석]



시계열 분석을 통해 평소와 다른 이상패턴을 찾아낼 수 있고 이를 기반으로 사람들의 생활 패턴 변화를 살펴볼 수 있다.

단순히 많이 간 곳으로 분석하면 언제나 1위는 비슷하다.

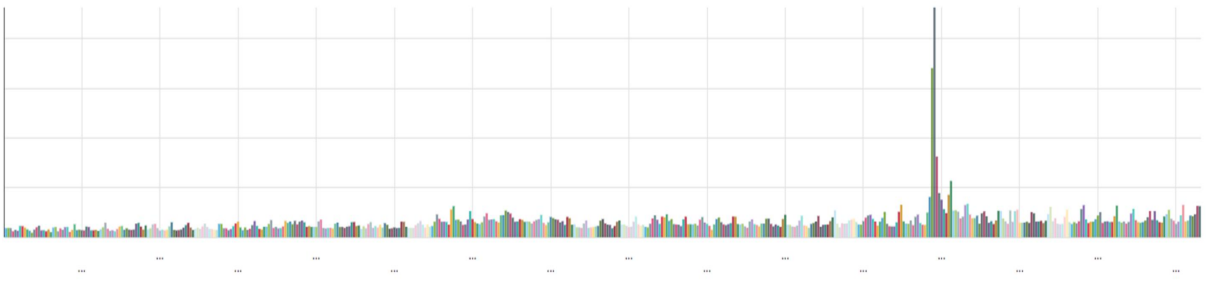
평소보다 갑자기 더 많이 방문한 곳 혹은 카테고리를 분석해야 패턴의 변화를 살펴 볼 수 있다.

23

이동 패턴 분석 [핫카테고리 분석]

1월 26일부터 약 1주간 빨래방의 급증!?
얼어붙은 세탁기에 빨래방 '한파 특수'…동파 막으려면?
(http://news.jtbc.joins.com/article/article.aspx?news_id=NB11581868)

빨래방 카테고리 일간 길안내 수 (2017. 01 ~2018. 04)



24

이동 패턴 분석 [핫카테고리 분석]

2018년 설 기간 핫 카테고리 분석

4일	5일	6일	7일	8일	9일	10일
					2018 동계 올…	하나로클럽, 떡/한과, 국립묘지, 초콜릿
11일	12일	13일	14일	15일	16일	17일
하나로클럽, 납골당, 국립묘지, 장례, 묘지, 정관장, 초콜릿	하나로클럽, 전통식품제조, 정관장	초콜릿, 정관장, 하나로클럽, 도축업, 정육점, 떡/한과	정관장, 초콜릿, 떡/한과, 하나로클럽, 과일,채소가게	설날 연휴 방앗간, 제사음식, 마을회관, 세차장	설날 요양원, 장례, 납골당, 마을회관, 화장터, 빌라, 아파트	설날 연휴 요양원, 장례, 납골당, 묘지, 스케이트장
18일	19일	20일	21일	22일	23일	24일
장례, 스케이트장, 납골당	스케이트장, 입시학원					
25일	26일	27일	28일	3월 1일	2일	3일
2018 동계 올…				삼일절		

25

Clustering

26

Living Zone Analysis

- 행정동 간 통행량을 기반으로 네트워크를 생성하고, 클러스터링을 통해서 생활권을 구분
  ** 이 작업에서 물리적 위치, 거리 정보는 별도로 사용하지 않음
  ** Fast greedy modularity optimization algorithm 사용
- 행정동 간 통행량은 물리적 거리와 매우 높은 상관관계를 보이기 때문에 인접한 곳들이 하나의 클러스터, 즉 생활권으로 표현됨
- 주요 분석 내용
  -각 생활권의 중심 지역(허브)
  -각 생활권 내부에서의 주요 통행
  -시간대에 따른 생활권의 변화
- 생활권 분석은 수요/공급 모델의 정교화 및 공공 정책 수립 근거자료 등으로 사용 가능

27

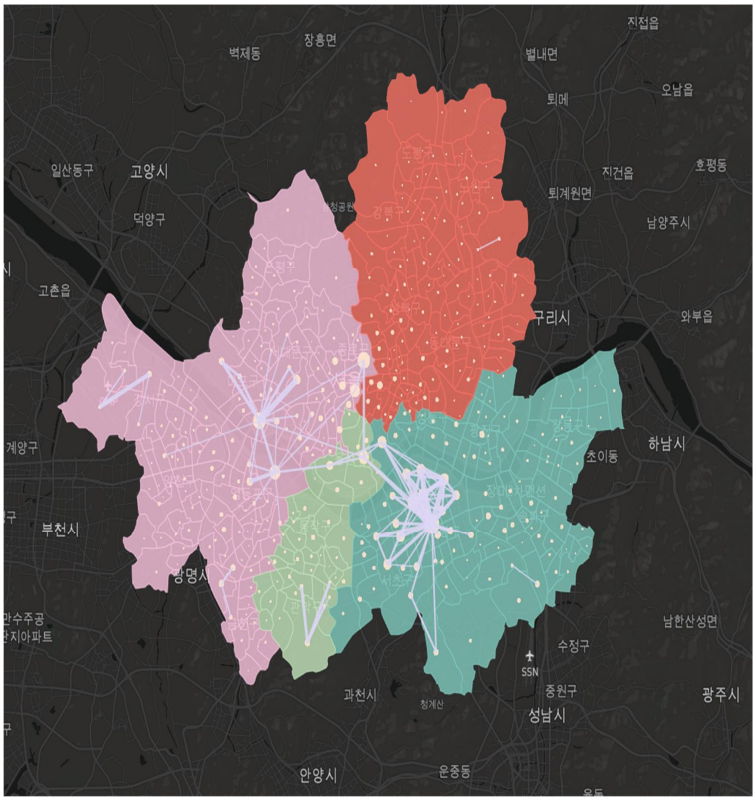
Living Zone Analysis

서울/모든 요일/모든시간
분석 기간:20180901~20181030
분석데이터: 카카오모빌리티 택시 출도착

4개 생활권(임의 명명):
동남권, 동북권, 서부권, 관악권

주요 거점:
강남, 홍대, 광화문, 서울대

서울대를 중심으로 하는 관악권이 별도의 생활권으로 분리되는 이유는?



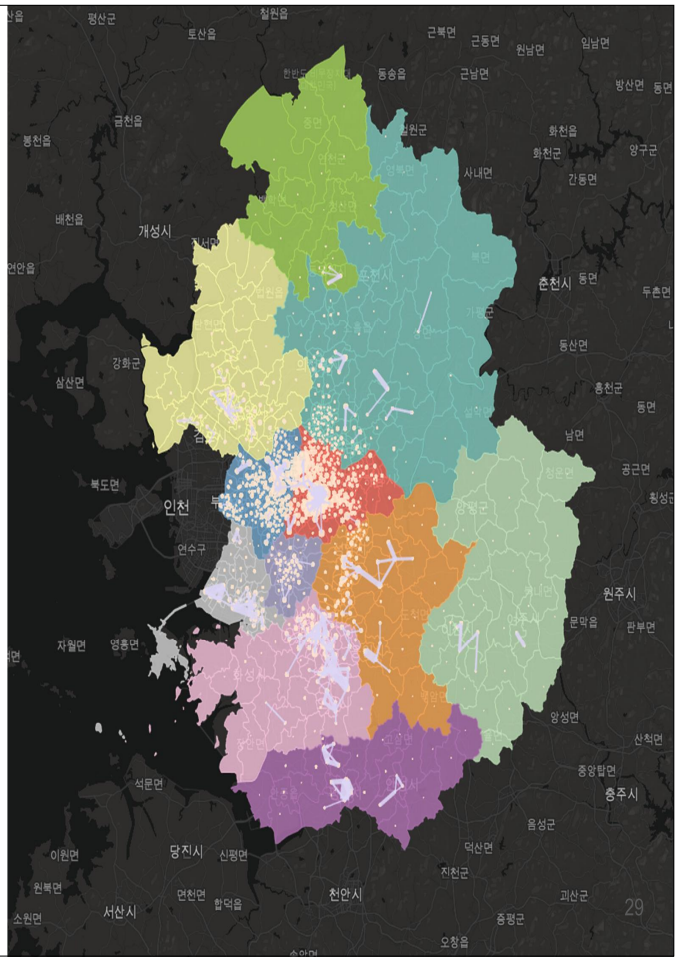
28

Living Zone Analysis

서울,경기/모든 요일/모든시간
분석 기간:20180901~20181030

11개 생활권:

고양-김포-파주권: 일산 중심 통행 많음
서울서부권
서울동부권
서울동북-의정부-양주-포천권
성남-용인-광주권: 광주 –성남 간 통행 많음
수원-화성-오산: 내부 통행 많음
안양권
안산-시흥권: 내부 통행 많음
여주-이천권
연천권
평택권: 평택, 송탄 근방 내부 통행 많음



Summary

- **모빌리티 데이터:** 우리의 삶을 보여주는 중요한 데이터 중 하나
- 모빌리티 데이터 분석
  - 모빌리티 데이터 + 서비스 데이터
  - 다양한 방법론 적용
- 모빌리티 데이터 분석의 결과
  - 보다 빠르고 정확하고 편리한 이동 제공
  - 이동 최적화를 통한 사회적 비용 감소
  - 각 사용자에게 적합한 서비스 제공
  - 행위에 대한 다양한 가설을 검증

30

Thank you.

QnA

31

제7회

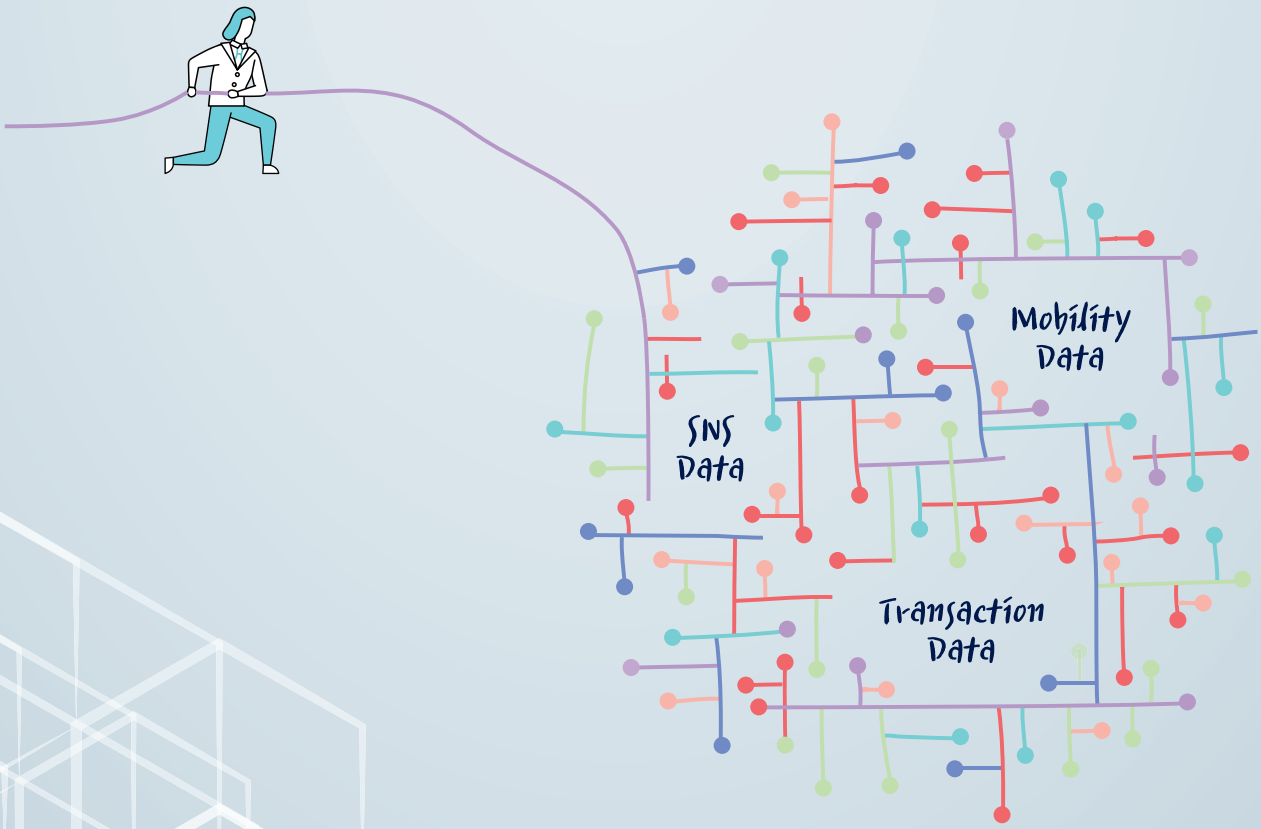
KOSSDA

데이터 페어

2부: 빅데이터 활용 사례

정치연구와 빅데이터

한규섭 교수, **노선혜** 연구원 (서울대 언론정보학과)



pollab
An Analytic Solution to Strategic Communication

빅데이터를 활용한 정치분석

노선혜 (서울대 언론정보학과)

