



**INTRO** 

TITLE

TEAM

BACKGROUND

**PROCESS** 

ROAD-MAP

CRAWLING

**PROCESSING** 

**ANALYSIS** 

**EXIT** 

**INSIGHT** 





2

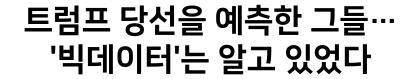
<u>빅데이터로</u> 분석해보는 2012년도 대선

"누가 王이 될 상인가"

신동원, 유하진, 임훈, 정다운, 홍승원



# DATA & FUN STOR

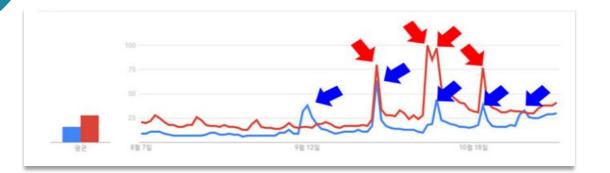


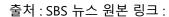
미국의 45대 대통령으로 도널드 트럼프가 당선되면서 미국은 물론 전세계 언론들이 이른바 '멘붕'이다. 여론조사를 토대로 한 대부분 미국 언론사의 선거 결과 예측 시스템들은 힐러리 클린턴의 당선 가능성을 90% 이상으로 점쳤지만, 결과는 반대로 나왔다.



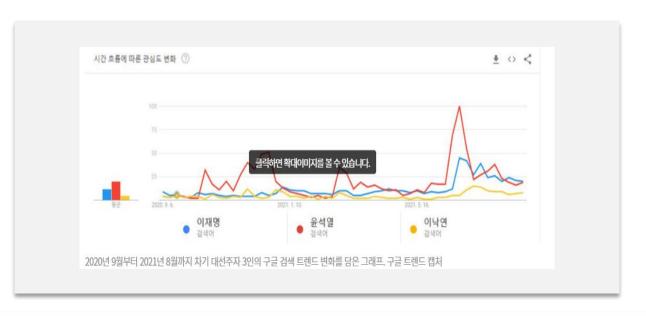
우 교수 개인적으로는 트럼프가 마음에 들지 않았지만, 빅데이터 분석결과는 트럼프의 승리로 나왔다는 것이다. 예상 득표율은 트럼프 54~52%, 힐러리 48~46%, 선거인단 수는 트럼프 285~275, 힐러리 263~253이였다.







# 화두로 떠오르는 BIG DATA



#### [대선후보빅데이터분석] 총 언급량...윤석열 19만건, 이재명 15만건

음 천영준시사저널 컬럼니스트 (chunyg@sisajournal.com) □ ② 승인 2021.03.13 08:00 □ 로수 1639

□홈 □ 오피니언 □ 기고

#### 빅데이터를 통해 대선을 예측할 수 있을까?

□ 전인호 (주)얌전한고양이 대표이사 □ 입력 2021.08.28 17:21 □ 수정 2021.08.30 10:01 □ □ 댓글 1

ELECTION & PREDICTION

[스피치 인사이트] 여야 빅2, 이재명·윤석열 후보 언론 및 소셜활동 비교







# DATA VISUALISATION

# 빅데이터가 가져온 정치 패러다임의 변화

**Employee Designation** 

4차 산업혁명시대 IT기술의 발전과 확산에 따른 소셜미디어의 진화와 빅데이터의 정치적 상호작용은 정당, 정치인, 유권자 모두의 정치적 인식과 행위에 영향을 줌으로써 정치과정의 패러다임을 변화시키고 있다.

#### 출처

1) Norris, P. Democratic Phoenix: Reinventing Political Axtivism, Cambridge, UK: Cambridge University Press, 2002; Norris, P. Preaching to the Converted?: Pluralism, Participation and Party Websites, Party Politics. 9(1), 2003, 21-45쪽.; Norris, P. Political Activism: New Challenges, New Opportunities, In The Oxford Handbook of Comparative Politics, Boikes, C. & Stokes, S.(eds.). Oxford: Oxford University Press, 2007.; 이원태 외 3인, 디지털 컨버전스 시대 미디어 플랫폼의 진화와 정치참여 연구: 디지털 컨버전스 기반 미래연구 시리즈, 정보통신정책연구원, 2009, 73-74쪽.



