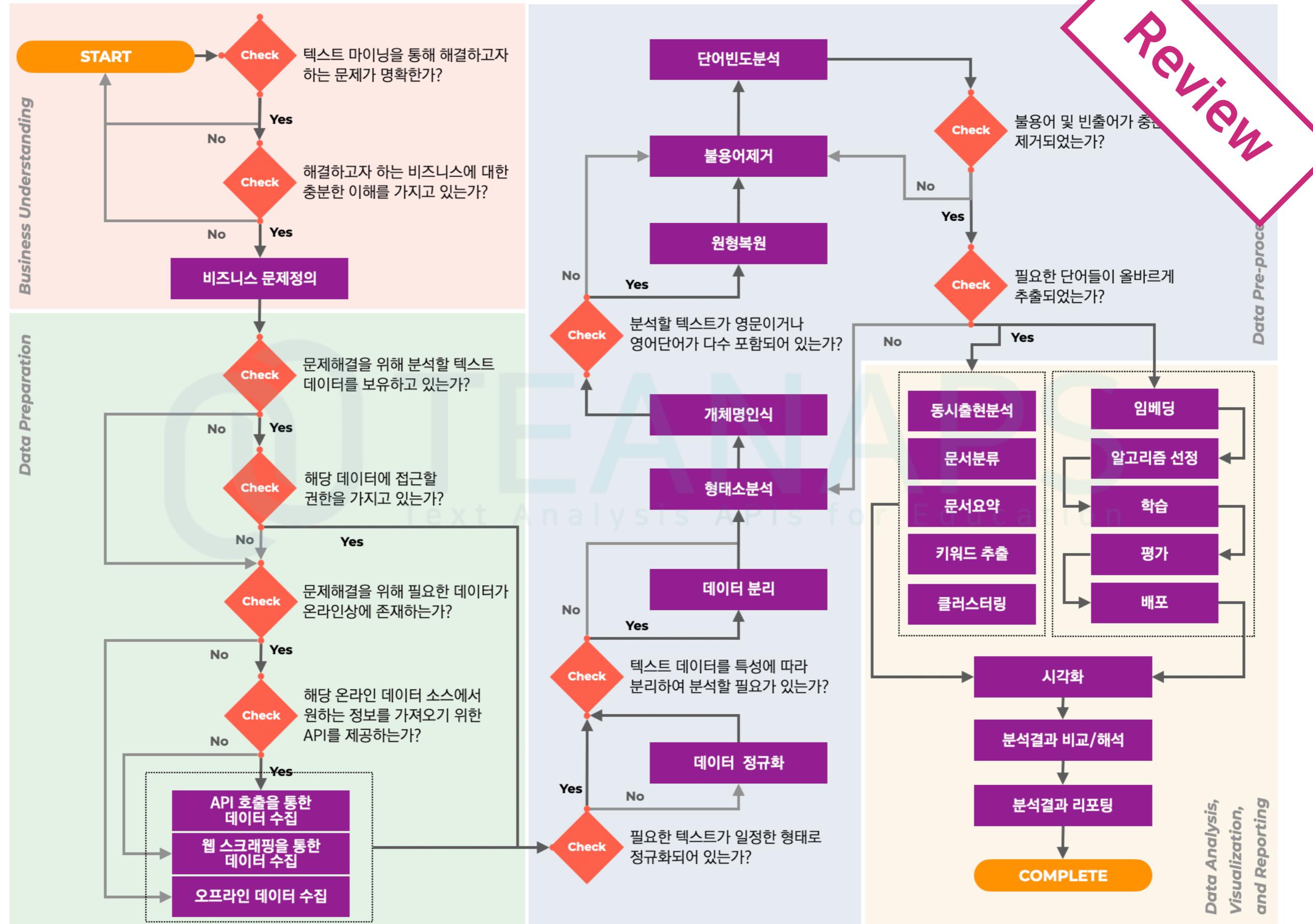


ADVANCED TEXT MINING

by FINGEREDMAN (fingeredman@gmail.com)



WEEK 06

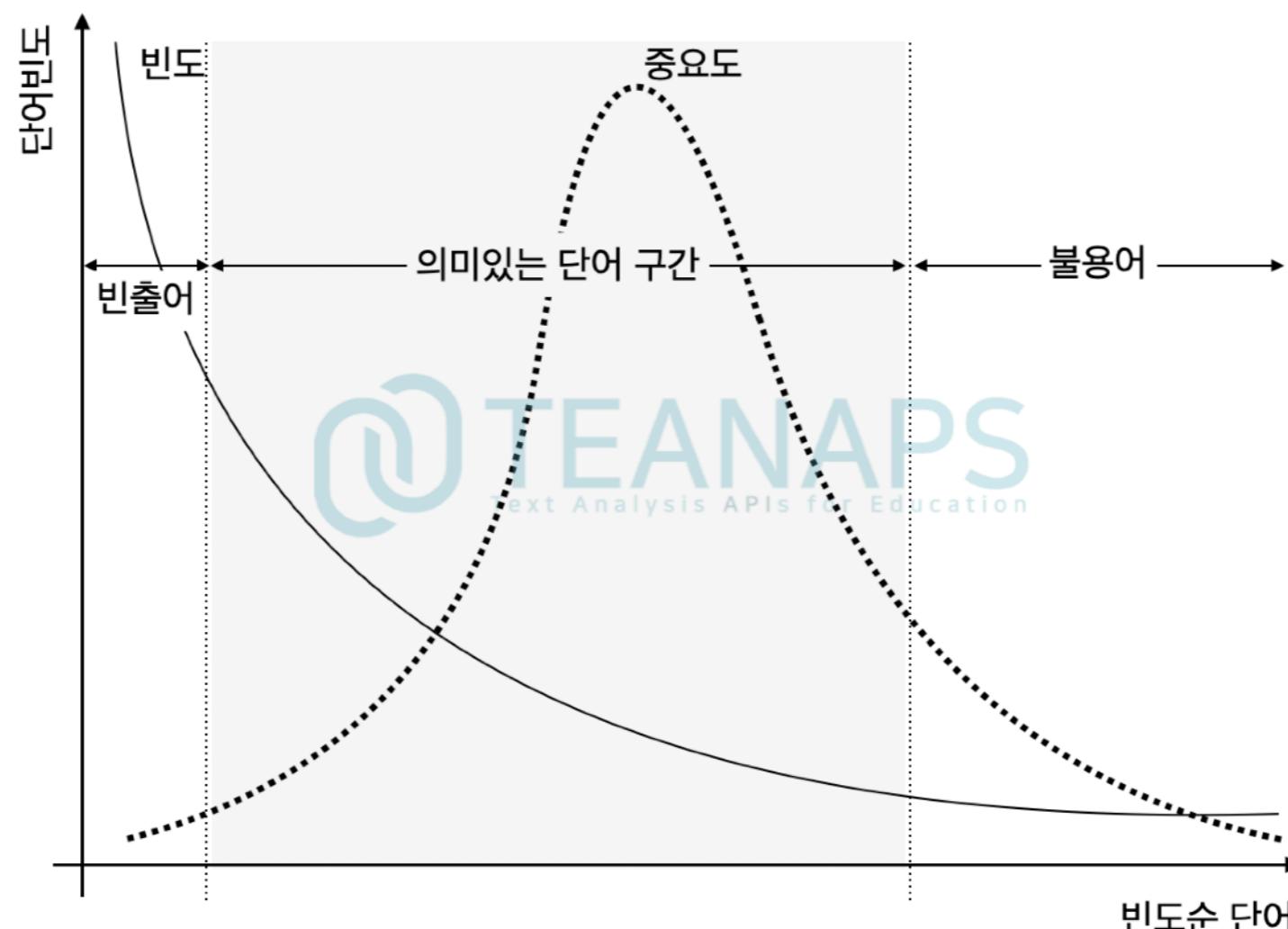
Representing Text Data

단어 가중치: 단어빈도

단어빈도 (Term Frequency, TF)

- 특정 단어가 문서에 출현한 횟수로 단어의 특징을 표현하는 가장 간단한 방법
- 간단하지만 가장 빠르게 문서를 표현하고 파악할 수 있으며 기초통계와 같이 분석 전 반드시 거쳐야 하는 과정
- 단어가 너무 희귀한 경우 큰 의미를 부여하기 어려우며, 너무 흔한 경우 의미가 과도하게 부여될 가능성이 있음

$$\text{TermFrequency} = \text{count}(word | document)$$

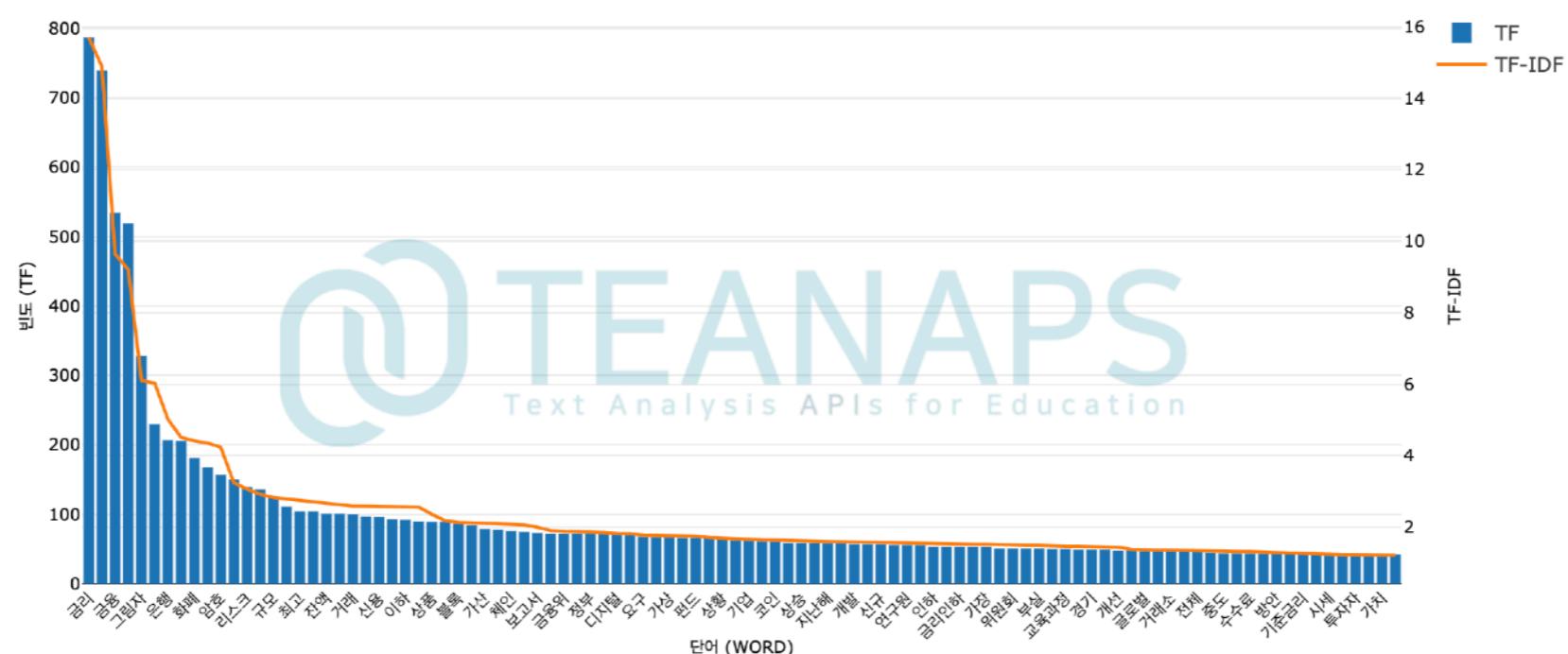


단어 가중치: 단어빈도

| TF-IDF (TF-Inverse Document Frequency)