1

빅데이터 여론 추정

2

여론조사심의 위원회 등록 여론조사: 2017년 대선

조사기관	의뢰기관	등록일	구분1	구분2	응답률	문재인	안철수	문재인-안철수
한국리서치	한국일보	4/10/2017	유무선혼합	전화면접	19.3	37.7	37.0	0.7
리얼미터	이데일리	4/10/2017	유무선혼합	ARS/전화면접	11.8	41.1	34.8	6.3
조원씨앤아이	쿠키뉴스	4/10/2017	유무선혼합	ARS	5.8	40.6	34.4	6.2
알앤써치	(주)데일리안	4/12/2017	무선RDD	ARS	4.2	42.3	37.0	5.3
한국리서치	JTBC	4/12/2017	유무선혼합	전화면접	22.3	38.0	38.3	-0.3
리얼미터	MBN,매일경제	4/12/2017	유무선혼합	ARS/전화면접	9.8	44.8	36.5	8.3
리서치뷰	프레시안	4/13/2017	무선RDD	ARS	9.8	46	36.5	9.5
한국갤럽	한국갤럽	4/13/2017	유무선혼합	전화면접	23	40	37	3
리얼미터	MBN,매일경제	4/14/2017	유무선혼합	ARS	10.6	45.4	30.7	14.7

3

역대 총선 출구조사 흑역사

	19^{th} National Assembly ('12)		18^{th} National Assembly ('08)		17^{th} National Assembly ('04)	
	S	DP	S	DP	S	DP
Lower Bound	131	131	155	75	92	157
Upper Bound	147	147	178	93	114	182
Election Outcome	152 (X)	127 (X)	153 (X)	81 (O)	121 (X)	152 (X)

4

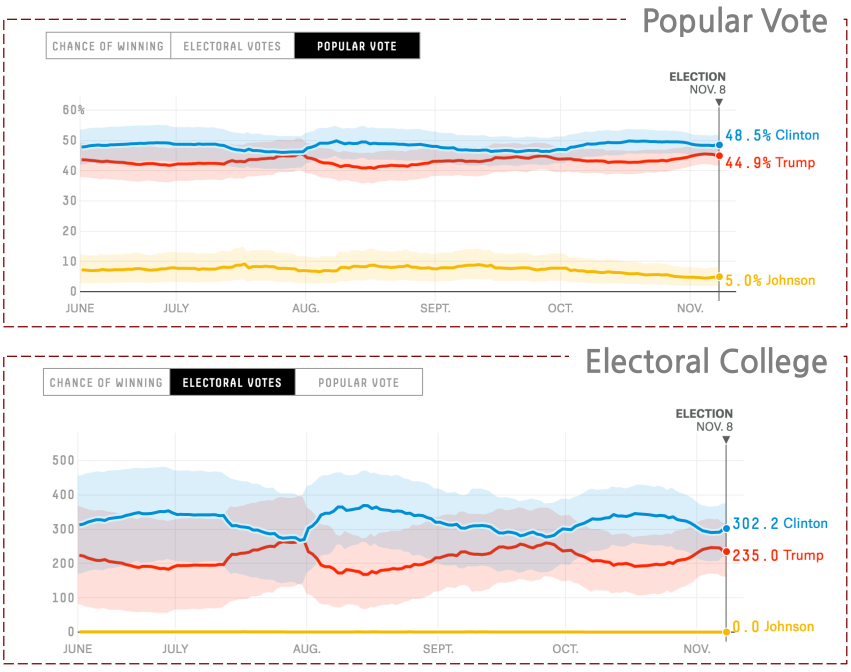
여론조사는 미국대선을 잘못 예측했는가?

2016 미국 대선 선거인단 예측

		FiveThirtyEight	Wall Street Journal
힐러리	227	302	332
트럼프	304	235	206

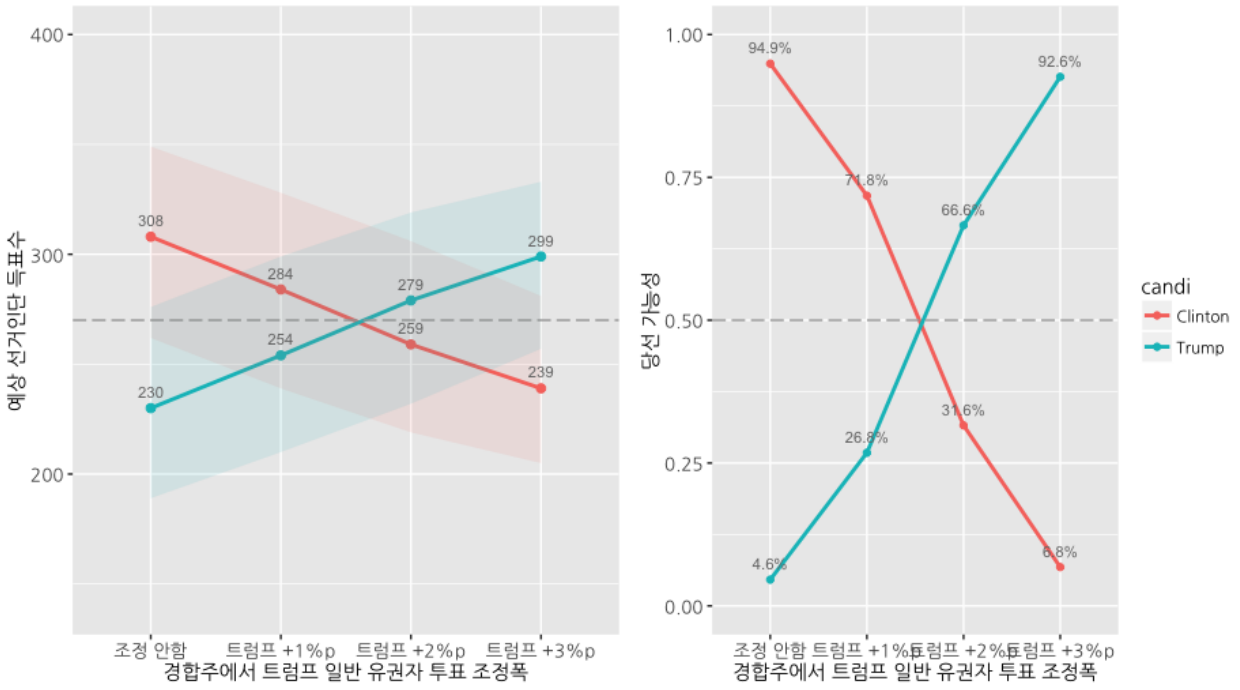
5

여론조사 기반 선거인단 추정 (미국 대선)

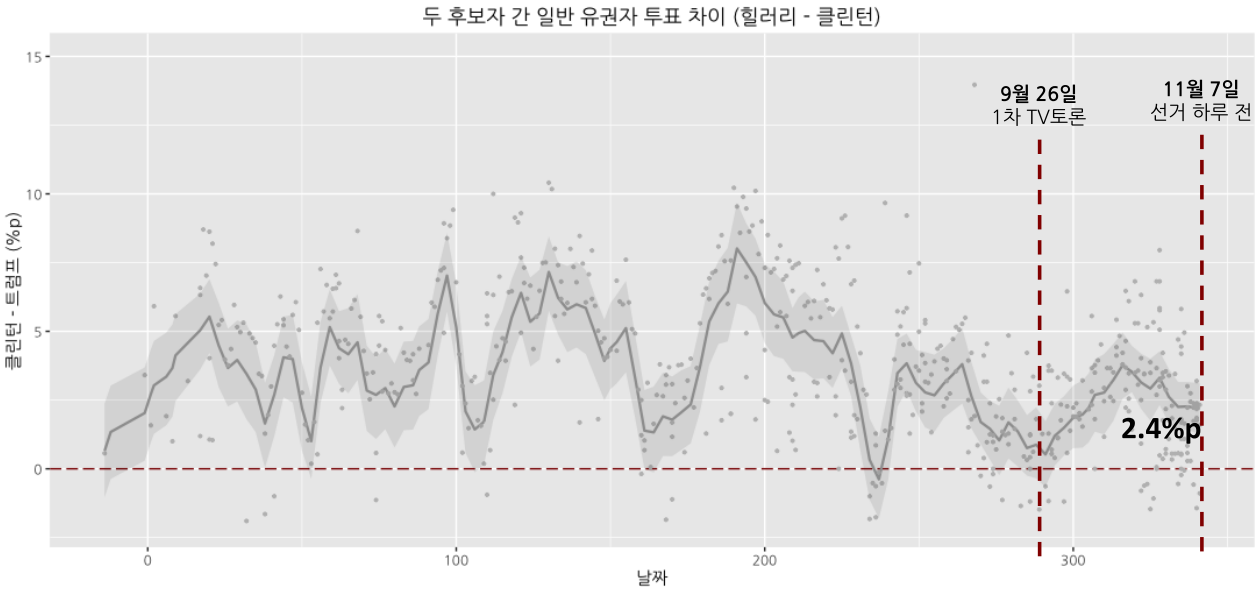


6

2016년 미국 대선 - 선거인단 득표 추정

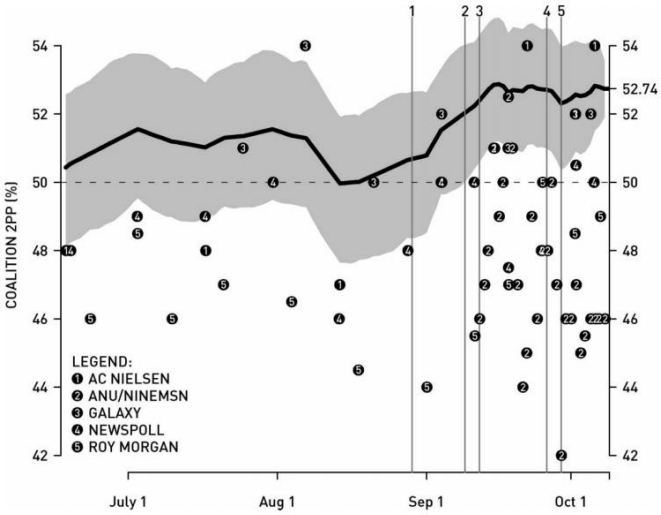


2016년 미국 대선 - 득표율 추정



여론조사 어떻게 할 것인가?