# 빅데이터와 대선

데이터 전략팀을 운영한 오바마

2012년 재선을 위한 오바마 캠프는 IT조직을 꾸렸는데, 짐 메시나(Jim Messina) 선거운동본부장을 총지휘자로 <u>'테크놀로지팀' '데이터 분석팀' '뉴미디어팀(디지털팀)'</u>으로 구성

테크놀로지팀

데이터분석팀

뉴미디어팀



- 선거 마지막 주 경합 주 친구에게 메시지 전송유도

->메시지 수신자 중 20%가 요청 수락40)

# ANALYS

#### 출처

Hispanic Business.com, Nov.15, Obama's Victory Is a Win for Big Data.; Harvard Business Review Blog Network, Nov.13, 2012: The First Big Data Election.; 국민일 보, 2012. 11. 8자, 빅데이터 선거전략 빅히트 시카고사단의 승리.

http://blogs.itmedia.co.jp/business20/ (검색일: 2019.7.15.)

CIO Korea, 2012. 11. 15자. 버락 오바마의 승리, 그 뒤에는 빅데이터가 있었다



**ROAD-MAP** 

**CRAWLING** 

**PROCESSING** 







DO VOTE! 2022

# ANALYSIS

# ROAD-MAP



INSIGHT 힐러리와 트럼프 대선

#### 각 후보의 언론 노출 분석

뉴스 헤드라인에 노출된 데이터 양은 지지도와 관련이 있을까?

INSIGHT 빅데이터와 키워드

#### 연관 단어 구성 분석

언론의 노출된 단어와 지지도는 연관이 있을까?

INSIGHT 오바마 대선

#### 유권자의 성향 분석

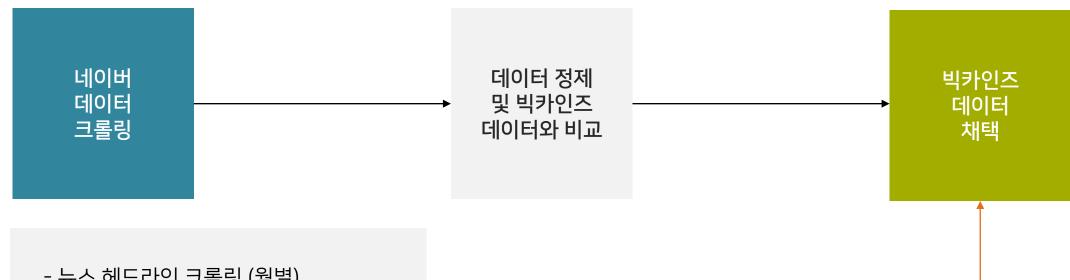
유권자의 성향은 지지도&언론 노출량와 관련이 있을까?





## CRWALING

네이버 키워드 크롤링 & 빅카인즈



- 뉴스 헤드라인 크롤링 (월별)
- 6~12월 데이터 통합

기 간: 2012년 6월 1일 - 2012년 12월 17일

키워드: 문재인 or 박근혜

- 데이터 총량 -> 노출되는 뉴스 헤드라인 기준

- 각 후보 연관 검색어 -> 전체 아티클 기준

털 : 네이버

- \*중복 기사는 링크 기준으로 제거
- \*단어가 헤드라인에 없으나 본문에 있는 경우 > 결측
- \*네이버 크롤링 수 제한으로 기간별 데이터량과 총 기간별 데이터 량의 차이 발생





# **PROCESSING**

연관 단어 구성 분석

3만개 기사 키워드 데이터 중복값 제거 INDEX 키워드 추출 후 COUNT 최빈 키워드 10개 추출 (이상값 제거) 대선 후보자 대표 키워드 추출