- 역문서빈도 (Inverse Document Frequency, IDF) : 단어가 출현한 문서가 적을수록 단어의 가중치를 낮게 표현하는 방법 (희박성)
 $IDF = 1 + \log(\text{document}/\text{count(word/document)})/\log(\text{document})$
- TF와 IDF 개념을 통합하여, 단어가 문서에 출현한 횟수와 희박성을 동시에 활용해 가중치를 표현하는 방법
 $TF - IDF = \text{Frequency} * IDF$
- 지프의 법칙 (Zipf's law)
 - 1) 자연어에 나타나는 단어들을 출현 횟수가 높은 순으로 정렬하면, 단어의 출현 횟수는 순위에 반비례함
 - 2) 가장 사용 빈도가 높은 단어는 두 번째 단어보다 빈도가 약 두 배 높으며, 세 번째 단어보다는 빈도가 세 배 높음

단어빈도 및 TF-IDF (TF & TF-IDF)

* 김수인, 김재원, and 배휘동, 왜 프로그래밍에는 창의성이 필요하다고 할까요, 2017.5.25., <https://medium.com/elice/>.

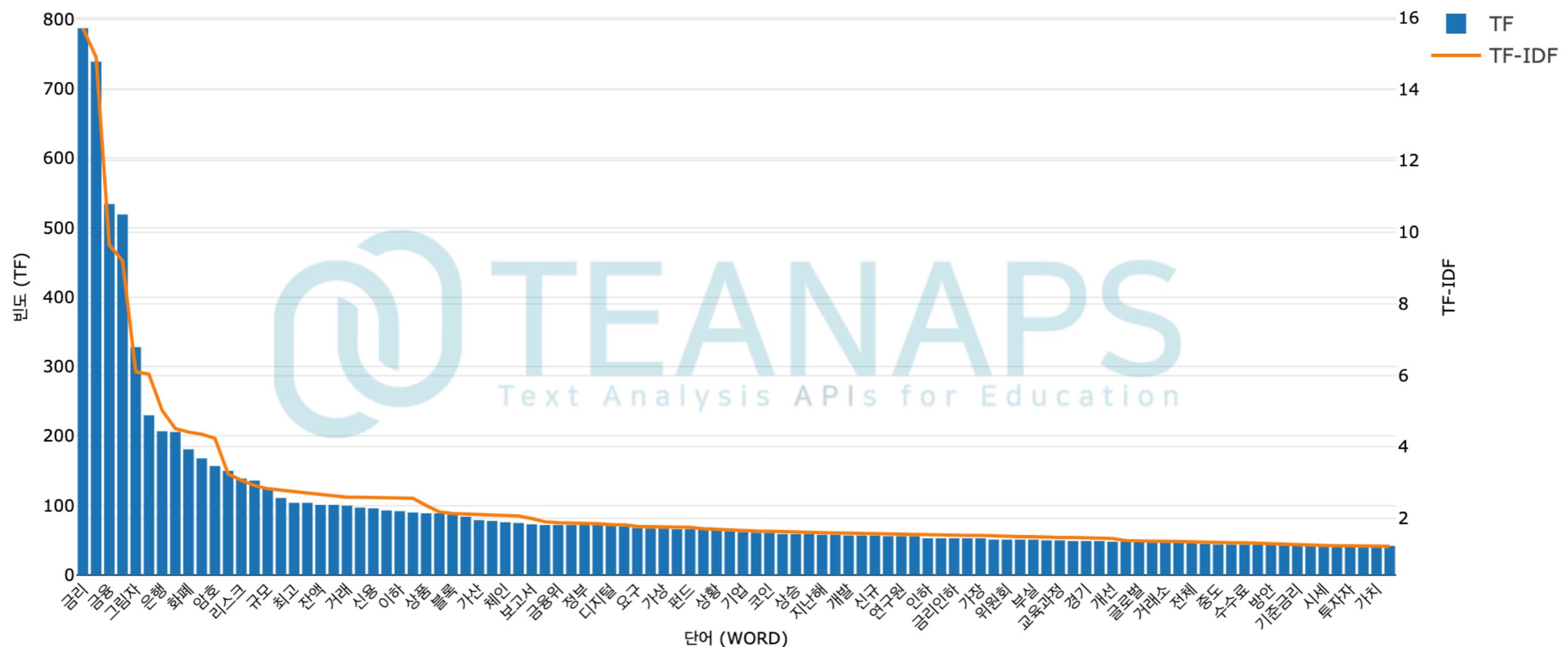
** references

*** references

단어 가중치: 단어빈도

TF & TF-IDF

단어빈도 및 TF-IDF (TF & TF-IDF)



문서를 단어 가중치로 표현하는 방법

문서 내 단어의 빈도 계산하기

OhmyNews

"벚꽃 상춘객 올까봐 불도 꺼... 올해는 제발 참아달라"

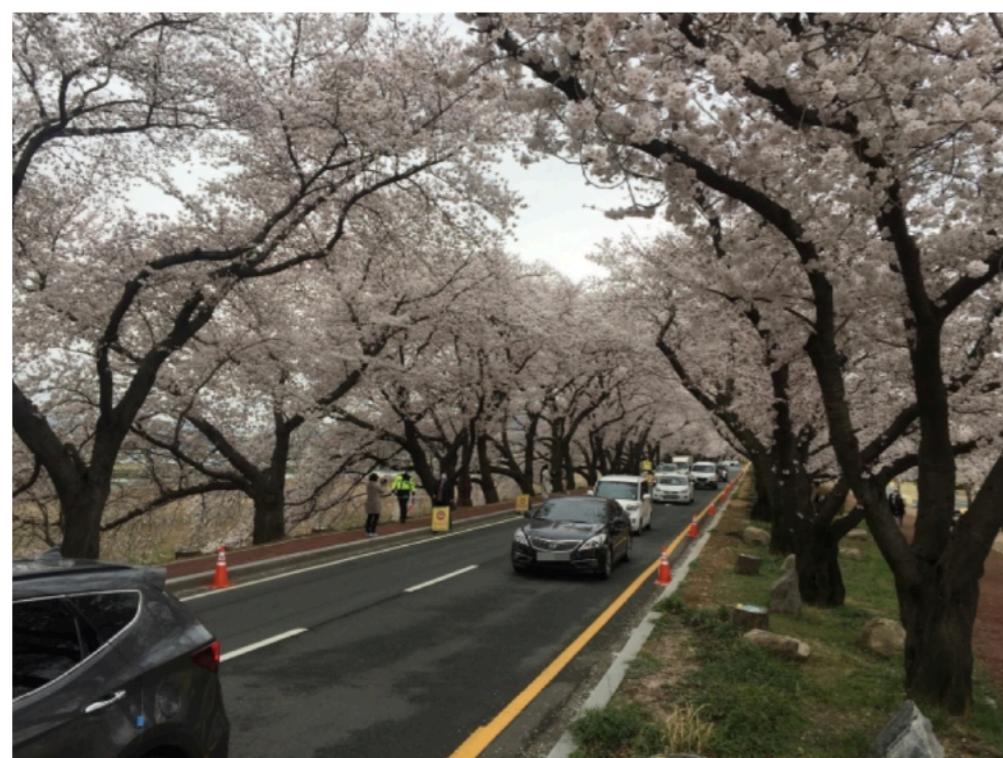
기사입력 2020.03.28. 오후 8:30 기사원문 스크랩 본문듣기 설정

좋아요 27 댓글 14

요약봇 가 둘러보기

[사진] 경주 벚꽃터널, 주정차 금지... 상춘객 몰릴 것에 대비해 야간경관조명도 꺼

[오마이뉴스 한정환 기자]



▲ 28일 주말 오전, 조금은 한산한 경주 흥무로 벚꽃길 모습

© 한정환

천년고도 경주에 벚꽃이 활짝 피었다. 해마다 이맘때쯤이면 경주는 벚꽃 관광객 맞이로 분주했다. 그러나 코로나19 확산으로, 올해는 사정이 달라졌다.

28일 오전 주말을 맞아 벚꽃길로 유명한 흥무로 벚꽃터널을 찾아보았다. 예년에 비해 벚꽃 상춘객들은 많이 줄었다. 코로나19 여파로 많은 관광객이 경주 방문을 자제하고 있는 듯 보인다. 그러나 벚꽃 상춘객이 몰릴 것에 대비하여 벚꽃터널에는 도로 양방향 100m 간격으로 경찰관들이 배치되어 불법 주정차 단속을 하고 있다. 사진을 찍기 위해 차에서 잠시 내리는 것도 안 된다.

관광도시 경주의 특성상 다른 도시처럼 벚꽃 명소를 통제할 수가 없다. 도시 전체가 관광지이고, 대부분 가로수에 벚꽃이 심어져 있어 통제를 하게 되면, 도시 전체가 봉쇄가 되어야 하기 때문에 현수막을 걸어 두고 홍보만 하고 있다. 사진으로나 마 경주 벚꽃 소식을 전하기로 했던 '경주시 벚꽃 알리미'도 잠정 중단된 상태이다.

경주시 관계자는 "지난 22일 일제히 불을 밝힌 야간경관조명도 27일 저녁부터 벚꽃 상춘객이 몰릴 것에 대비하여 일제히 불을 껐다. 벚꽃은 해마다 피니, 올해는 집단 감염이 우려되는 벚꽃 나들이를 내년으로 미루고, 코로나19 확산 방지를 위해 모두가 힘을 합쳐달라"라고 말했다.