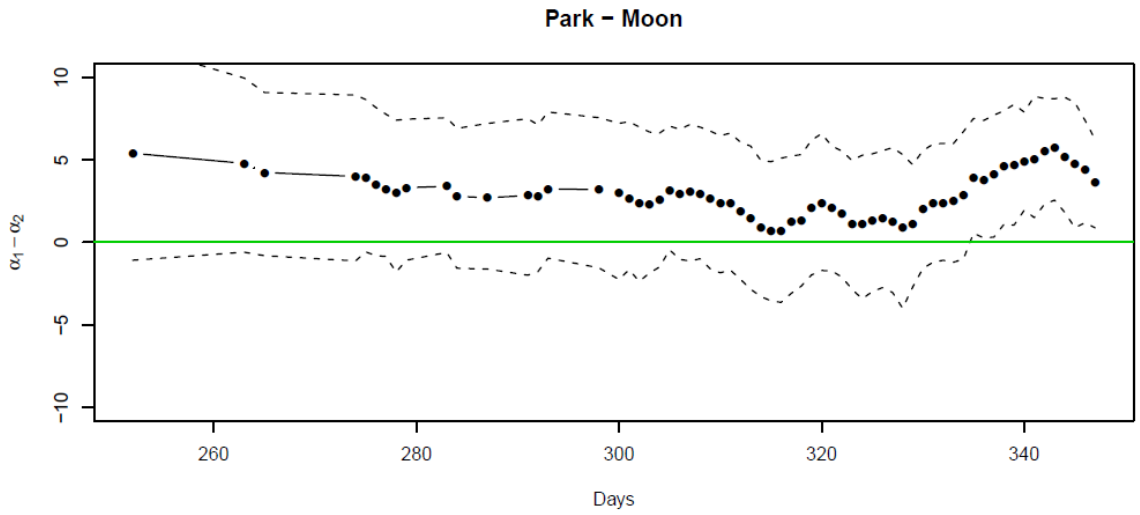
Pooling the polls over an election campaigns



사이먼 잭맨 (Simon Jackman)
호주 국립대학교 정치학과 교수
前 스탠포드대학교 정치학과 교수

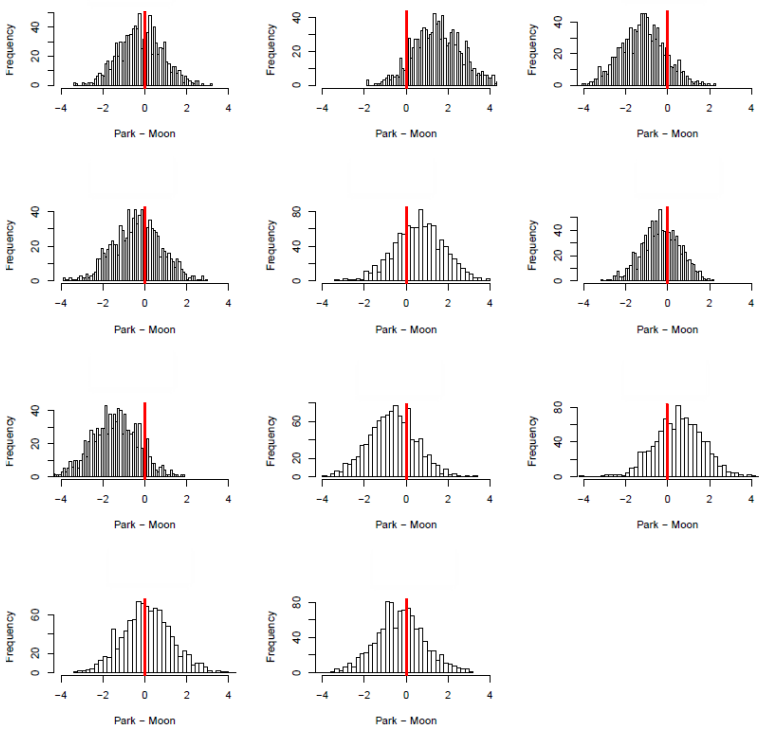
9

2012년 대선 예측



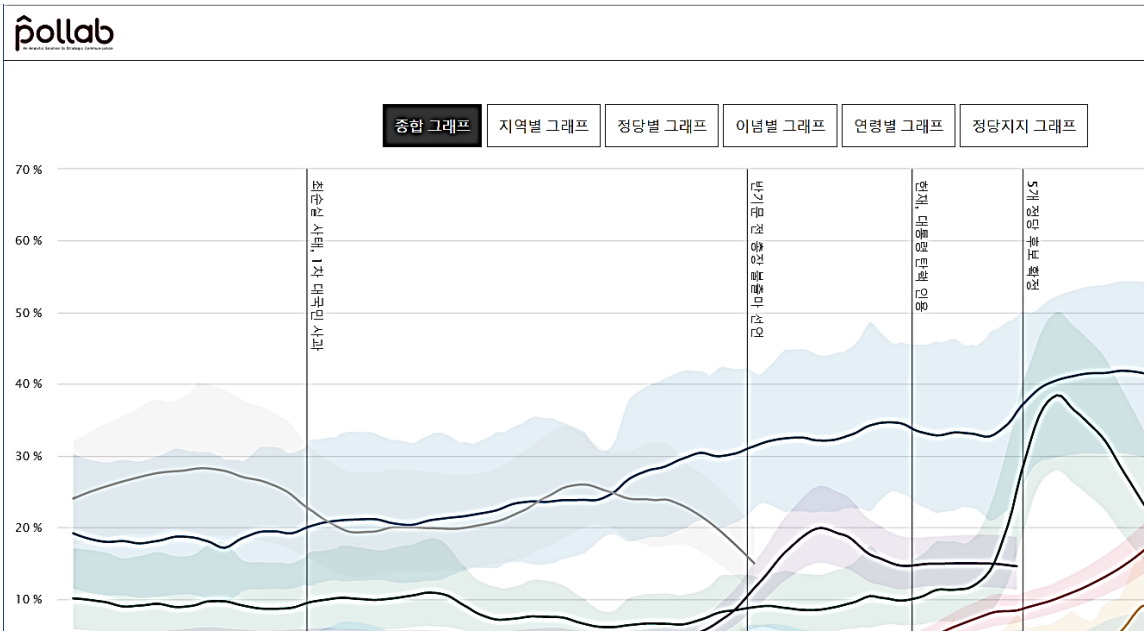
10

2012년 대선 여론조사 - 조사기관 효과



11

pollab 2017 대선 지지율 지수



12

2017 대선 지지율 지수: 최종결과

여론조사 공표 금지 직전 최종

서울대 pollab

	여론조사 공표 금지 직전	등락 (5/2대비)	등락 (4/28 대비)
문재인	41.1%	0.2%p▲	0.9%p▼
안철수	20.8%	0.4%p▼	5.1%p▼
홍준표	18.1%	1.1%p▼	3.4%p▲
심상정	8.6%	1.9%p▼	0.9%p▲
유승민	5.1%	0.1%p▲	0.0%p—

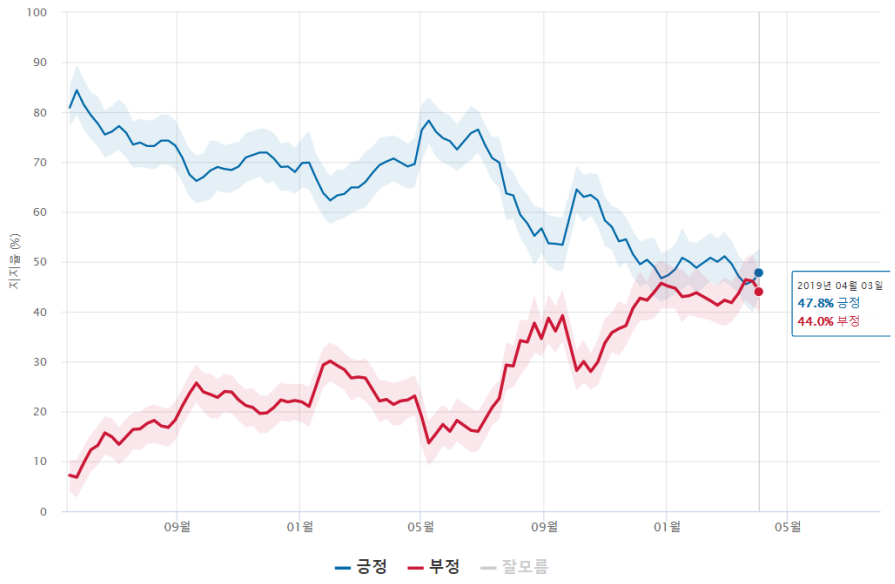
여론조사 공표 금지 직전 최종 vs. 실제 결과

서울대 pollab 지지율 지수

	여론조사 공표 금지 직전	실제결과	오차
문재인	41.1%	41.1%	0.0%p—
홍준표	18.1%	24.0%	5.9%p▲
안철수	20.8%	21.4%	0.6%p▲
유승민	5.1%	6.8%	1.7%p▲
심상정	8.6%	6.2%	2.4%p▼

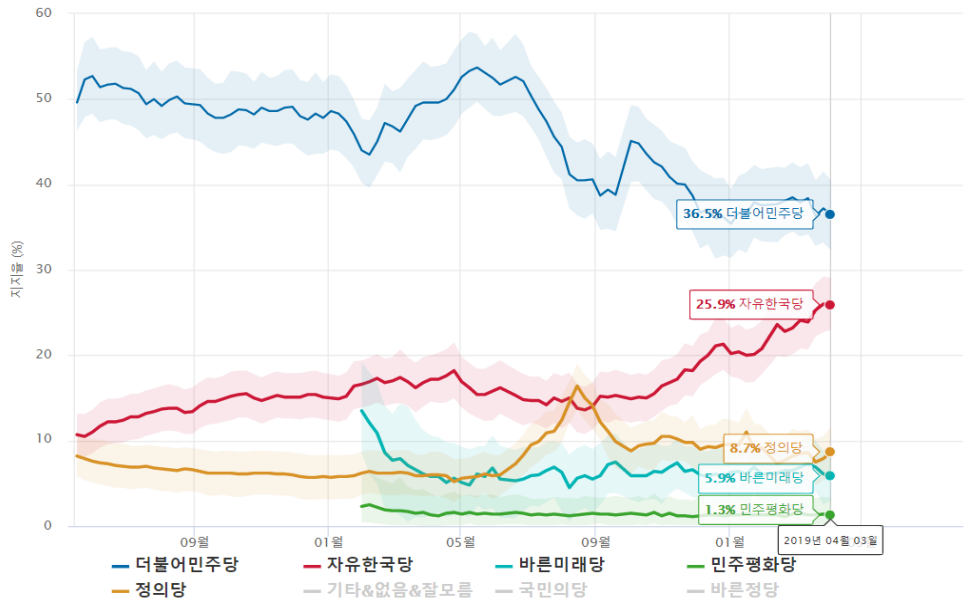
13

대통령 국정운영 평가



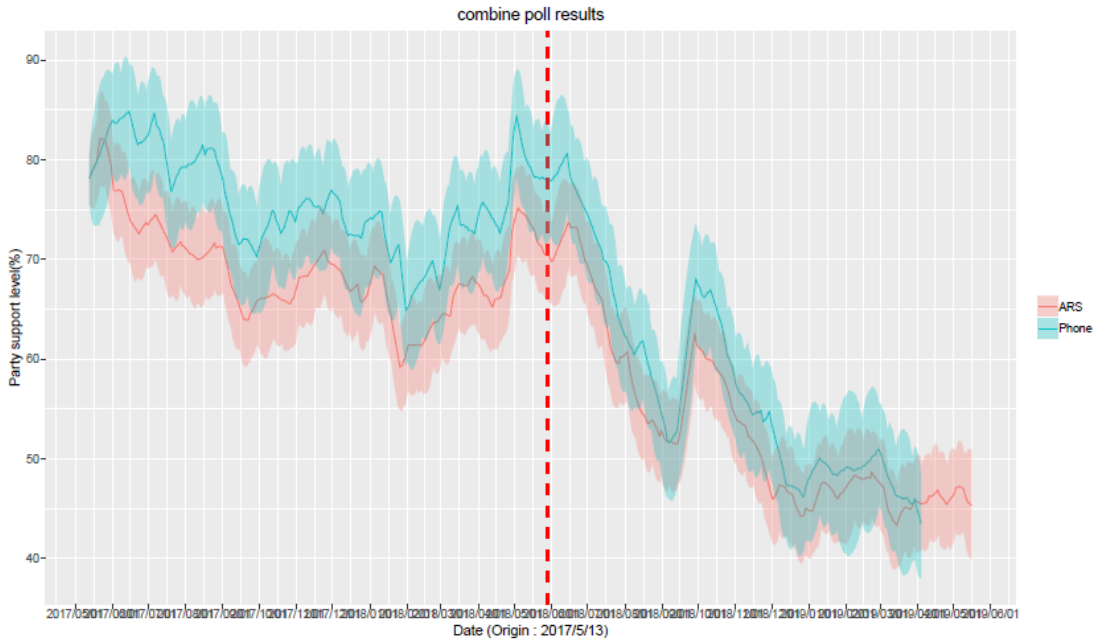
14

정당 지지율 평가



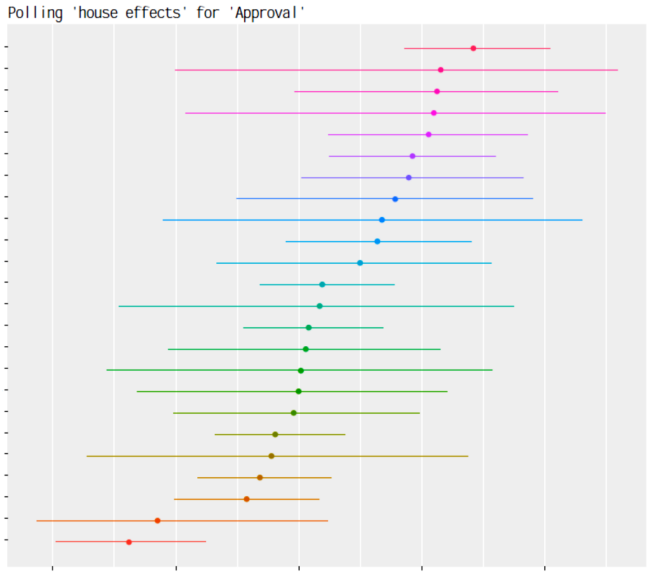
15

대통령 지지율 조사: ARS vs. 면접조사



16

대통령 지지율 조사: 조사기관 효과 (House Effects)