#### 전처리 과정

- 키워드 인덱스 제작
- :SET 변수 생성 후 키워드 입력
- 키워드 COUNTING
- : INDEX를 기사에 대입-> 중복값 제거
- 많은 키워드 작업량:

8만여개 \* 기사 3만개 = 24억

# **PRECESSINIG**

대선후보 지지도 데이터 전처리 과정

1

#### **POWER BI**

PDF 파일->데이터 변환

) 대선	후보 지지도 - 대	나자 구도	(자유응	답)								◎ 양지	구도 🧳	<u>/</u>
201	2년 7월 1~4주	표본수	김두관	김문수	문재인	박근혜	손학규	안철수	정몽준	기타	모름 없음	박근혜	문재인	모름 없음
	전체	6,100	2%	1%	13%	37%	2%	23%	1%	1%	20%	48%	32%	20%
지역별	서울 인천/경기 강원 대전/충청 광주/전라	1,270 1,733 187 617 626	2% 1% 2% 2% 6%	2% 2% 1% 1%	13% 13% 14% 11%	33% 33% 43% 42% 12%	2% 2% 1% 2% 5%	25% 24% 16% 25% 34%	1% 1% 1% 0%	1% 1% 2% 1% 2%	21% 22% 21% 17% 23%	47% 47% 48% 52% 21%	33% 32% 27% 30% 56%	20% 21% 25% 18% 23%
	대구/경북 부산/울산/경남 제주 남성	632 968 67	1% 2% 2% 3%	1% 1% 1% 1%	8% 12% 14%	59% 45% 45%	0% 1% 1% 2%	13% 20% 17% 25%	1% 1% 0%	0% 1% 0%	16% 17% 20%	66% 56% 53%	16% 25% 33%	18% 19% 14%
성별	여성 여성	3,020 3,080	1%	1%	11%	35%	1%	25%	1%	1%	23%	46% 50%	35% 29%	19% 21%
연령별	19~29세 30대 40대 50대 60세 이상	1,130 1,256 1,346 1,146 1,223	1% 2% 3% 3% 2%	1% 1% 2% 2% 1%	18% 21% 13% 7% 4%	22% 22% 34% 49% 59%	1% 1% 2% 3% 2%	36% 32% 26% 14% 7%	0% 1% 1% 1% 0%	1% 1% 2% 1%	20% 19% 18% 20% 23%	37% 35% 45% 60% 64%	39% 44% 38% 22% 15%	24% 21% 17% 19% 21%
직업별	농/임/어업 자영업 블루칼라 화이트칼라 가정주부 학생 무직/기타	194 846 722 1,782 1,484 577 495	2% 4% 2% 2% 1% 0% 2%	1% 2% 1% 1% 1% 2% 1%	6% 11% 11% 18% 9% 19% 8%	49% 42% 41% 27% 45% 19% 50%	1% 2% 2% 2% 1% 1% 2%	15% 20% 22% 30% 17% 41% 10%	1% 1% 2% 1% 0% 0%	1% 1% 1% 1% 1% 0%	23% 18% 17% 18% 24% 19% 25%	56% 52% 52% 39% 55% 37% 59%	20% 29% 31% 42% 22% 41% 19%	24% 19% 17% 19% 23% 21% 23%
이념 성향별	보수 중도 진보 모름/의견없음	1,846 1,769 1,714 771	1% 2% 3% 1%	2% 1% 1% 1%	5% 14% 24% 4%	58% 33% 17% 39%	1% 2% 3% 1%	13% 29% 34% 13%	1% 1% 1% 1%	1% 1% 1% 1%	17% 18% 16% 40%	71% 45% 28% 45%	13% 35% 56% 16%	16% 20% 17% 39%
주요 지지 정당별	새누리당 민주통합당 지지정당 없음	1,994 1,343 2,440	1% 3% 2%	2% 1% 1%	2% 31% 9%	79% 10% 21%	1% 4% 2%	6% 36% 29%	1% 1% 0%	1% 1% 1%	9% 14% 34%	88% 17% 36%	4% 70% 29%	8% 13% 35%