경주시는 지난 2월 22일 첫 코로나19 확진자가 발생한 이래 현재까지 총 40명이 양성 판정을 받았으며, 이 중에서 사망 1명, 완치 10명을 제외한 29명이 현재 자가격리 및 생활치료센터에 입소하여 치료를 받고 있다.

문서를 단어 가중치로 표현하는 방법

| 형태소 분석을 통한 단어주머니 생성

구분	문장
문장 01	천년고도 경주에 벚꽃이 활짝 피었다.
문장 02	해마다 이맘때쯤이면 경주는 벚꽃 관광객 맞이로 분주했다.
문장 03	그러나 코로나19 확산으로, 올해는 사정이 달라졌다.
문장 04	8일 오전 주말을 맞아 벚꽃길로 유명한 흥무로 벚꽃터널을 찾아보았다.
문장 05	예년에 비해 벚꽃 상춘객들은 많이 줄었다.
문장 06	코로나19 여파로 많은 관광객이 경주 방문을 자제하고 있는 듯 보인다.
문장 07	그러나 벚꽃 상춘객이 몰릴 것에 대비하여 벚꽃터널에는 도로 양방향 100m 간격으로 경찰관들이 배치되어 불법 주정차 단속을 하고 있다.
문장 08	사진을 찍기 위해 차에서 잠시 내리는 것도 안 된다.
문장 09	관광도시 경주의 특성상 다른 도시처럼 벚꽃 명소를 통제할 수가 없다.
문장 10	도시 전체가 관광지이고, 대부분 가로수에 벚꽃이 심어져 있어 통제를 하게 되면, 도시 전체가 봉쇄가 되어야 하기 때문에 현수막을 걸어 두고 홍보만 하고 있다.
문장 11	사진으로나마 경주 벚꽃 소식을 전하기로 했던 '경주시 벚꽃 알리미'도 잠정 중단된 상태이다.
문장 12	경주시 관계자는 "지난 22일 일제히 불을 밝힌 야간경관조명도 27일 저녁부터 벚꽃 상춘객이 몰릴 것에 대비하여 일제히 불을 껐다.
문장 13	벚꽃은 해마다 피니, 올해는 집단 감염이 우려되는 벚꽃 나들이를 내년으로 미루고, 코로나19 확산 방지를 위해 모두가 힘을 합쳐달라"라고 말했다.
문장 14	경주시는 지난 2월 22일 첫 코로나19 확진자가 발생한 이래 현재까지 총 40명이 양성 판정을 받았으며, 이 중에서 사망 1명, 완치 10명을 제외한 29명이 현재 자가격리 및 생활치료센터에 입소하여 치료를 받고 있다.

문서를 단어 가중치로 표현하는 방법

| 형태소 분석을 통한 단어주머니 생성



문서를 단어 가중치로 표현하는 방법

형태소 분석을 통한 단어주머니 생성

구분	유니그램 (품사=NNG NNP, 불용어 제거)	단어주머니 (74 단어)
문장 01	고도, 경주, 벚꽃	벚꽃 경주 도로 대부분 제외 저녁
문장 02	이맘때, 경주, 벚꽃, 관광객, 분주	도시 단속 잠정 자제
문장 03	코로나, 확산, 올해, 사정	코로나 내년 자제
문장 04	오전, 주말, 벚꽃, 길, 유명, 흥무, 벚꽃, 터널	상춘객 관광지 자가 입소 관광 관계자 고도 유명
문장 05	예년, 벚꽃, 상춘객	전체 사진 치료 터널 경찰관 경관 격리 감염 간격 나들이 센터 상태 생활 홍보 현재 현수막 판정 특성 집단 중단 주정차 양성 여파 우려 완치 오전 예년 우려 완치
문장 06	코로나, 여파, 관광객, 경주, 방문, 자제	이맘때 여파 우려 완치 오전 예년 우려 완치
문장 07	벚꽃, 상춘객, 대비, 벚꽃, 터널, 도로, 양방향, 간격, 경찰관, 배치, 불법, 주정차, 단속	대비 확산 나들이 센터 상태 생활 홍보 현재 현수막 판정 특성 집단 중단 주정차 양성 여파 우려 완치 오전 예년 우려 완치
문장 08	사진	관광객 발생 모두 방문 봉쇄 방지 배치 분주 불법 사망 사정 명소 흥무 경찰관 경관 격리 감염 간격 나들이 센터 상태 생활 홍보 현재 현수막 판정 특성 집단 중단 주정차 양성 여파 우려 완치 오전 예년 우려 완치
문장 09	관광, 도시, 경주, 특성, 도시, 벚꽃, 명소, 통제	경찰관 경관 격리 감염 간격 나들이 센터 상태 생활 홍보 현재 현수막 판정 특성 집단 중단 주정차 양성 여파 우려 완치 오전 예년 우려 완치
문장 10	도시, 전체, 관광지, 대부분, 가로수, 벚꽃, 통제, 전체, 봉쇄, 현수막, 흥보	전체 봉쇄 방지 배치 분주 불법 사망 사정 명소 흥무 경찰관 경관 격리 감염 간격 나들이 센터 상태 생활 홍보 현재 현수막 판정 특성 집단 중단 주정차 양성 여파 우려 완치 오전 예년 우려 완치
문장 11	사진, 경주, 벚꽃, 소식, 경주시, 벚꽃, 알리, 잠정, 중단, 상태	경찰관 경관 격리 감염 간격 나들이 센터 상태 생활 홍보 현재 현수막 판정 특성 집단 중단 주정차 양성 여파 우려 완치 오전 예년 우려 완치
문장 12	경주시, 관계자, 야간, 경관, 조명, 저녁, 벚꽃, 상춘객, 대비	경주시 관계자 야간 경관 조명 저녁 벚꽃 상춘객 대비 경찰관 경관 격리 감염 간격 나들이 센터 상태 생활 홍보 현재 현수막 판정 특성 집단 중단 주정차 양성 여파 우려 완치 오전 예년 우려 완치
문장 13	벚꽃, 올해, 집단, 감염, 우려, 벚꽃, 나들이, 내년, 코로나, 확산, 방시, 위해, 모두	벚꽃 올해 집단 감염 우려 벚꽃 나들이 내년 코로나 확산 방시 위해 모두 경찰관 경관 격리 감염 간격 나들이 센터 상태 생활 홍보 현재 현수막 판정 특성 집단 중단 주정차 양성 여파 우려 완치 오전 예년 우려 완치
문장 14	경주시, 지난, 코로나, 확진, 발생, 현재, 양성, 판정, 사망, 완치, 제외, 격리, 생활, 치료, 센터, 입소, 치료	경주시 지난 코로나 확진 발생 현재 양성 판정 사망 완치 제외 격리 생활 치료 센터 입소 치료 경찰관 경관 격리 감염 간격 나들이 센터 상태 생활 홍보 현재 현수막 판정 특성 집단 중단 주정차 양성 여파 우려 완치 오전 예년 우려 완치

문서를 단어 가중치로 표현하는 방법

형태소 분석을 통한 단어주머니 생성

구분	벚꽃	경주	도시	코로나	상춘객	경주시	전체	사진	치료	터널	통제	올해	대비	확산	관광객	발생	모두	방문	봉쇄	방지	배치	분주	불법	사망	사정	명소	흥무
문장 01	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
문장 02	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	
문장 03	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	
문장 04	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
문장 05	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
문장 06	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	
문장 07	2	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	
문장 08	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
문장 09	1	1	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
문장 10	1	0	2	0	0	0	2	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
문장 11	2	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
문장 12	1	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
문장 13	2	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	
문장 14	0	0	0	1	0	1	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	

단어 가중치: 네트워크 중심성

동시출현 분석 (Co-word Analysis)

- 문서에 서로 다른 두 단어의 동시출현 횟수와 네트워크 중심성을 통해 단어의 특징을 표현하는 방법
- 두 단어 사이의 동시출현을 연관성의 척도로 취급하고, 그 관계를 네트워크 중심성으로 표현하여 가중치를 계산함
- **연관어** (공기어, Co-word) : 하나의 문서에서 함께 출현하여 서로 밀접한 의미관계를 가지는 단어