17

텍스트 빅데이터 분석

18

우리동네 공약지도 (중앙선거관리위원회): 언론보도 + 지방의회 회의록

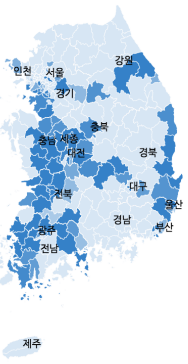
언론기사 분석 | 지방의회 회의록 분석

전국

#교육 1위 | #소방서 2위 | #청소년 3위 | #안전 4위 | #학교 5위 | #경찰 6위 | #학생 7위 | #조류독감 8위 | #농협 9위 | #보건소 10위 | #아파트 11위
#기업 12위 | #장애인 13위 | #일자리 14위 | #국비 15위 | #중소기업 16위 | #메르스 17위 | #어린이 18위 | #전통시장 19위 | #중국 20위 | #농기센터 21위
#여성 22위 | #어르신 23위 | #화재 24위 | #구제역 25위 | #고등학교 26위 | #건강 27위 | #관광객 28위 | #장학금 29위 | #관광 30위 | #환경 31위
#농업 32위 | #봉사활동 33위 | #가족 34위 | #청년 35위 | #국회 36위 | #문화재단 37위 | #저소득 38위 | #혁신도시 39위 | #학부모 40위 | #폭염 41위
#소년체전 42위 | #가을 43위 | #안전점검 44위 | #노조 45위 | #어린이집 46위 | #대학생 47위 | #도민체전 48위 | #아이 49위 | #KTX 50위

시·도별 비교 (KTX)

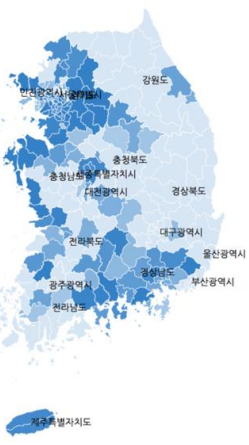
구·시·군별 비교 (KTX)



지역별 이슈 지도

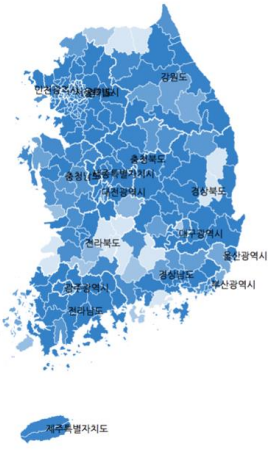
"미세먼지"

구·시·군별 이슈 지도 (미세먼지)



"일자리"

구·시·군별 이슈 지도 (일자리)



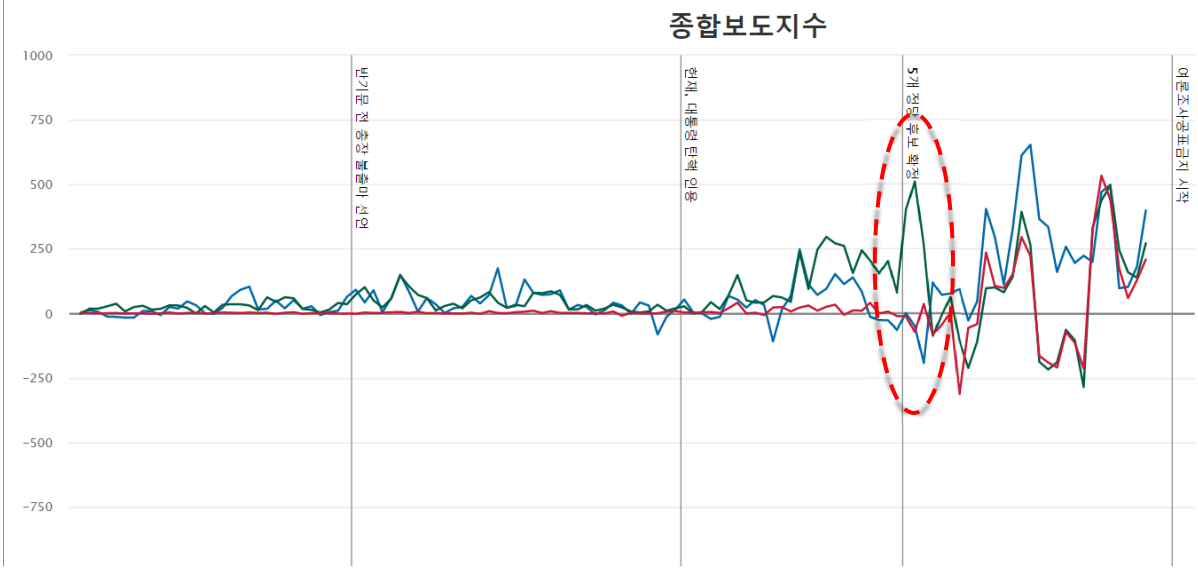
"창조경제"

구·시·군별 이슈 지도 (창조경제)

2017 대통령 선거 pollab보도지수

언론 논조 분석 (기계학습)

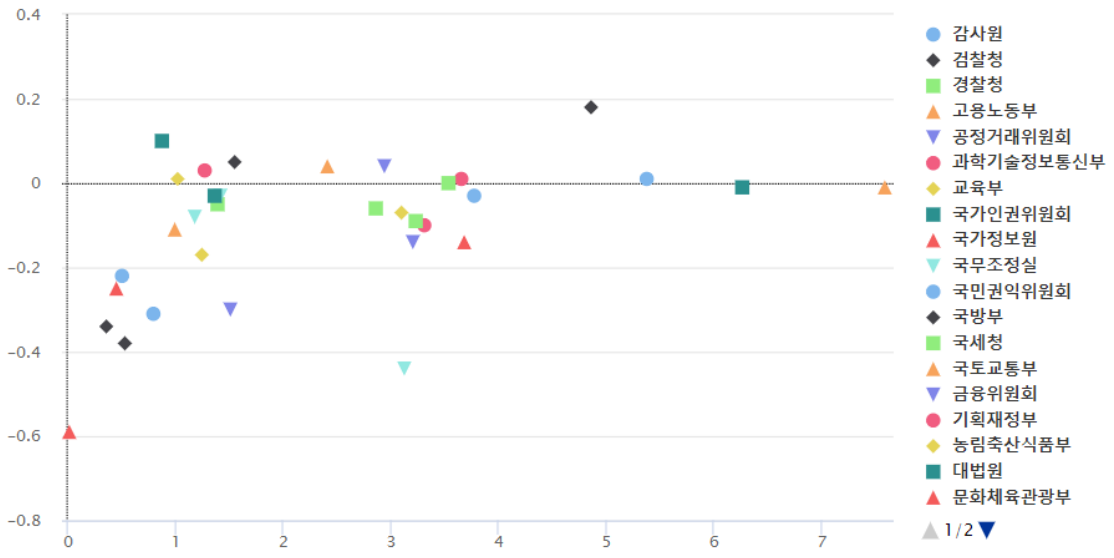
pollab



21

33개 공공기관 신뢰도 조사

공공기관 언론 논조 분석 (기계학습)

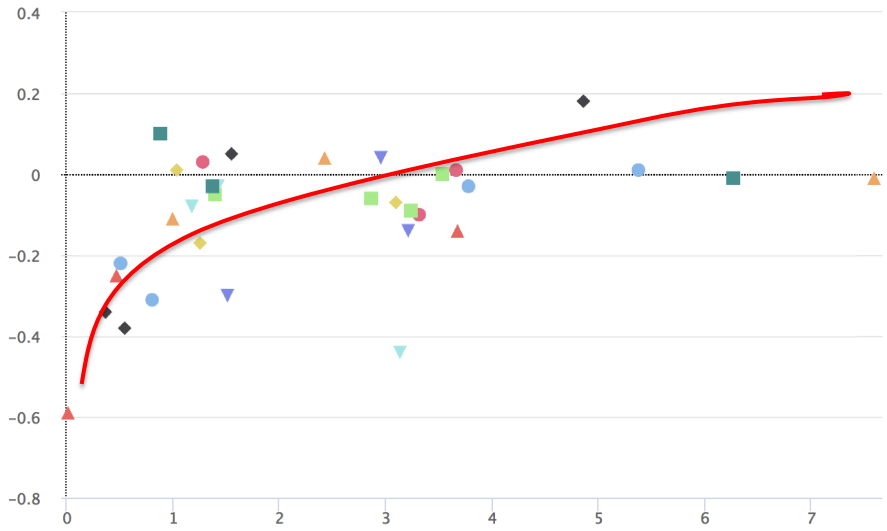


22

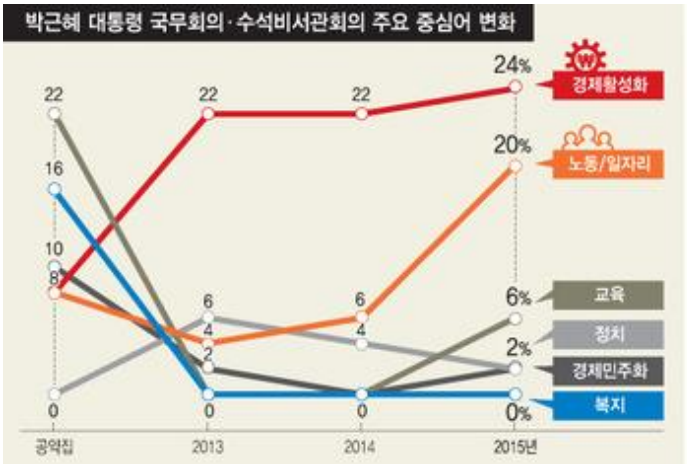
33개 공공기관 신뢰도 조사

공공기관 언론 논조 분석 (기계학습)