3-	A <sup>B</sup> <sub>C</sub> Column1		A <sup>B</sup> C Column2	A <sup>B</sup> <sub>C</sub> Column3		A <sup>B</sup> <sub>C</sub> Column4
1		nuil	n n	et .	rul	대선 후보다자 구도
2		mil	조사시기	표분수		박근해
3	주간		06製 1平(30-03)	1,525		37%
4	주간 연구		08製 2平(06~10)	1,516		36%
5	<b>42</b>		08製 3平(19~17)	1,222		37%
6	425		08製 4草(20-24)	1,561		41%
7	주간		08월 5주(27~91)	1,555		40%
8	42		09활 1주(03~07)	1,521		40%
9	주간		09월 2주(10~14)	1,551		42%
10.	주간		09월 3주(17~21)	1,591		39%
11	추간		09월 4주(24~28)	1,577		39%
12	주간		10월 1주(01-05)	1.248		40%
13	<b>42</b> t		10월 2주(08~12)	1,551		40%
14	주간		10월 3주(15~18)	1,539		36%
15	<b>42</b> t		10個 4平(22~26)	1,561		37%
16	주간		11월 1주(29~02)	1,544		38%
17	주간		11월 2平(05-09)	1,534		39%
18	주간		11월 3주(12~16)	1,545		39%
15	22)		11#4Z(19-21)	1.541		1016

모름/없음

#### **EXCEL**

데이터 전처리 (전체추이, 월별 상세 데이터)

49% 12% 19% 전체 null 48% 41% 12% 47% 34% 19% 지역별 서울 44% 43% 13% 13% 49% 33% 17% 지역별 선물 44% 45% 42% 12% 56% 18% 25% 지역별 안전/경기 46% 42% 12% 55% 18% 29% 20% 지역별 강원 55% 34% 11% 51% 29% 20% 지역별 대전/세종/51% 37% 11% 19% 54% 27% 지역별 대전/세종/51% 36% 16% 16% 58% 16% 15% 지역별 부산/울산/56% 33% 10% 58% 28% 14% 지역별 부산/울산/56% 33% 10% 48% 16% 50% 35% 16% 연령별 19-29세 36% 48% 16% 50% 30% 21% 연령별 30-39세 33% 57% 10% 16% 50% 30% 21% 연령별 30-39세 33% 57% 10% 16% 50% 36% 48% 16% 연령별 40-49세 44% 45% 11% 36% 48% 16% 연령별 50-59세 60% 30% 10% 47% 36% 17% 연령별 50-59세 60% 30% 10% 61% 23% 16% 연령별 남성 46% 43% 12% 57% 12% 21% 성별 남성 46% 43% 12% 57% 12% 21% 성별 남성 46% 48% 38% 12%

전체주이 전체열별 pro\_6 pro\_7 pro\_8 9월 10월 11월 12-1월 12-2

9 null

- 1. 전체 후보 지지율
- 2. 특성별 지지율 변화(월 단위)





### **PRECESSINIG**

대선후보 지지도 데이터 전처리 과정

# 한국갤럽<sup>ᆐᄼᆀ</sup>정치지표

Political index

제25~28호 • 2012년 7월 통합

- 역대 대통령 분기별 직무 수행 평가 1988~2012
- 주요 월간 지표
- 정당 지지도
- 2012 대통령 선거 후보 지지도
- 최근 6개월 지표 비교 2012년 7월 1~4주(2~27)

휴대전화 RDD 조사 평균 응답률 19% 유효표본 전국 성인 6,100명 표본오차 ±1.3%포인트(95% 신뢰수준)

#### **PICK DATA**

"올해 12월에는 대통령 선거가 있는데요, 귀하는 누가 다음 번 대통령이 되는 것이 좋다고 생각하십니까? 그럼 누구에게 조금이라도 더 호감이 가십니까? 양자 간에는 누가 대통령이 되는 것이 더 좋다고 생각하십니까?"