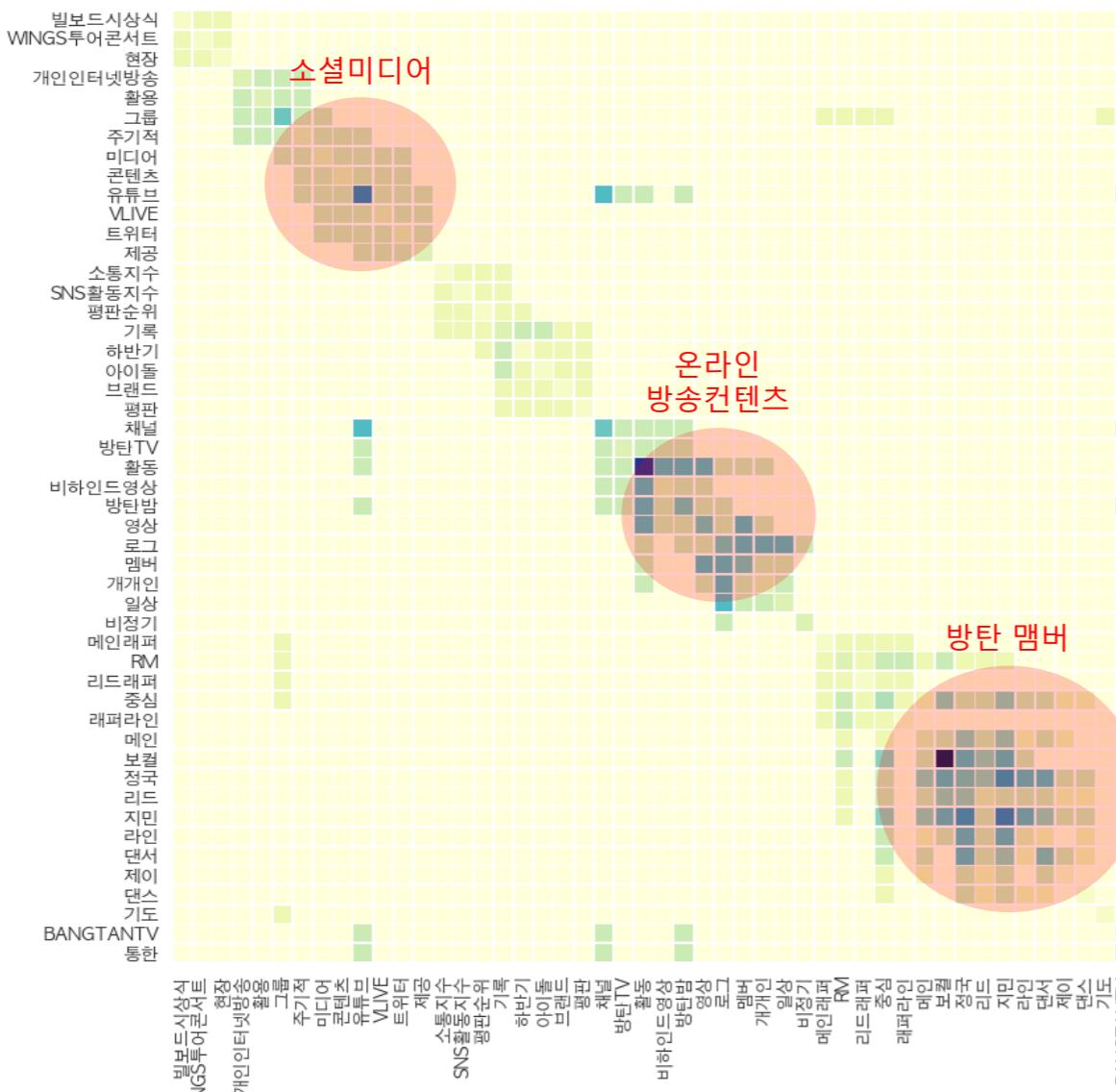


표1 '아쿠르트 아줌마' 연관어 변화

아쿠르트 아줌마는 여전히 '아쿠르트'와의 연관도가 가장 높지만 2016년 들어 '커피' 및 '크림치즈' 제품 연관어와 '10일'이라는 키워드가 등장. 아쿠르트 아줌마는 '배달하는' 역할에서 맛난 제품을 위해 '만나고' '찾고' '발견하는' 대상으로 변화 중.

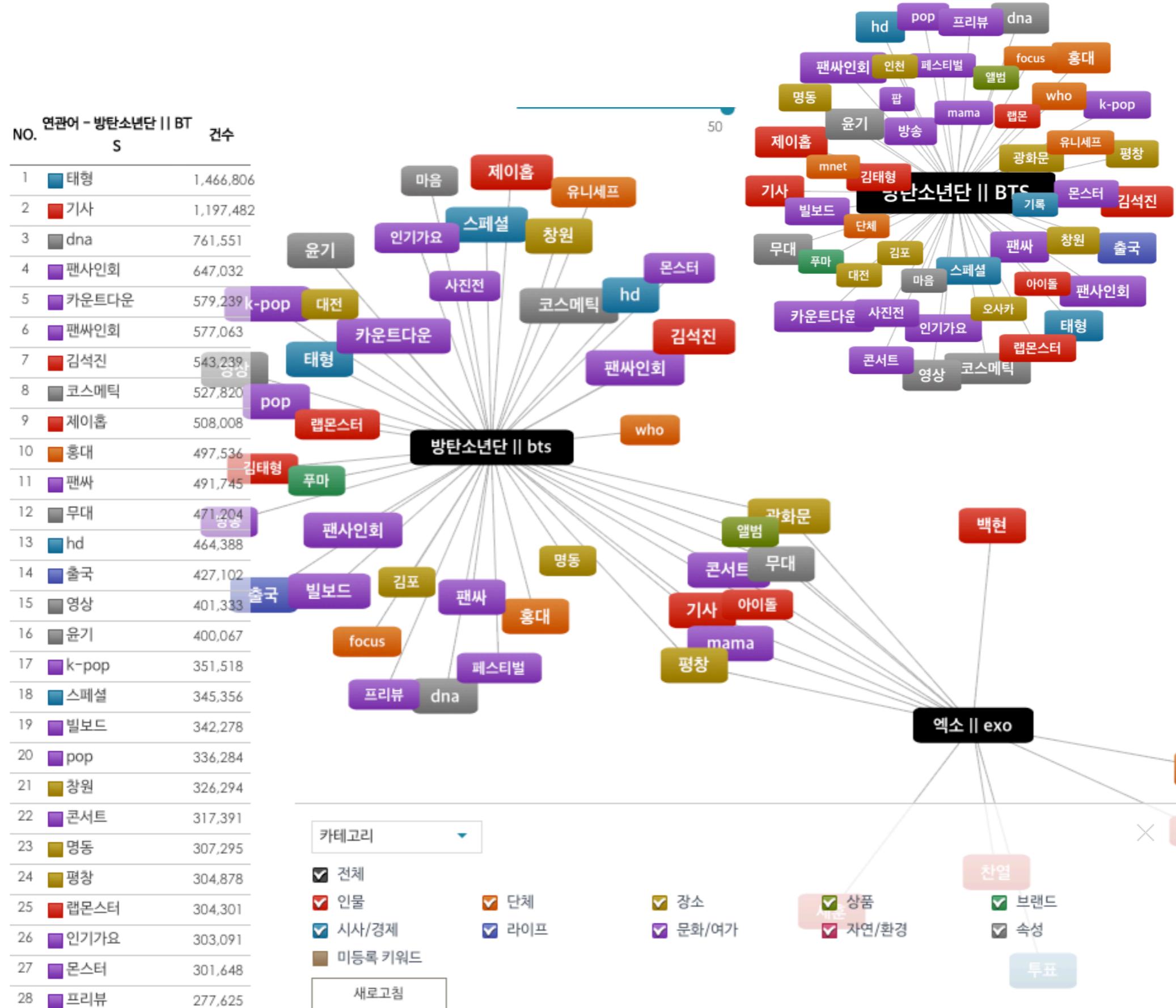
2013년		2014년		2015년		2016년		
No.	연관어	언급 비중	No.	연관어	언급 비중	No.	연관어	언급 비중
1	아쿠르트	21.3%	1	아쿠르트	26.3%	1	아쿠르트	26.6%
2	먹다	4.9%	2	건강	4.5%	2	집	4.7%
3	아침	4.4%	3	아침	4.0%	3	아침	4.4%
4	엄마	4.2%	4	집	3.6%	4	맛	3.9%
5	집	3.5%	5	제품	3.4%	5	먹다	3.4%
6	오다	2.8%	6	엄마	3.3%	6	사다	2.8%
7	사다	2.7%	7	맛	2.7%	7	주다	2.8%
8	주다	2.5%	8	같다	2.6%	8	다니다	2.7%
9	구입하다	2.4%	9	우유	2.6%	9	엄마	2.6%
10	아이	2.4%	10	주다	2.2%	10	우유	2.1%
11	아쿠르트 주다	2.3%	11	먹다	2.2%	11	만나다	2.1%
12	배달하다	2.3%	12	만나다	2.0%	12	제품	2.0%
13	수입	2.3%	13	사다	1.9%	13	사진	2.0%
14	다니다	2.1%	14	알다	1.9%	14	나오다	2.0%
15	얼려먹다	2.0%	15	배달하다	1.8%	15	팔다	1.9%
16	살다	2.0%	16	다니다	1.8%	16	지나가다	1.8%
17	제품	2.0%	17	하루야채	1.7%	17	하나	1.7%
18	세븐	1.8%	18	나누다	1.7%	18	판매	1.7%
19	가다	1.8%	19	지나가다	1.6%	19	일하다	1.6%
20	자녀	1.8%	20	세븐	1.5%	20	오다	1.6%
21	만나다	1.8%	21	수입	1.5%	21	찾다	1.6%
22	마시다	1.7%	22	찾다	2.3%	22	음료	1.5%
23	유산균	1.7%	23	노인	1.4%	23	마시다	1.4%
24	일하다	1.7%	24	마시다	1.4%	24	길	1.4%
...				
29	팔다	1.4%	29	물다	1.3%	29	배달하다	1.3%
29	구입하다	1.0%	29	구입하다	1.0%	29	구입하다	1.0%

상승 키워드 하락 키워드 신규 키워드

* 전병진, 신한은행 파이낸스: 텍스트 마이닝 기초, 2018.12.12.

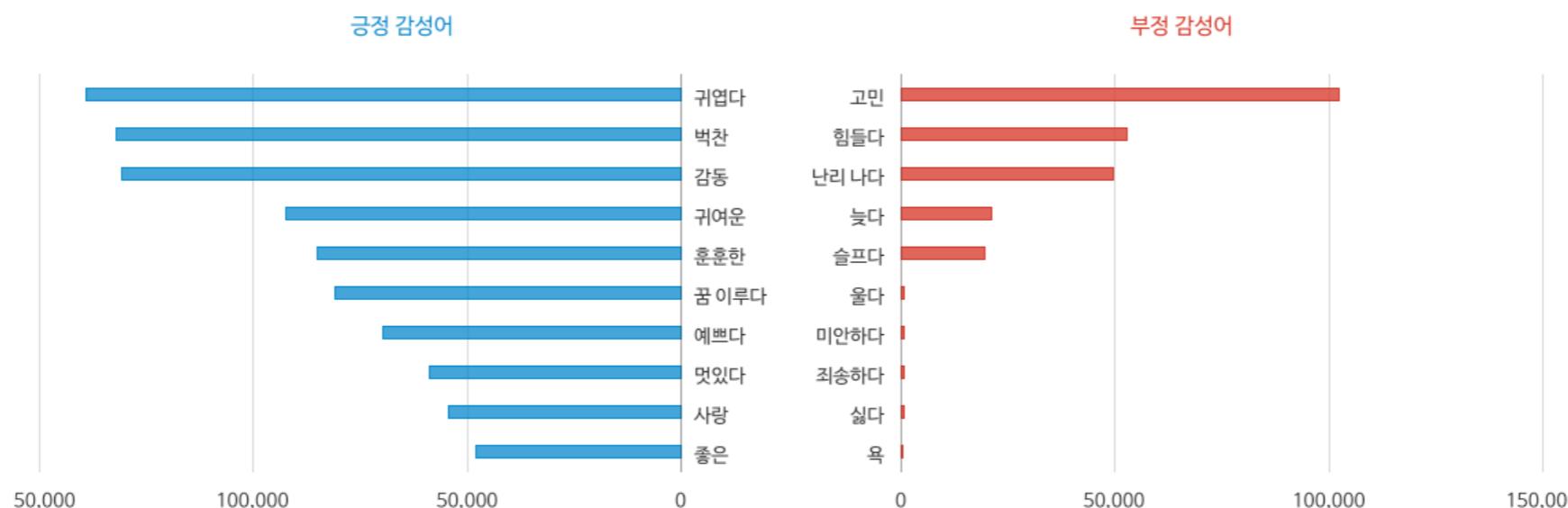
** 백경혜(DBR), "매력을 소비하는 나는 덕후! 즐거움을 위해 기꺼이 지갑을 연다", 2017.1., http://dbr.donga.com/article/view/1203/article_no/7935/.