공공기관 언론보도 논조 – 언론논조와 공공기관 평가



박근혜 대통령 회의 및 신년사 분석

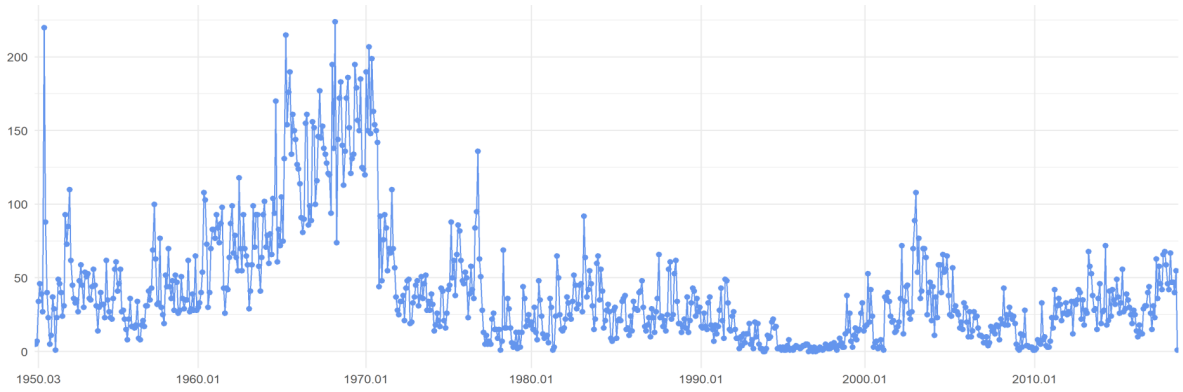


박근혜 대통령 회의 및 신년사 분석

25

북한 노동신문 및 조선중앙통신 분석

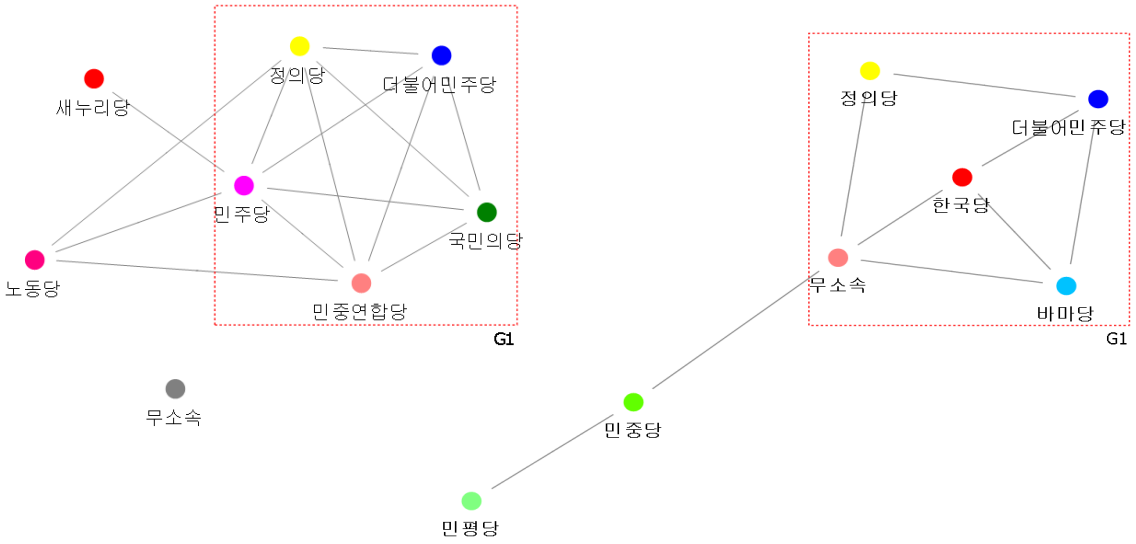
미국에 대한 노동신문 논조



26

국회의원선거, 지방선거 TV 토론 분석

후보자 및 정당 간 네트워크



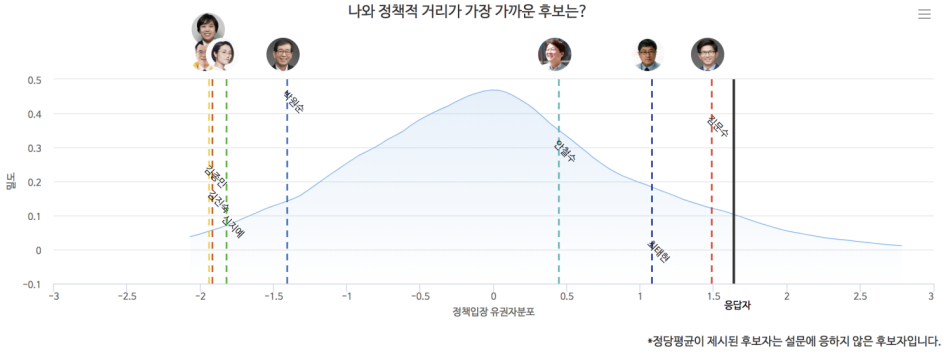
27

나에게 딱 맞는 후보찾기

28

“나에게 딱 맞는 후보찾기”

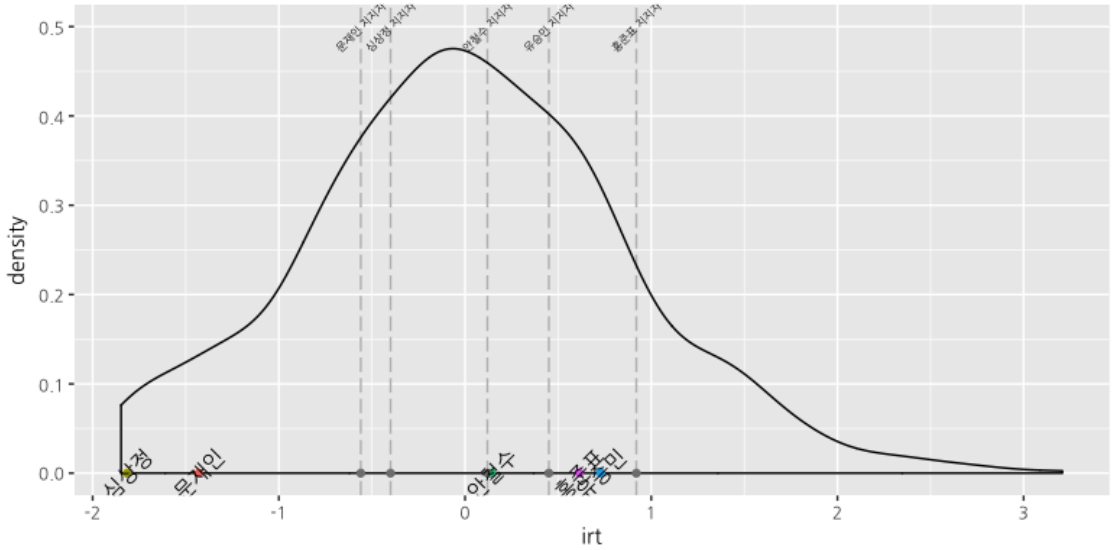
“나에게 딱 맞는 후보찾기”



*정당평균이 제시된 후보자는 설문에 응하지 않은 후보자입니다.

31개 설문으로 추정한 귀하의 정책입장 점수는 **1.64**이며,
서울유권자 100명 중 귀하는 **95**번째로 진보적, **5**번째로 보수적입니다.