#### ◎ 대선 후보 지지도 - 다자 구도(자유응답)

201	2012년 7월 1~4주		김두관	김문수	문재인	박근혜	손학규	안철수	정몽준	기타	모름 없음
	전체		2%	1%	13%	37%	2%	23%	1%	1%	20%
지역별	서울 인천/경기 강원 대전/충청 광주/전라 대구/경북 부산/울산/경남	1,270 1,733 187 617 626 632 968	2% 1% 2% 2% 6% 1% 2%	2% 2% 1% 1% 1% 1%	13% 13% 14% 11% 17% 8% 12%	33% 33% 43% 42% 12% 59% 45%	2% 2% 1% 2% 5% 0% 1%	25% 24% 16% 25% 34% 13% 20%	1% 1% 1% 0% 1% 1%	1% 1% 2% 1% 2% 0% 1%	21% 22% 21% 17% 23% 16% 17%
성별	제주 남성 여성	67 3,020 3,080	2% 3% 1%	1% 2% 1%	14% 14% 11%	45% 35% 39%	1% 2% 1%	17% 25% 21%	0% 1% 1%	0% 1% 1%	20% 17% 23%
연령별	19~29세 30대 40대 50대 60세 이상	1,130 1,256 1,346 1,146 1,223	1% 2% 3% 3% 2%	1% 1% 2% 2% 1%	18% 21% 13% 7% 4%	22% 22% 34% 49% 59%	1% 1% 2% 3% 2%	36% 32% 26% 14% 7%	0% 1% 1% 1% 0%	1% 1% 2% 1%	20% 19% 18% 20% 23%
직업별	농/임/어업 자영업 블루칼라 화이트칼라 가정주부 학생 무직/기타	194 846 722 1,782 1,484 577 495	2% 4% 2% 2% 1% 0% 2%	1% 2% 1% 1% 1% 2% 1%	6% 11% 11% 18% 9% 19% 8%	49% 42% 41% 27% 45% 19% 50%	1% 2% 2% 2% 1% 1% 2%	15% 20% 22% 30% 17% 41% 10%	1% 1% 2% 1% 0% 0%	1% 1% 1% 1% 1% 0% 1%	23% 18% 17% 18% 24% 19% 25%
이념 성향별	보수 중도 진보 모름/의견없음	1,846 1,769 1,714 771	1% 2% 3% 1%	2% 1% 1% 1%	5% 14% 24% 4%	58% 33% 17% 39%	1% 2% 3% 1%	13% 29% 34% 13%	1% 1% 1% 1%	1% 1% 1% 1%	17% 18% 16% 40%
주요 지지 정당별	새누리당 민 <u>주통</u> 합당 지지정당 없음	1,994 1,343 2,440	1% 3% 2%	2% 1% 1%	2% 31% 9%	79% 10% 21%	1% 4% 2%	6% 36% 29%	1% 1% 0%	1% 1%	9% 14% 34%

#### ◎ 양자 구도 ②

◎ 양작 구포 ②									
박근혜	문재인	모름 없음							
48%	32%	20%							
4/%	33%	20%							
47%	32%	21%							
48%	27%	25%							
52%	30%	18%							
21%	56%	23%							
66%	16%	18%							
56%	25%	19%							
53%	33%	14%							
46%	35%	19%							
50%	29%	21%							
37%	39%	24%							
35%	44%	21%							
45%	38%	17%							
60%	22%	19%							
64%	15%	21%							
56%	20%	24%							
52%	29%	19%							
52%	31%	17%							
39%	42%	19%							
55%	22%	23%							
37%	41%	21%							
59%	19%	23%							
71%	13%	16%							
45%	35%	20%							
28%	56%	17%							
45%	16%	39%							
88%	%	8							
1	%	13 /0							
	%	35%							