*** references



NO.	연관어 - 엑소 exo	건수
1	투표	539,724
2	콘서트	469,536
3	백현	373,347
4	찬열	363,191
5	세훈	342,956
6	엑소엘	333,724
7	mama	302,523
8	시우민	185,113
9	그룹	153,995
10	대상	153,270
11	티켓	147,756
12	멜론	144,508
13	파워	136,761
14	video	134,713
15	평창	134,577
16	광화문	133,053
17	단독콘서트	120,636
18	앨범	118,598
19	레이	111,334
20	멤버	107,329
21	가수	106,322
22	디오	104,295
23	기사	98,662
24	아이돌	98,416
25	고척스카이돔	95,535
26	music	88,950
27	올림픽	88,537
28	종대	88,349

감성 키워드 순위



기간별 연관어 순위 : 방탄소년단 || BTS

2017/10/03 ~ 2017/11/03

 전체 트위터 블로그 커뮤니티 인스타그램 뉴스 확인

 일별 주별 월별 분기별


카테고리	순위	2017/10/03~2017/10/07		2017/10/08~2017/10/14		2017/10/15~2017/10/21		2017/10/22~2017/10/28		2017/10/29~2017/11/03	
		연관어	건수								
<input checked="" type="checkbox"/> 전체	1	방탄소년단	826,236	방탄소년단	987,650	방탄소년단	798,031	방탄소년단	549,312	방탄소년단	1,481,373
<input checked="" type="checkbox"/> 인물	2	태형	449,263	태형	552,652	기사	260,201	김석진	147,539	평창	289,317
<input checked="" type="checkbox"/> 단체	3	코스메틱	404,925	홍대	428,101	태형	217,125	태형	141,507	광화문	250,964
<input checked="" type="checkbox"/> 장소	4	기사	260,803	기사	423,304	출국	205,573	hd	108,600	콘서트	214,887
<input checked="" type="checkbox"/> 상품	5	명동	253,605	dna	362,319	푸마	187,739	무대	106,730	기사	205,638
<input checked="" type="checkbox"/> 브랜드	6	팬싸인회	207,776	카운트다운	355,310	김석진	165,320	타이페이	95,120	유니세프	205,279
<input checked="" type="checkbox"/> 라이프	7	팬싸인회	200,843	팬싸인회	329,385	dvd	128,894	잼	92,580	캠페인	131,943
<input checked="" type="checkbox"/> 시사/경제	8	마음	198,856	영상	297,121	hd	127,530	제이홉	77,808	무대	119,974
<input checked="" type="checkbox"/> 문화/여가	9	who	186,415	팬싸인회	283,711	dna	124,366	콘서트	77,700	스페셜	119,423
<input checked="" type="checkbox"/> 자연/환경	10	dna	184,025	팬싸	222,826	제이홉	117,073	윤기	76,071	리허설	113,370
<input checked="" type="checkbox"/> 속성	11	팬싸	175,731	출국	197,729	mama	107,045	대만	66,871	올림픽	109,404

표1 '야쿠르트 아줌마' 연관어 변화

야쿠르트 아줌마는 여전히 '야쿠르트'와의 연관도가 가장 높지만 2016년 들어 '커피' 및 '크림치즈' 제품 연관어와 '10일'이라는 키워드가 등장. 야쿠르트 아줌마는 '배달하는' 역할에서 맛난 제품을 위해 '만나고' '찾고' '발견하는' 대상으로 변화 중.

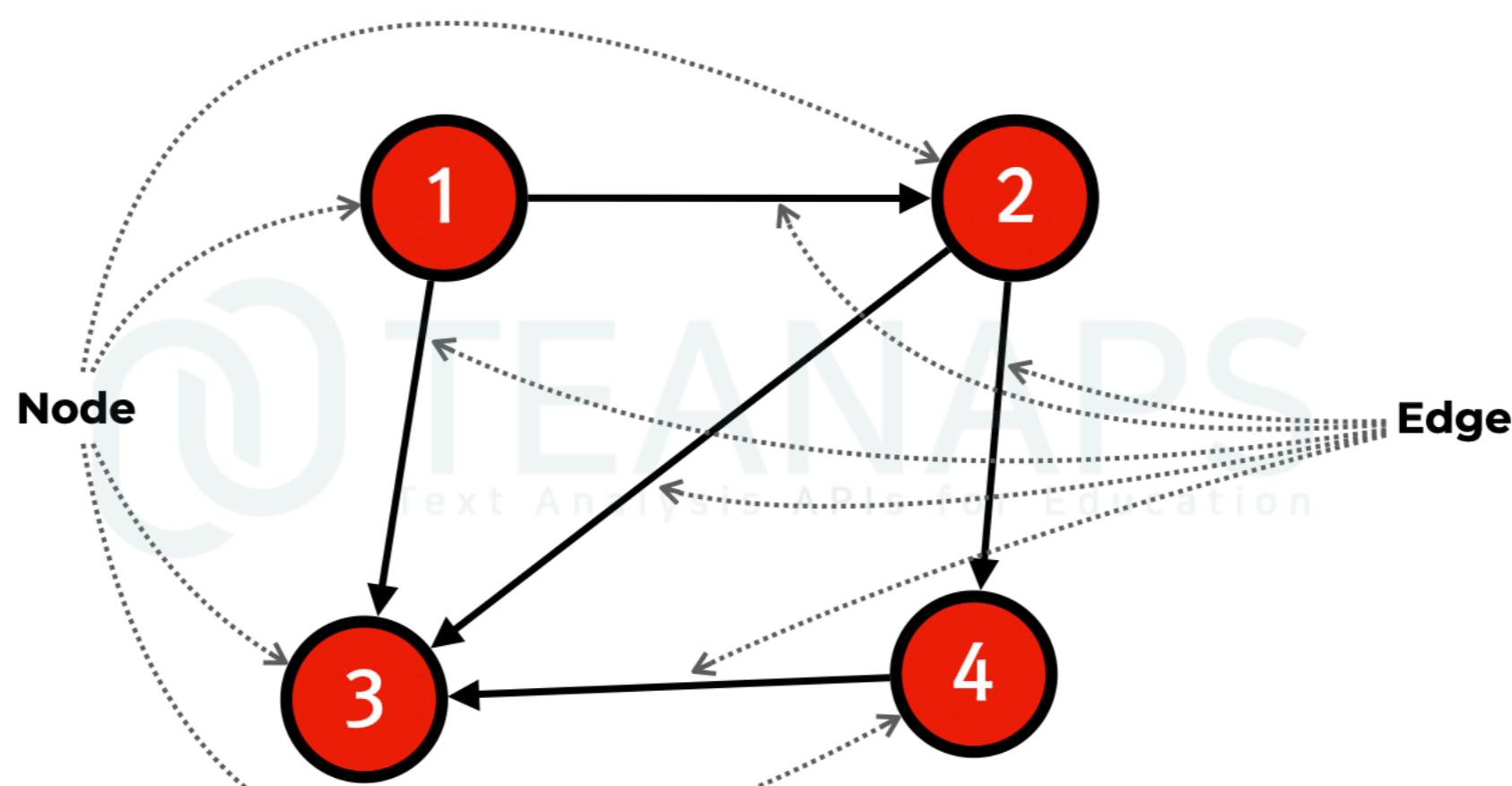
2013년			2014년			2015년			2016년		
No.	연관어	언급 비중	No.	연관어	언급 비중	No.	연관어	언급 비중	No.	연관어	언급 비중
1	야쿠르트	21.3%	1	야쿠르트	26.3%	1	야쿠르트	26.6%	1	야쿠르트	13.1%
2	먹다	4.9%	2	건강	4.5%	2	집	4.7%	2	콜드브루	8.2%
3	아침	4.4%	3	아침	4.0%	3	아침	4.4%	3	커피	7.4%
4	엄마	4.2%	4	집	3.6%	4	맛	3.9%	4	맛	6.6%
5	집	3.5%	5	제품	3.4%	5	먹다	3.4%	5	끼리	5.7%
6	오다	2.8%	6	엄마	3.3%	6	사다	2.8%	6	치즈	5.3%
7	사다	2.7%	7	맛	2.7%	7	주다	2.8%	7	과자	5.0%
8	주다	2.5%	8	같다	2.6%	8	다니다	2.7%	8	아메리카노	4.1%
9	구입하다	2.4%	9	우유	2.6%	9	엄마	2.6%	9	먹다	3.3%
10	아이	2.4%	10	주다	2.2%	10	우유	2.1%	10	크림치즈	3.1%
11	야쿠르트 주다	2.3%	11	먹다	2.2%	11	만나다	2.1%	11	라떼	2.8%
12	배달하다	2.3%	12	만나다	2.0%	12	제품	2.0%	12	만나다	2.7%
13	수입	2.3%	13	사다	1.9%	13	사진	2.0%	13	가격	2.4%
14	다니다	2.1%	14	알다	1.9%	14	나오다	2.0%	14	찾다	1.9%
15	얼려먹다	2.0%	15	배달하다	1.8%	15	팔다	1.9%	15	아침	1.8%
16	살다	2.0%	16	다니다	1.8%	16	지나가다	1.8%	16	10일	1.6%
17	제품	2.0%	17	하루야채	1.7%	17	하나	1.7%	17	엄마	1.5%
18	세븐	1.8%	18	나누다	1.7%	18	판매	1.7%	18	우유	1.4%
19	가다	1.8%	19	지나가다	1.6%	19	일하다	1.6%	19	팔다	1.3%
20	자녀	1.8%	20	세븐	1.5%	20	오다	1.6%	20	발견하다	1.3%
21	만나다	1.8%	21	수입	1.5%	21	찾다	1.6%	21	사다	1.2%
22	마시다	1.7%	22	찾다	2.3%	22	음료	1.5%	22	인기	1.2%
23	유산균	1.7%	23	노인	1.4%	23	마시다	1.4%	23	편의점	1.2%
24	일하다	1.7%	24	마시다	1.4%	24	길	1.4%	24	끼리딥앤크런치	1.1%
...				
29	팔다	1.4%	29	묻다	1.3%	29	배달하다	1.3%	29	구입하다	1.0%