2017년 대선: 정책투표

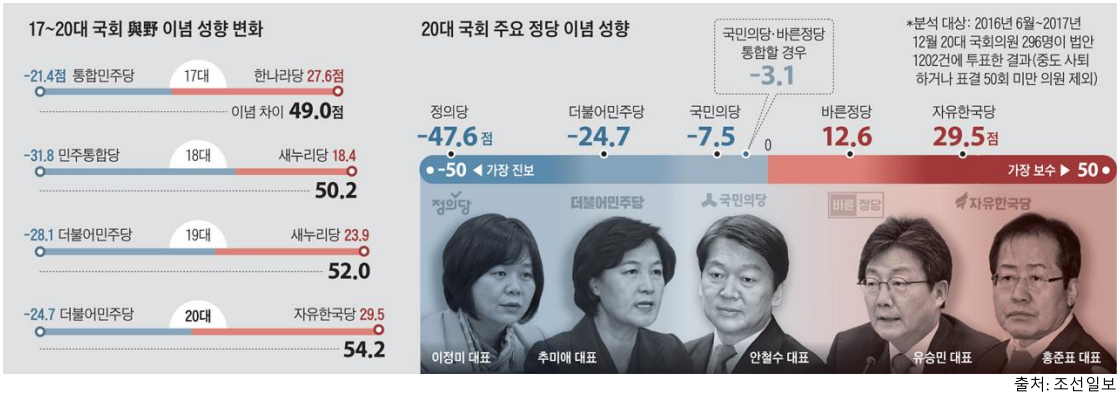


31

국회 표결 및 법안 발의 분석

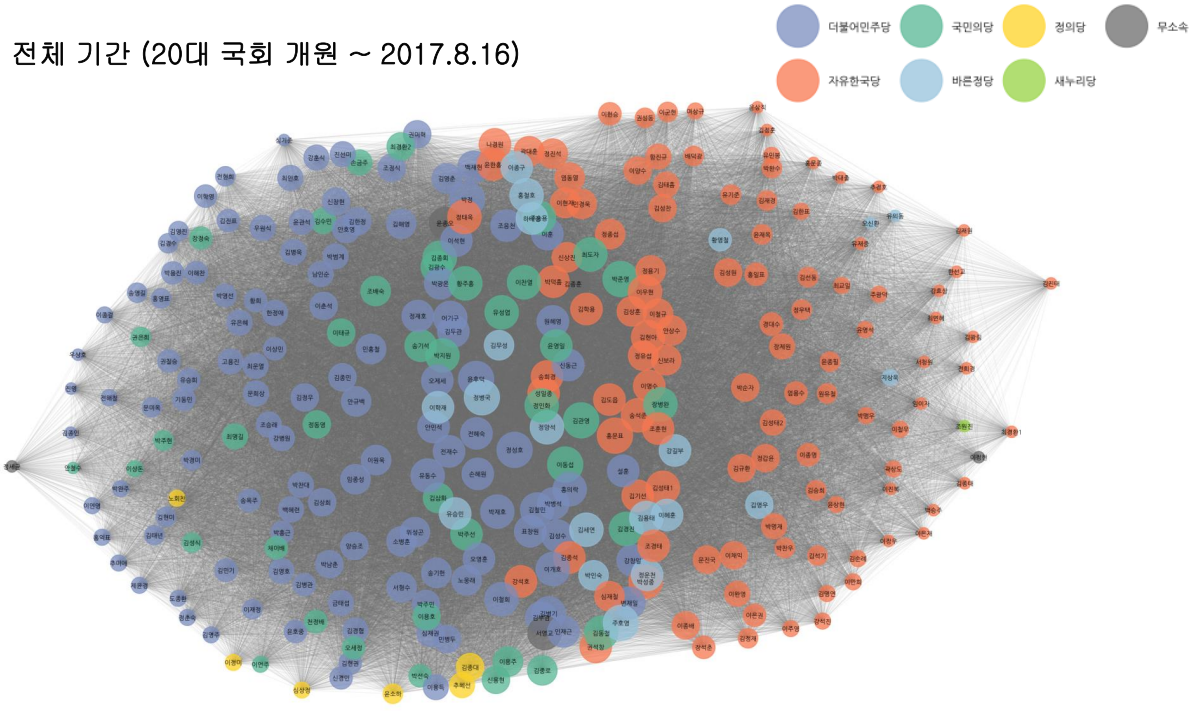
32

국회 표결 결과 분석



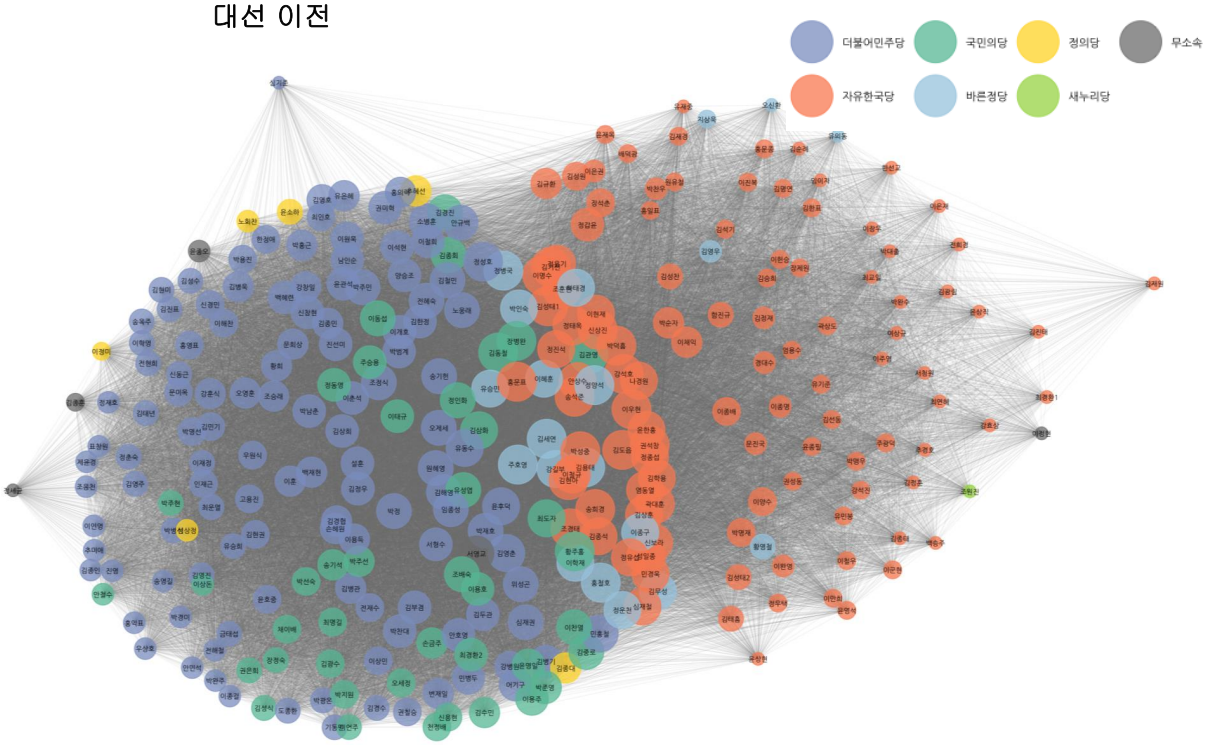
출처: 조선일보

공동 발의 분석



공동 발의 분석

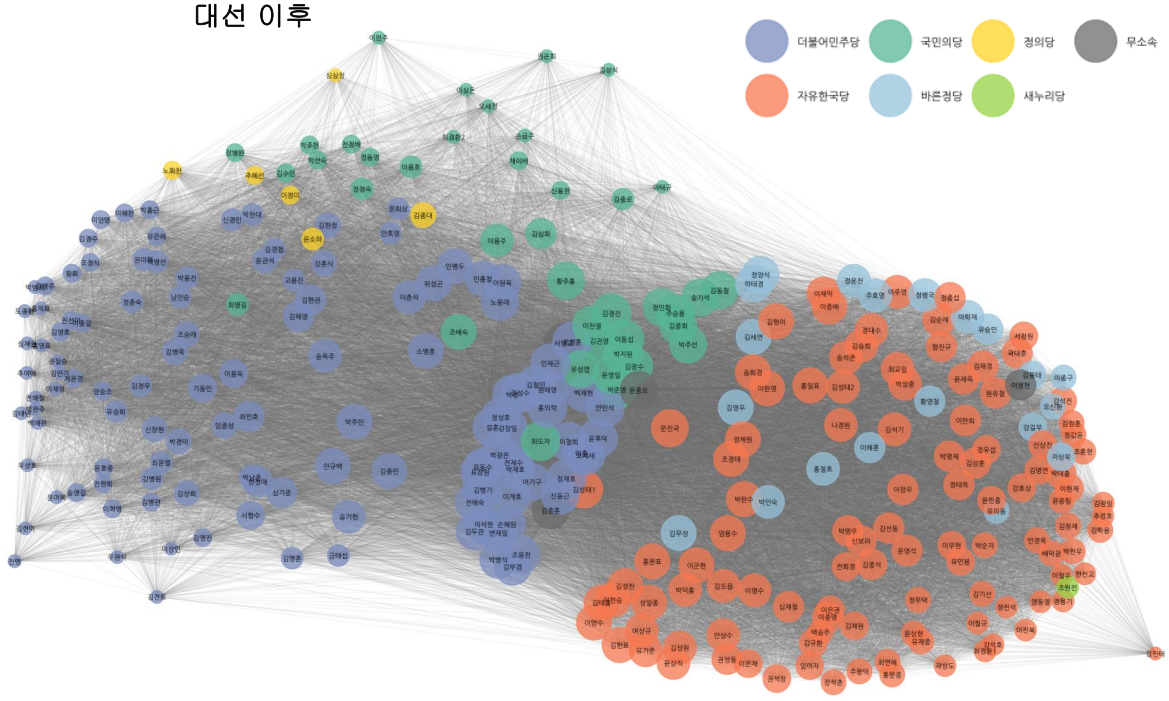
대선 이전



35

공동 발의 분석

대선 이후

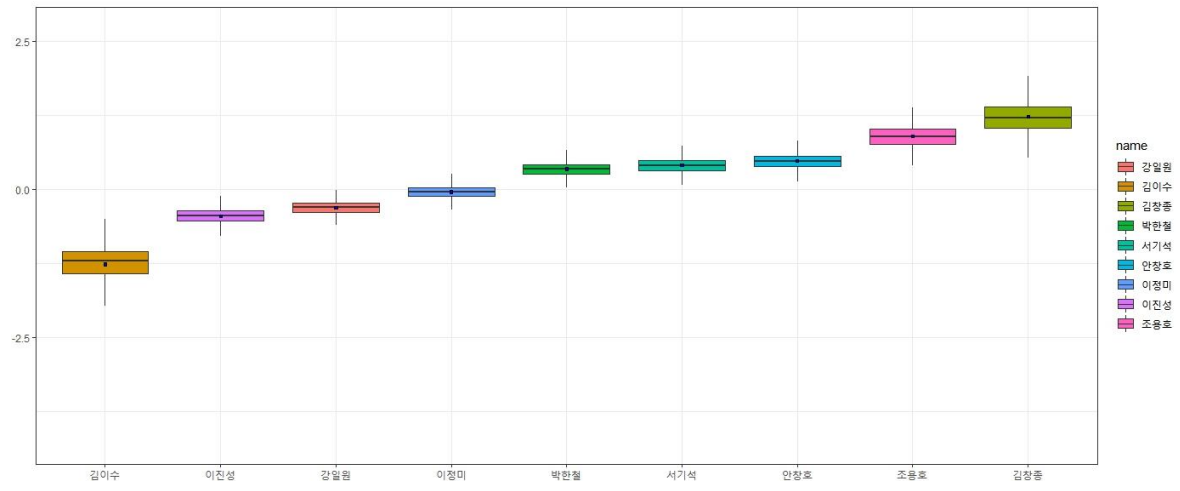


36

법원 판결 분석

37

헌법 재판관 분석



38

헌법 재판관 분석

박근혜 대통령 탄핵 심판 헌법재판관 구성 및 성향

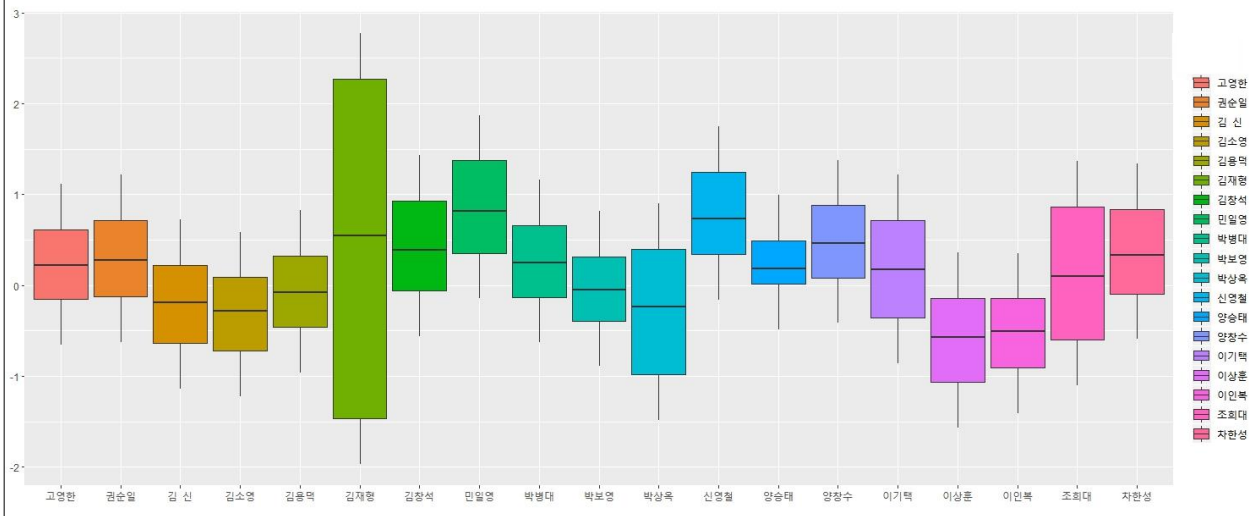
김이수 (63)	이정미 (54)	강일원 (57)	서기석 (63)	이진성 (60)	박한철 (63)	안창호 (59)	조용호 (61)	김창종 (59)
전남고, 서울대 사법연수원장 민주통합당 2018.9.19	마산여고, 고려대 대전고법 부장판사 이용훈 전대법원장 2017.3.13	용산고, 서울대 서울고법 부장판사 여야 합의 2018.9.19	경남고, 서울대 서울중앙지법원장 박근혜 대통령 2019.4.14	경기고, 서울대 광주고법원장 양승태 대법원장 2018.9.19	제물포고, 서울대 서울동부지검장 이명박 대통령 2017.1.31	대전고, 서울대 서울고검장 새누리당 2018.9.19	중앙고, 건국대 서울고법원장 박근혜 대통령 2019.4.14	영신고, 경북대 대구지법원장 양승태 대법원장 2018.9.19