# **PRECESSINIG**

대선후보 지지도 데이터 전처리 과정

1

#### **POWER BI**

PDF 파일->데이터 변환

) 대선	후보 지지도 - 대	나자 구도	(자유응	답)								◎ 양지	구도 🧳	
201	2년 7월 1~4주	표본수	김두관	김문수	문재인	박근혜	손학규	안철수	정몽준	기타	모름 없음	박근혜	문재인	모름 없음
	전체	6,100	2%	1%	13%	37%	2%	23%	1%	1%	20%	48%	32%	20%
지역별	서울 인천/경기 강원 대전/충청 광주/전라 대구/경북 부산/울산/경남 제주	1,270 1,733 187 617 626 632 968 67	2% 1% 2% 2% 6% 1% 2% 2%	2% 2% 1% 1% 1% 1% 1%	13% 13% 14% 11% 17% 8% 12% 14%	33% 33% 43% 42% 12% 59% 45%	2% 2% 1% 2% 5% 0% 1%	25% 24% 16% 25% 34% 13% 20% 17%	1% 1% 1% 0% 1% 1% 1%	1% 1% 2% 1% 2% 0% 1%	21% 22% 21% 17% 23% 16% 17% 20%	47% 47% 48% 52% 21% 66% 56% 53%	33% 32% 27% 30% 56% 16% 25% 33%	20% 21% 25% 18% 23% 18% 19%
성별	남성 여성	3,020 3,080	3% 1%	2% 1%	14% 11%	35% 39%	2% 1%	25% 21%	1% 1%	1% 1%	17% 23%	46% 50%	35% 29%	19% 21%
연령별	19~29세 30대 40대 50대 60세 이상	1,130 1,256 1,346 1,146 1,223	1% 2% 3% 3% 2%	1% 1% 2% 2% 1%	18% 21% 13% 7% 4%	22% 22% 34% 49% 59%	1% 1% 2% 3% 2%	36% 32% 26% 14% 7%	0% 1% 1% 1% 0%	1% 1% 2% 1%	20% 19% 18% 20% 23%	37% 35% 45% 60% 64%	39% 44% 38% 22% 15%	24% 21% 17% 19% 21%
직업별	농/임/어업 자영업 블루칼라 화이트칼라 가정주부 학생 무직/기타	194 846 722 1,782 1,484 577 495	2% 4% 2% 2% 1% 0% 2%	1% 2% 1% 1% 1% 2% 1%	6% 11% 11% 18% 9% 19% 8%	49% 42% 41% 27% 45% 19% 50%	1% 2% 2% 2% 1% 1% 1%	15% 20% 22% 30% 17% 41% 10%	1% 1% 2% 1% 0% 0%	1% 1% 1% 1% 1% 0%	23% 18% 17% 18% 24% 19% 25%	56% 52% 52% 39% 55% 37% 59%	20% 29% 31% 42% 22% 41% 19%	24% 19% 17% 19% 23% 21% 23%
이념 성향별	보수 중도 진보 모름/의견없음	1,846 1,769 1,714 771	1% 2% 3% 1%	2% 1% 1% 1%	5% 14% 24% 4%	58% 33% 17% 39%	1% 2% 3% 1%	13% 29% 34% 13%	1% 1% 1% 1%	1% 1% 1% 1%	17% 18% 16% 40%	71% 45% 28% 45%	13% 35% 56% 16%	16% 20% 17% 39%
주요 지지 정당별	새누리당 민주통합당 지지정당 없음	1,994 1,343 2,440	1% 3% 2%	2% 1% 1%	2% 31% 9%	79% 10% 21%	1% 4% 2%	6% 36% 29%	1% 1% 0%	1% 1% 1%	9% 14% 34%	88% 17% 36%	4% 70% 29%	8% 13% 35%