상승 키워드 하락 키워드 신규 키워드

단어 가중치: 네트워크 중심성

그래프 (Graph) 기본개념

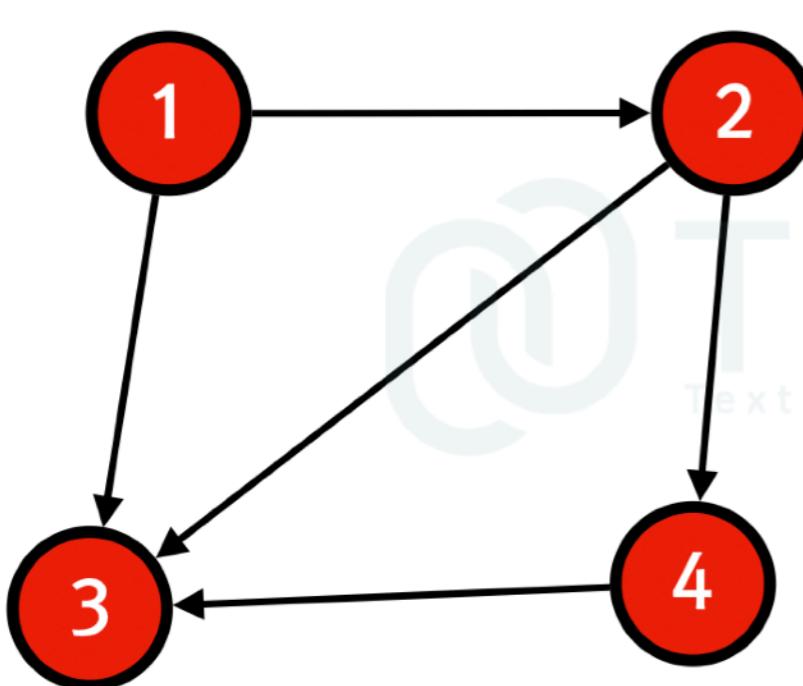
- 노드 (node, vertex, point) : 관계를 가지는 그래프 요소
- 엣지 (edge, line, arc) : 관계로 연결된 한 쌍의 노드
- 방향성 그래프 (directed graph) : 화살표를 이용해 방향이 표시된 그래프
- 비방향성 그래프 (undirected graph) : 방향성이 없는 그래프



단어 가중치: 네트워크 중심성

그래프 (Graph) 기본개념

방향성 그래프 (*directed graph*)



엣지리스트 (edge list)

Vertex Vertex

1	2
1	3
2	3
2	4
3	4

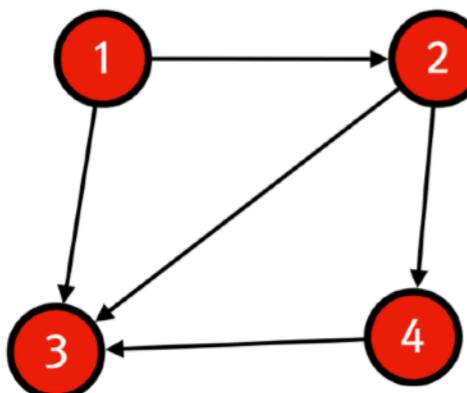
인접행렬 (adjacency matrix)

Vertex	1	2	3	4
1	-	1	1	0
2	0	-	1	1
3	0	0	-	0
4	0	0	1	-

단어 가중치: 네트워크 중심성

그래프 (Graph) 기본개념

방향성 그래프 (*directed graph*)

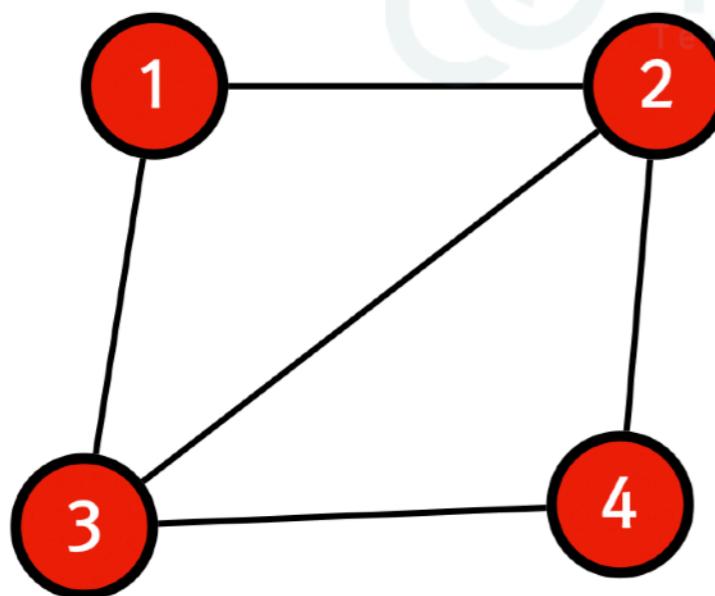


엣지리스트 (*edge list*)

Vertex Vertex

1	2
1	3
2	3
2	4
3	4

비방향성 그래프 (*undirected graph*)



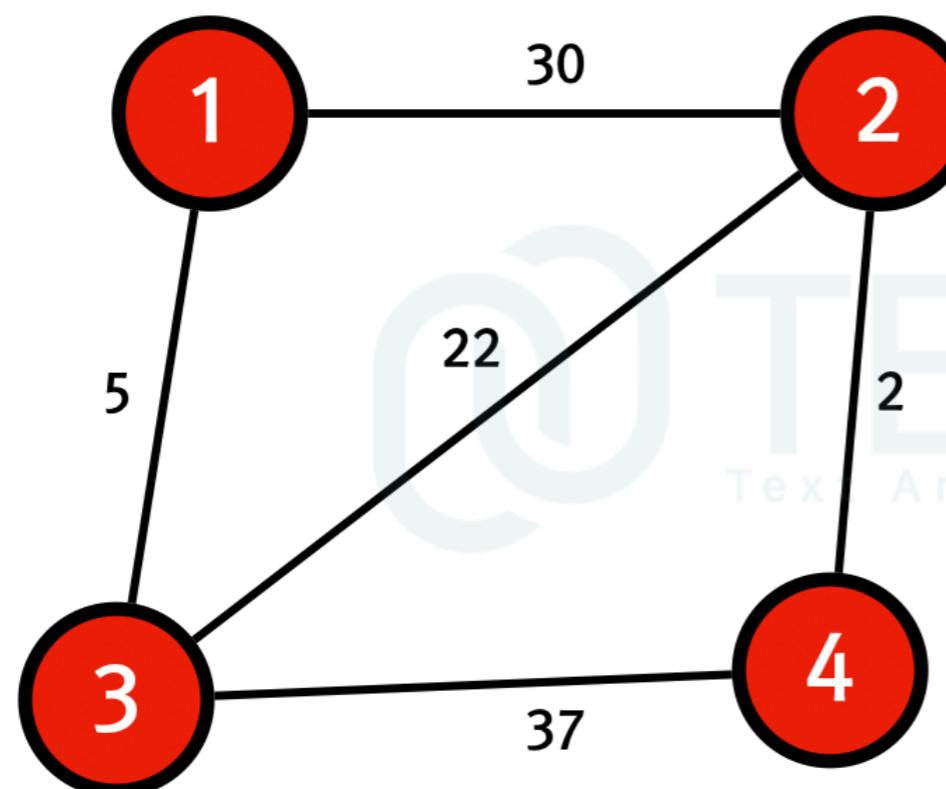
인접행렬 (*adjacency matrix*)

Vertex	1	2	3	4
1	-	1	1	0
2	1	-	1	1
3	1	1	-	1
4	0	1	1	-

단어 가중치: 네트워크 중심성

그래프 (Graph) 기본개념

- 경로 (path) : 간선에 의하여 연결된 노드들의 순차적 배열
- 최단 경로 (shortest path) : 그래프의 두 노드 간의 가장 짧은 경로
- 엣지 리스트 (edge list) : 노드와 노드 관계(경로)를 짹지어 목록으로 만든 것
- 가중치 (weight) : 네트워크에서 연결 관계의 강도를 나타내는 값



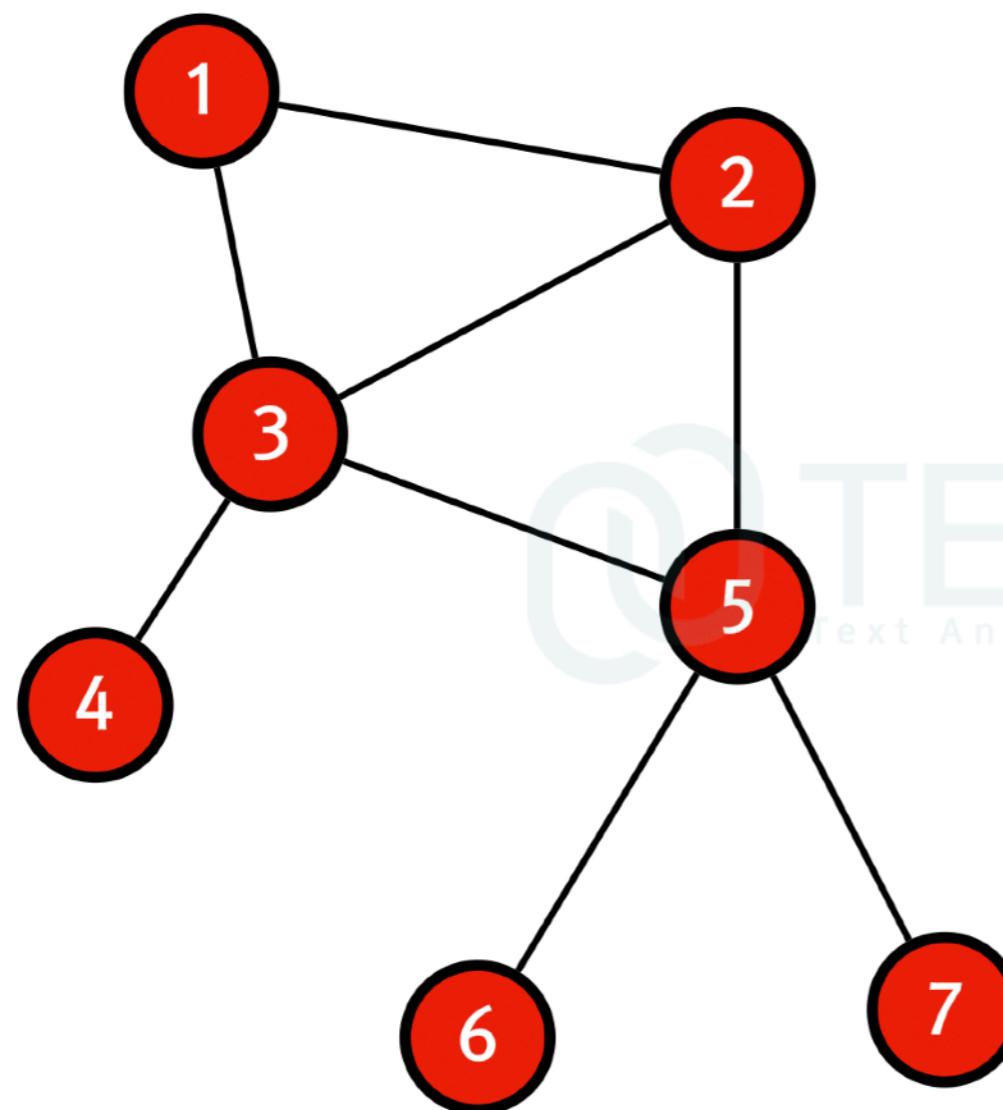
엣지리스트 (edge list)