출처: KBS

진보적 ← → 보수적

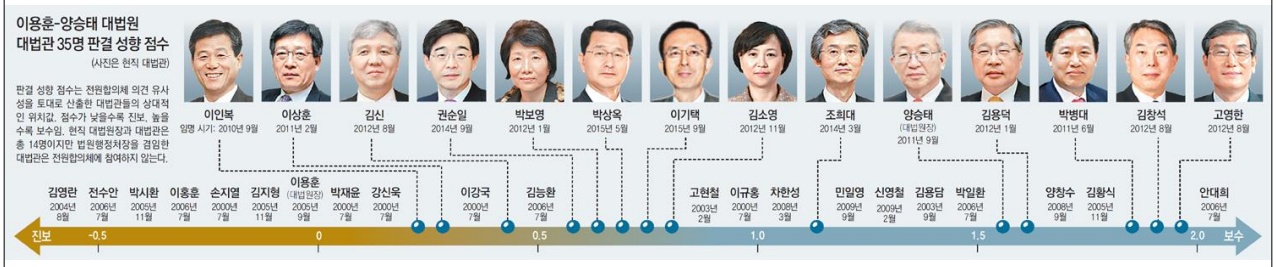
대법원 전원합의체 판결 분석

대법관 35명 판결 성향 점수



대법원 전원합의체 판결 분석

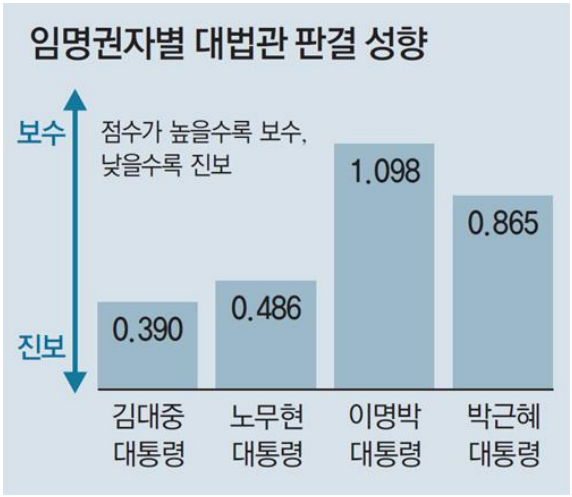
대법관 35명 판결 성향 점수



출처: 동아일보

대법원 전원합의체 판결 분석

임명권자별 대법관 판결 성향



출처: 동아일보

pollab
An Analytic Solution to Strategic Communication

감사합니다

한규섭 (서울대 언론정보학과)
kyuhahn@snu.ac.kr

노선혜 (서울대 언론정보학과)
shno205@snu.ac.kr

제7회

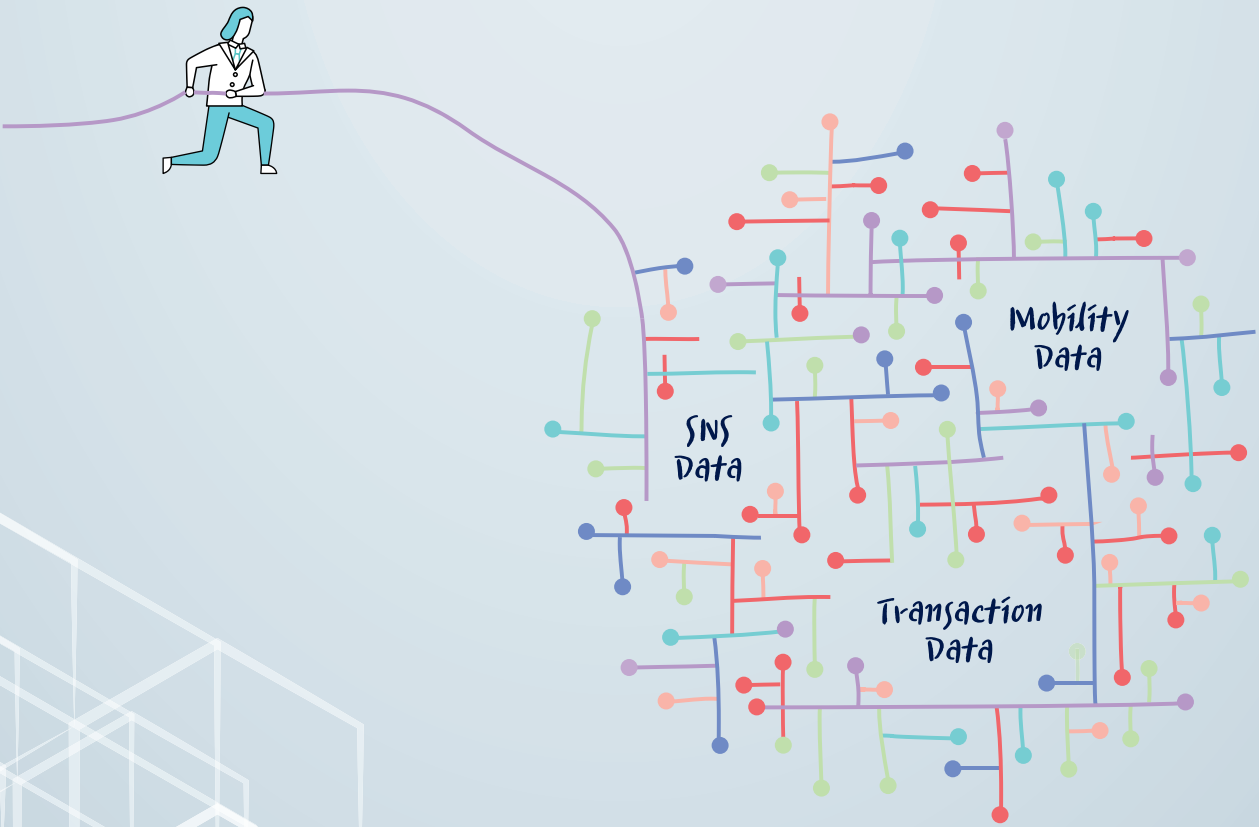
KOSSDA

데이터 페어

2부: 빅데이터 활용 사례

문학권력에 대한 세 가지 데이터 분석

이원재 교수 (카이스트 문화기술대학원)



문학권력에 대한 세 가지 데이터 분석

사회, 시장, 문학, 그리고 통계

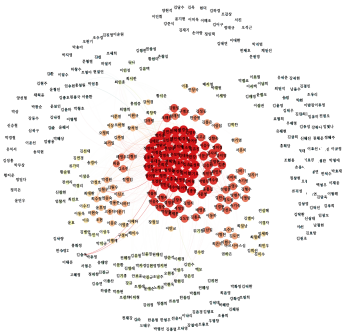
이원재 KAIST @ KOSSDA 데이터페어 062718

KAIST Graduate School of Culture Technology

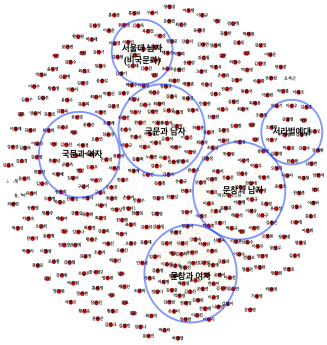
사회

(인연 기반 네트워크)

KAIST Graduate School of Culture Technology



=10.2+0.53



KAIST Graduate School of Culture Technology

統計

KAIST Graduate School of Culture Technology

[장정일 칼럼] '지하철 시(詩)'와 문학권력

신경숙 사태 이후 문학권력 논쟁이 분분했지만, 지금까지 논해진 것들은 아무것이나 넣으면 만두 속이 되는 것처럼 편의적이었다. 예컨대 10월 29일치 한국일보 기사(▶ 숫자로 확인된 문학권력의 실체)로도 소개된 전봉관 · 이원재 · 김병준의 정량적 · 통계적 연구는 문학권력에 대해 아무 것도 말해 주지 못하는 '통계의 마술'일 뿐이다. 거대 출판사의 상업주의 전략을 문학권력으로 적시하고 나면, 지하철 같

KAIST Graduate School of Culture Technology

염종선. 2015. "창비를 둘러싼 표절과 문학권력론 성찰." 43(4):66-92.

문학권력 관련 연구논문에 대하여

최근 카이스트의 한 연구진은 「문예지를 매개로 한 한국 소설가들의 사회적 지형」이라는 논문을 발표했고 일부 일간지에서 이에 관해 보도한 바 있다.[31] 1994년부터 2014년까지 21년간 소설을 대상으로

KAIST Graduate School of Culture Technology

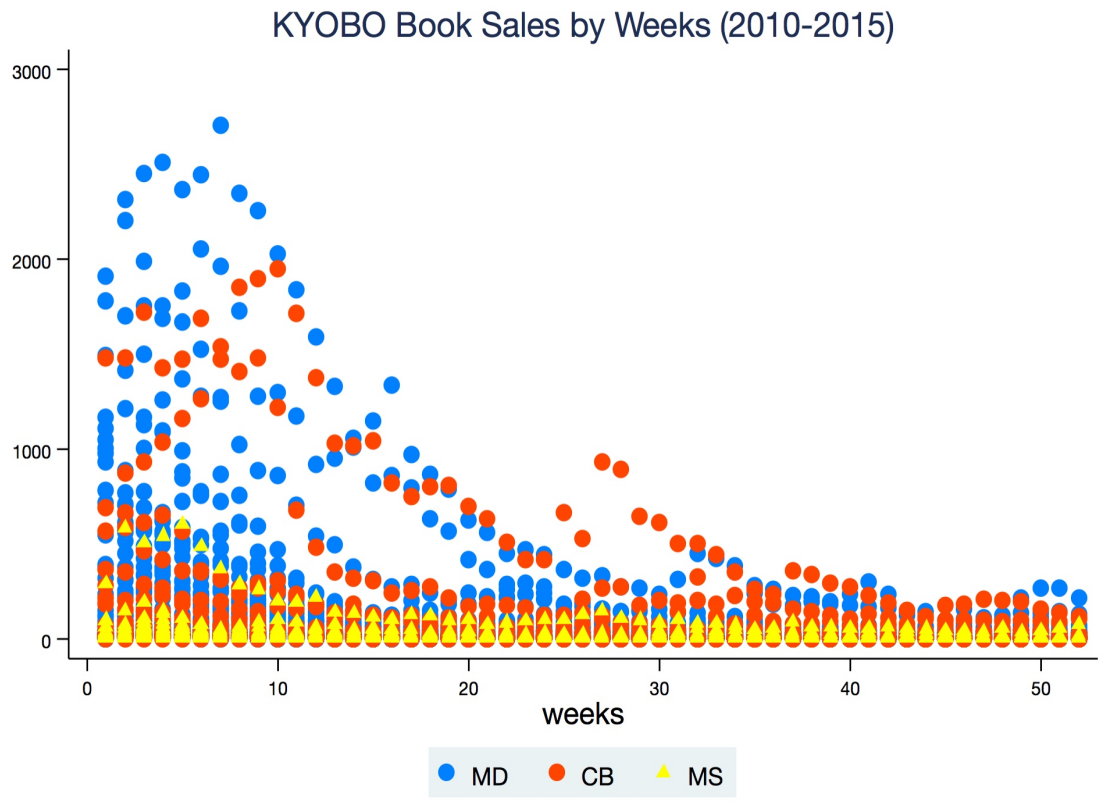
3대 출판사에 관한 내용이 섞여 있어서 일괄적으로 말하기 곤란하지만, 창비에 국한한다면 자사 출신 작가라고 편파적으로 밀어준다는 통념은 사실이 아님이 밝혀졌다. 또한 다른 각도에서 살펴보면, 논문은 "3대 계간지 중복 '주도작가' 총 26명"[32]의 명단을 밝히고 이들이 3대 계간지에 작품을 게재하거나 비평적으로 언급되고 단행본 출판을 한 현황을 적고 있는데, 문예지들에서 당대의 주요 작가들에 발표지면을 주고 비평적으로 언급한 것이 왜 문제인지는 밝히지 않았다. 상업적 동기에서 엉터리 작품들을 밀어주었다면 문제이겠지만 그에 대한 언급이 없는 것은 양적인 분석방법의 한계로 여겨진다. 더군다나 "3대 계간지 모두에 중복된 주도작가가 26명이나 된다는 것은 그만큼 소설 분야만큼은 3대 계간지가 획일화되었음을 의미한다"[33]라는 단언은 무슨 뜻인지 알 수 없다. 지난 21년간 얼마만큼의 모집단 수 중에서 26명 정도면 '획일화'를 확정할 수 있는 수준이라는 것인가.[34]

KAIST Graduate School of Culture Technology

시장

(교보문고 단행본 판매량)