3.	A <sup>R</sup> C Columnia	A <sup>®</sup> C Column2 *	A <sup>E</sup> <sub>C</sub> Column3 -	A <sup>B</sup> C Column4
1	. As	e nul	Pail .	대선 후보다자 구도
2	n n	조사시기	<b>亜</b> 替今	박근데
3	주간	06製 1平(30~03)	1,525	37%
4	주간	08월 2주(06~10)	1,516	36%
5	주간	08製 3주(13~17)	1,222	37%
6	42	08製 4卒(20~24)	1,561	41%
7	주간	08월 5주(27~31)	1,555	40%
8	주간	09월 1주(03~07)	1,521	40%
g	주간	09월 2주(10~14)	1,551	42%
0	주간	09個 3年(17~21)	1,591	39%
1	주간	09部 4平(24-28)	1,577	39%
12	주간	10日 1年(01-05)	1,248	40%
	주간	10個 2平(08~12)	1,551	40%
4	주간	10월 3주(15~18)	1,539	36%
5	주간	10個 4平(22-26)	1,561	37%
6	주간	11部 1주(29~02)	1,544	38%
17	주간	11월 2주(05~09)	1,534	39%
8	주간	11월 3주(12~16)	1,545	39%
0	조가	11944209-20	1.541	1004

모름/없음

28%

33%

23% 12% 9 null

19-29세 30-39세 40-49세 50-59세

전체주이 전체열별 pro\_6 pro\_7 pro\_8 9월 10월 11월 12-1월 12-2

전체

지역별 지역별 지역별

#### **EXCEL**

데이터 전처리 (전체추이, 월별 상세 데이터)

- 강원 55% 34% 11% 1. 전체 후보지지율 대전/세종/51% 37% 11% 1. 전체 후보지지율 광주/전라 18% 66% 16% 대구/경복 67% 24% 5%
  - 2. 특성별 지지율 변화(월 단위)





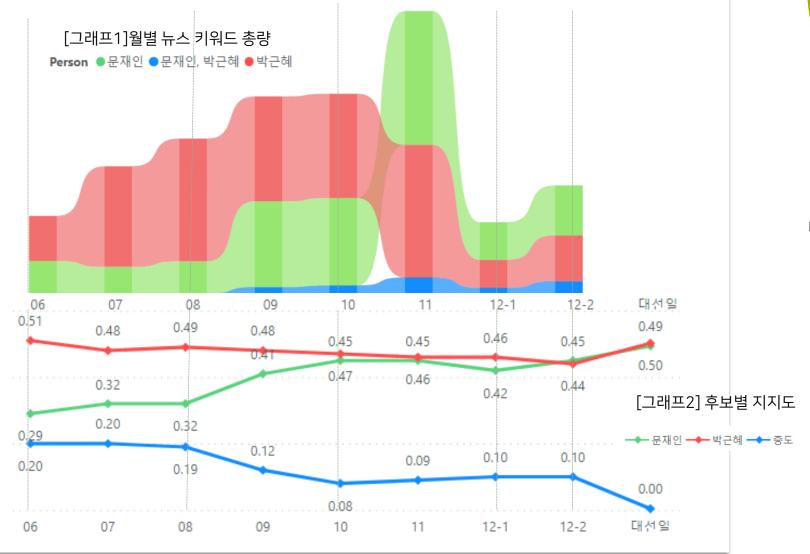




#### **DATA-VISUALIZATION**

#### 각 후보의 언론 노출 분석

뉴스 헤드라인에 노출된 데이터 양은 지지도와 관련이 있을까?