Vertex	Vertex	Weight
1	2	30
1	3	5
2	3	22
2	4	2
3	4	37

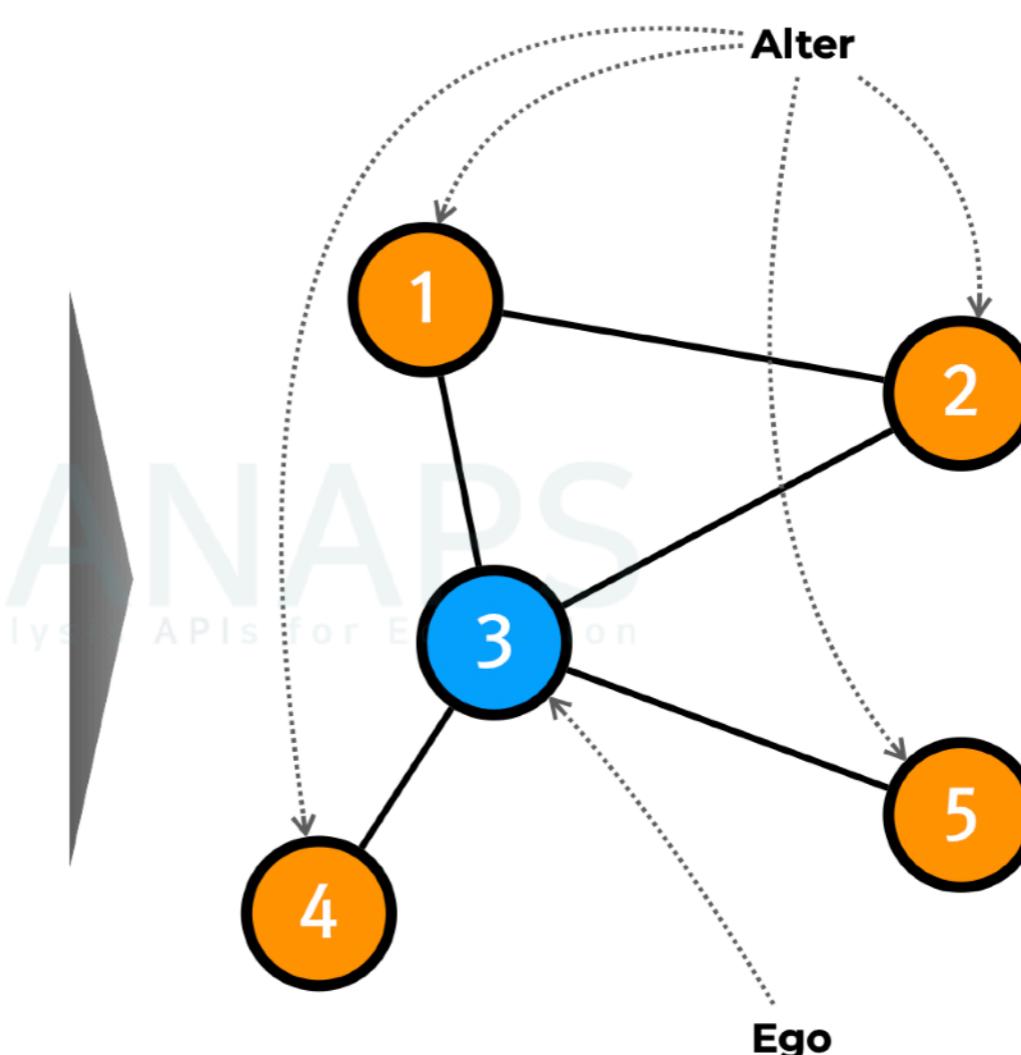
단어 가중치: 네트워크 중심성

그래프 (Graph) 기본개념

- **에고 네트워크 (ego network)** : 한 노드를 중심으로 다른 노드와의 연결관계를 표현한 네트워크



전체 네트워크 (Whole Network)

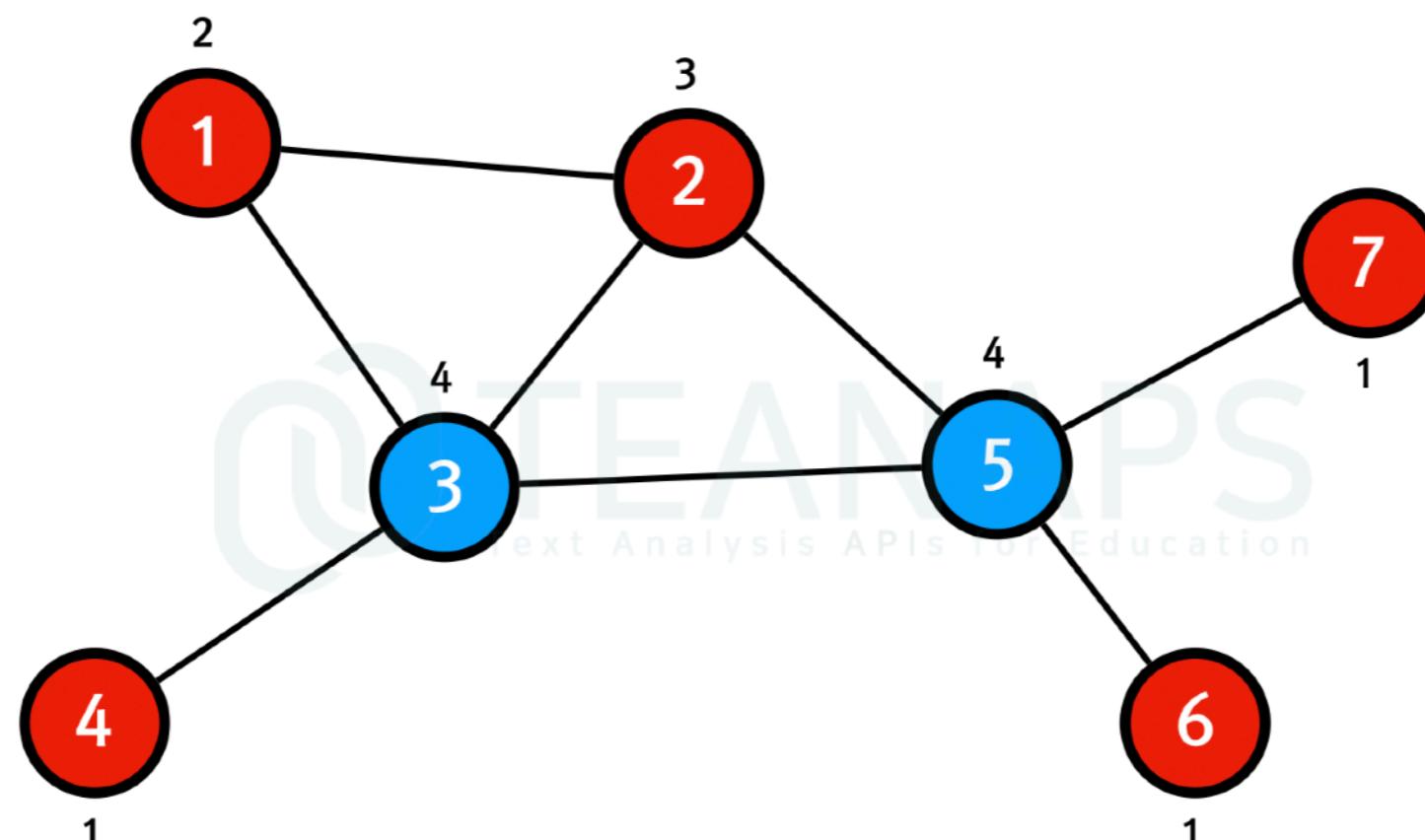


노드 3의 에고 네트워크

단어 가중치: 네트워크 중심성

연결 중심성 (Degree Centrality)

- 어떤 단어가 가장 많은 단어들과 같이 쓰였는가에 대한 척도
- 한 노드가 다른 노드와 연결된 엣지의 개수
- 비방향성 그래프에서는 한 노드로 연결될 수 있는 경로의 수
- 영향력 또는 인기도를 측정할 때 노드의 연결 정도의 척도로 사용
- 정보의 확산과 관련해 어느 노드가 중심이고, 다른 이웃 노드들에게 영향을 미치는지 평가할 때 사용



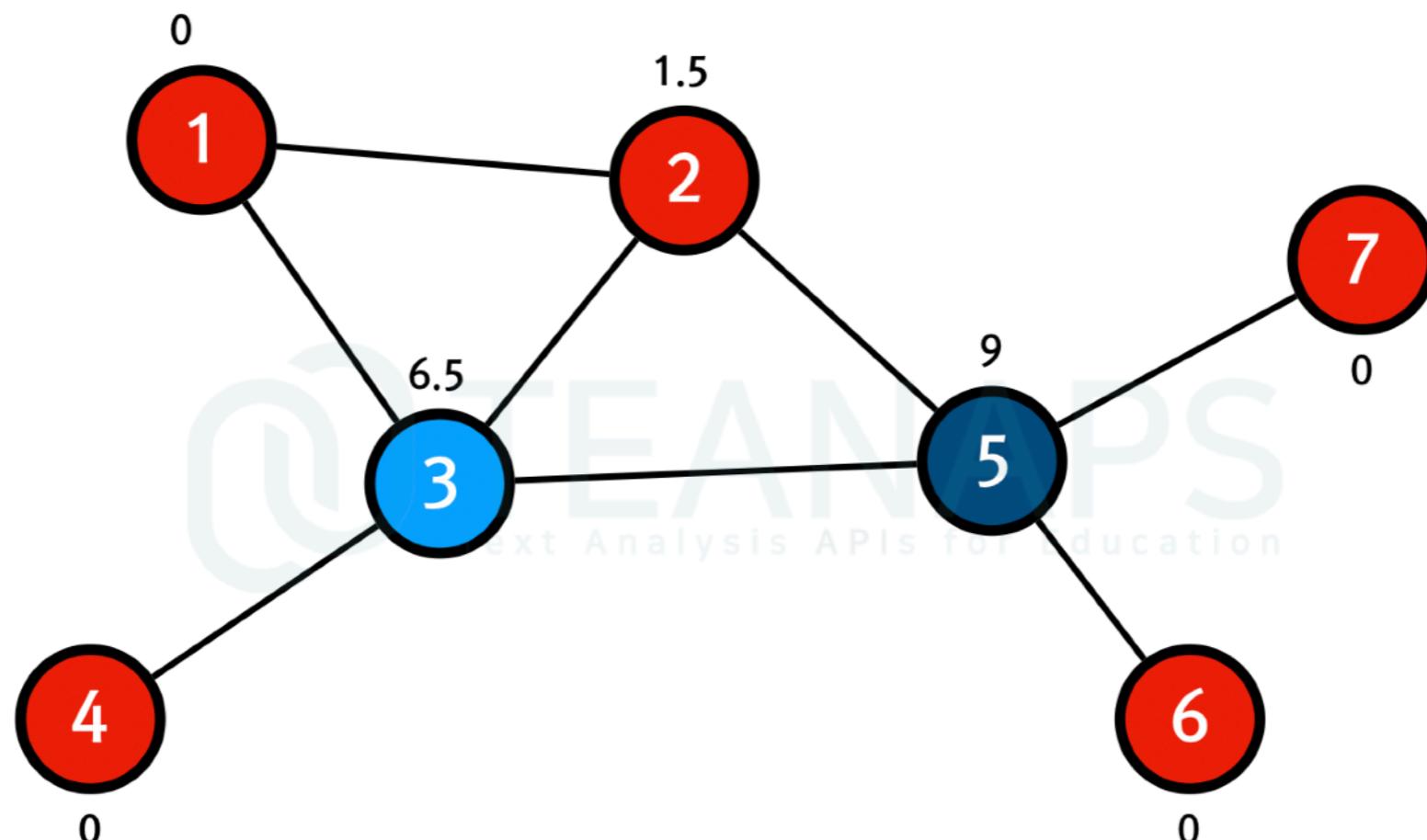
연결 중심성 (Degree Centrality)