KAIST Graduate School of Culture Technology

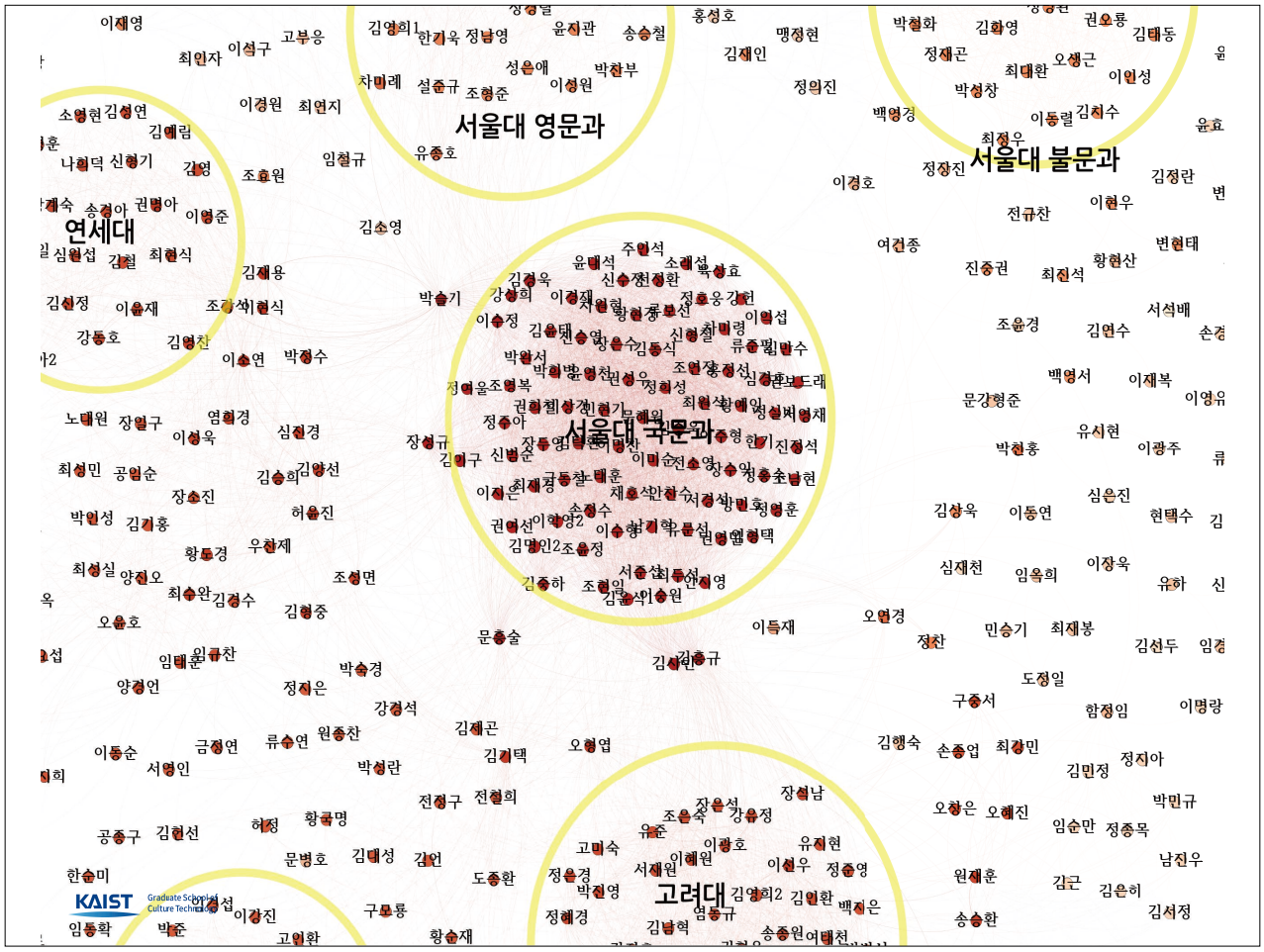


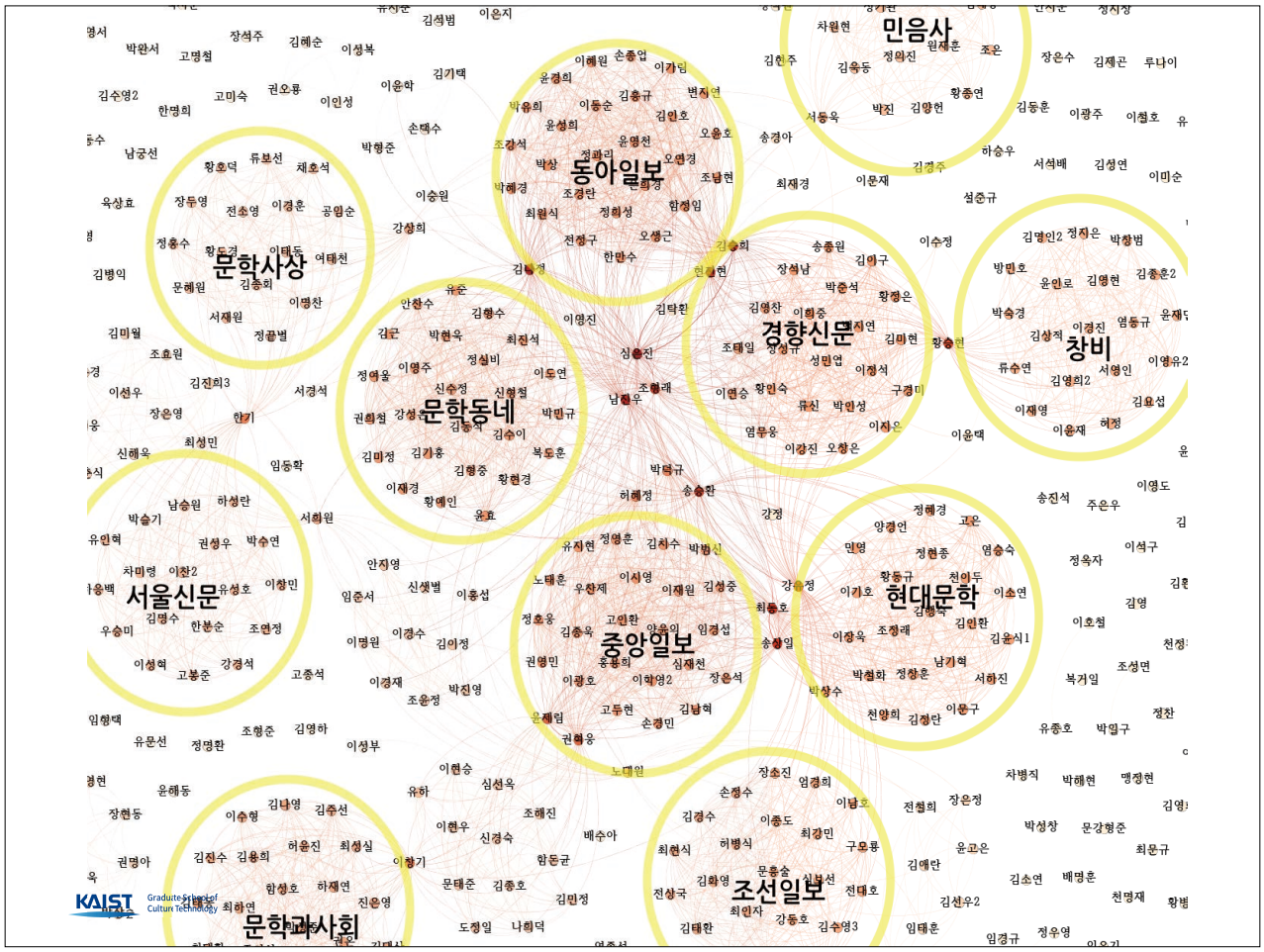
Source: KYOBO BOOK CENTRE 2016

KAIST Graduate School of Culture Technology

$$Y_{k,w+1} = Y_{k,w} + \mathbf{S}_{k,w}\phi + \delta O_{ij} + \mathbf{W}_j\beta + \mu_k + \omega_k + \tau_k + \varepsilon_{k,w+1}$$

KAIST Graduate School of Culture Technology


서울대 영문과
서울대 불문과
연세대
서울대 국문과
고려대
KAIST Graduate School of Culture Technology


민음사
동아일보
문학사상
경향신문
창비
문학동네
서울신문
현대문학
중앙일보
조선일보
문학과사회
KAIST Graduate School of Culture Technology

일주일 후의 판매량 증가분을 예측하는 시계열 고정효과 모형 (이원재, 김병준, 전봉관 2016)

	(1)	(2)	(3)
종속 변수: 판매량 증가분 (다음 주(週))			
판매량 증가분 (현재 주)	0.044***	0.042***	0.042***
	(0.007)	(0.007)	(0.007)
판매량 (현재 주)	-0.077***	-0.076***	-0.076***
	(0.002)	(0.002)	(0.002)
블로그 수 (현재 주)	-0.117	-0.113	-0.122
	(0.095)	(0.095)	(0.095)
언론 기사수 (현재 주)	0.753**	0.735**	0.746**
	(0.237)	(0.237)	(0.237)
창비 평론수 (현재 주)		-19.509***	-29.243***
		(3.030)	(4.803)
문사 평론수 (현재 주)		-2.818	-7.188
		(2.696)	(4.294)
문동 평론수 (현재 주)		2.918	-0.356
		(2.115)	(3.423)
작가-창비비평가 인구학적 유사성			10.194**
			(3.924)
작가-문사비평가 인구학적 유사성			4.322
			(3.399)
작가 문동비평가 인구학적 유사성			2.952
			(2.433)

KAIST Graduate School of Culture Technology

문학

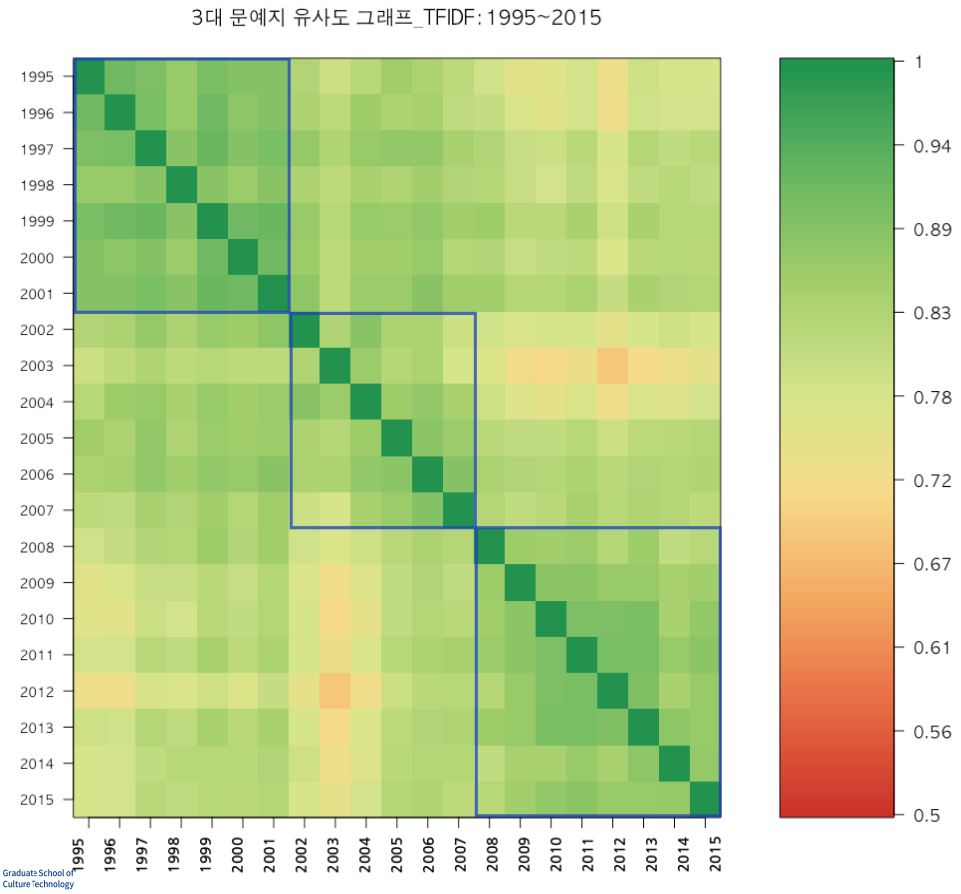
(개념어 추출)

KAIST Graduate School of Culture Technology

From the practical viewpoint, it is immediately apparent that **superhuman intelligence** and intuition are necessary to isolate this entire aggregate of structural relations (the expression of which is essential for understanding the works in question) simply by means of textual studies, no matter how profound and prolonged they may be.

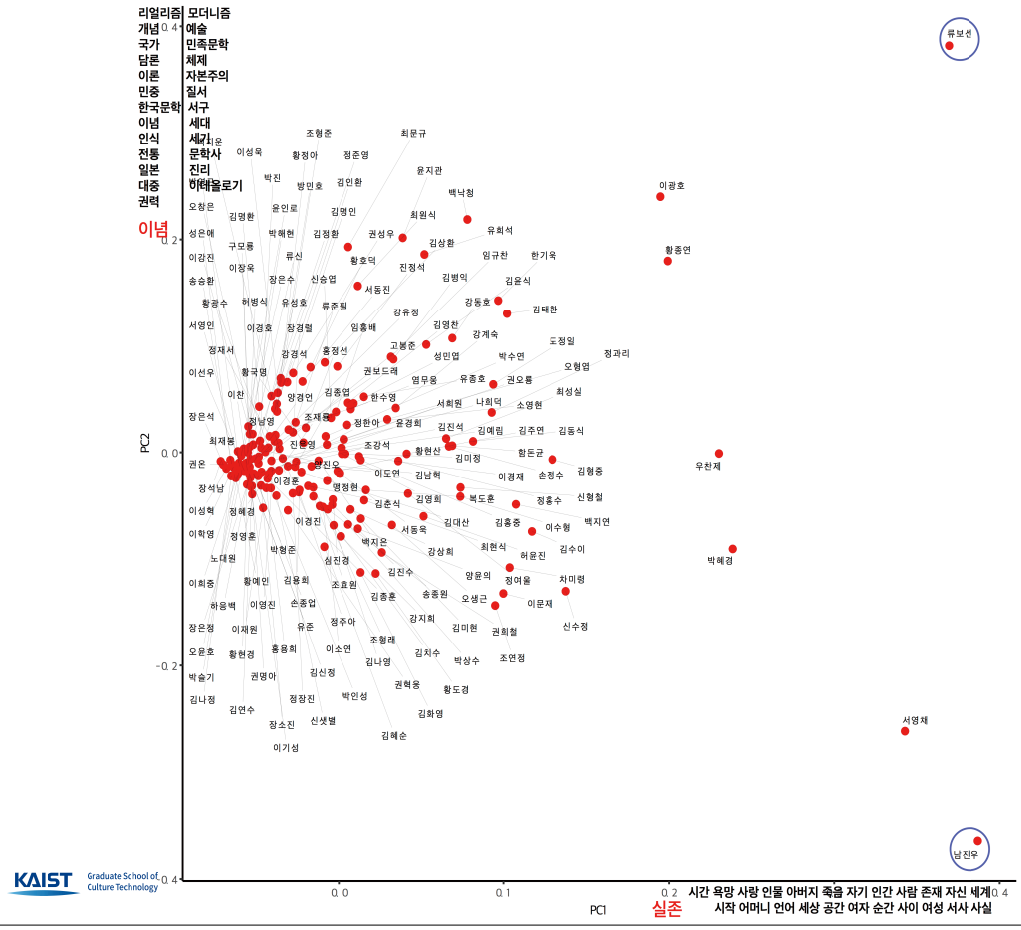
Lucien Goldman 1980

KAIST Graduate School of Culture Technology



KAIST Graduate School of Culture Technology

개념어 사용에 따른 비평가들의 유사성 분포 - 3대 문예지, 1995-2015



개념어 사용에 따른 비평가들의 유사성 분포 - 3대 문예지, 2008-2015