# **ANALYSIS**

빅데이터 증가량에 따라 영향을 받는 지지도 사건 안철수 11월 23일

- 박근혜 데이터 증가량 MAX 월 11월 지지도 MAX 월 12월 1주
- 문재인 데이터 증가량 MAX 8-9월 - 지지도 MAX 월 10월



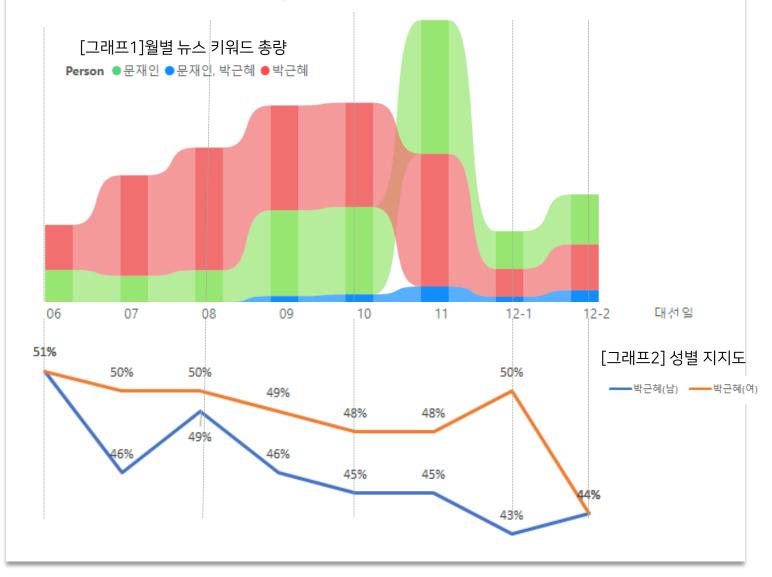


16

INSIGHT 힐러리와 트럼프 대선

#### 각 후보의 언론 노출 분석

뉴스 헤드라인에 노출된 데이터 양은 지지도와 관련이 있을까?



# ANALYSIS

- 박근혜 남자 MAX 지지도 6월 여자 MAX 지지도 7,8월



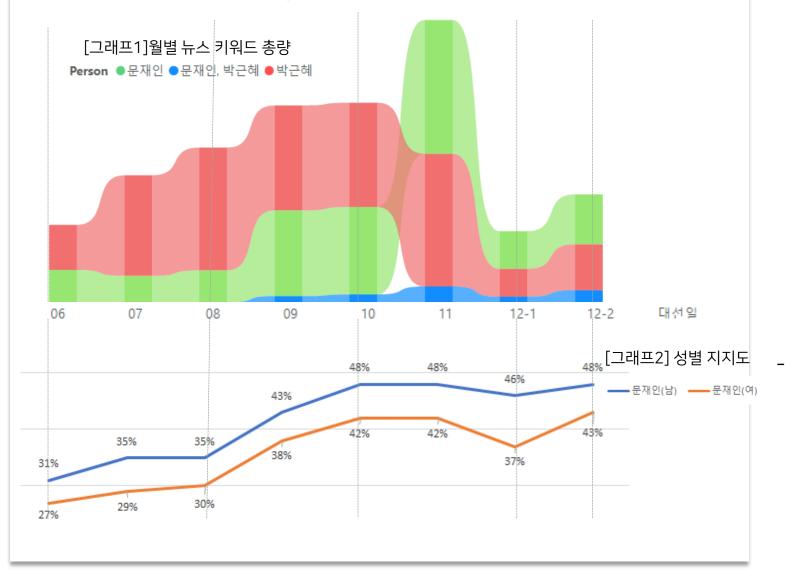




INSIGHT 힐러리와 트럼프 대선

#### 각 후보의 언론 노출 분석

뉴스 헤드라인에 노출된 데이터 양은 지지도와 관련이 있을까?





- 문재인 남자 MAX 지지도 10, 11, 12월(2주차) 여자 MAX 지지도 12월(2주차)







#### 컴퓨터 사양은 중요하다

돌리는데 시간이 너무 오래 걸린다

# INSIGHT



#### 적용하는 건 어렵다

이해가 되었는데 다시하는 게 왜 어려울까? 초반 알고리즘을 잘 짜는 건 중요합니다. 키워드 정제 와 기사 표본 추출을 통해 작업량 감소 시키기

#### 투표합시다

이렇게 복잡한 소중한 권리 행사합시다.