단어 가중치: 네트워크 중심성

매개 중심성 (Betweenness Centrality)

- 어떤 단어가 다른 단어들 사이의 연결고리 역할을 하는가에 대한 척도
- 네트워크 내에서 한 노드가 다른 노드들 사이의 경로에 위치하는 정도
- 각 노드가 다른 노드들 간의 최단거리(*shortest path*)에 등장하는 빈도

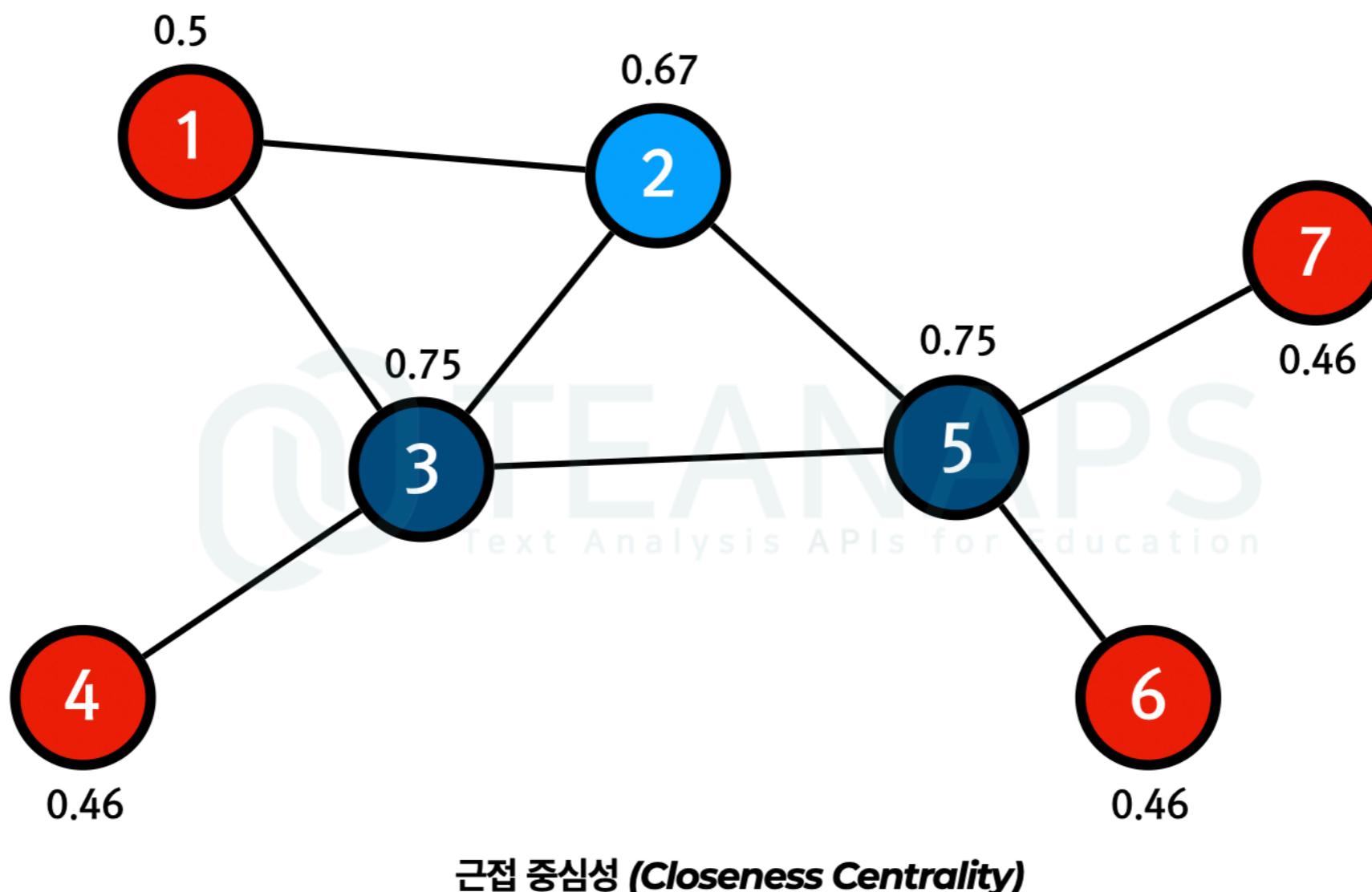
$$C_B(v) = \frac{i\text{와 } j \text{ 간의 최단경로 중 } v\text{를 지나는 경로의 수}}{i\text{와 } j \text{ 간의 최단경로의 수}} \quad i, j, v : \text{노드}$$



단어 가중치: 네트워크 중심성

근접 중심성 (Closeness Centrality)

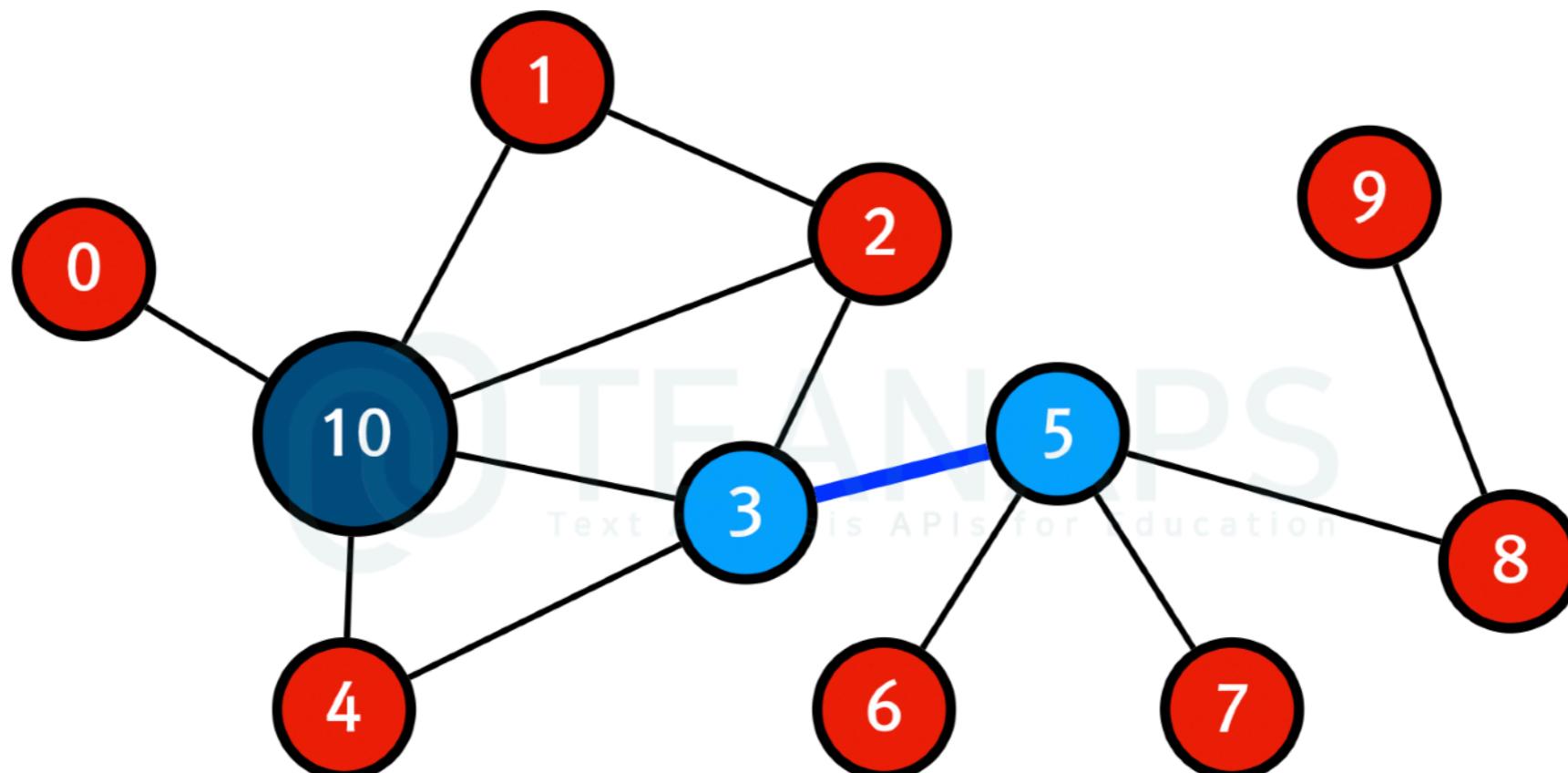
- 어떤 단어가 다른 단어들과의 가장 가까운 거리에 있는가에 대한 척도
- 한 노드에서 다른 모든 노드까지 모든 최단 경로의 평균 또는 이의 역수
- 모든 다른 노드에 도달하는데 까지 평균 소요 시간



단어 가중치: 네트워크 중심성

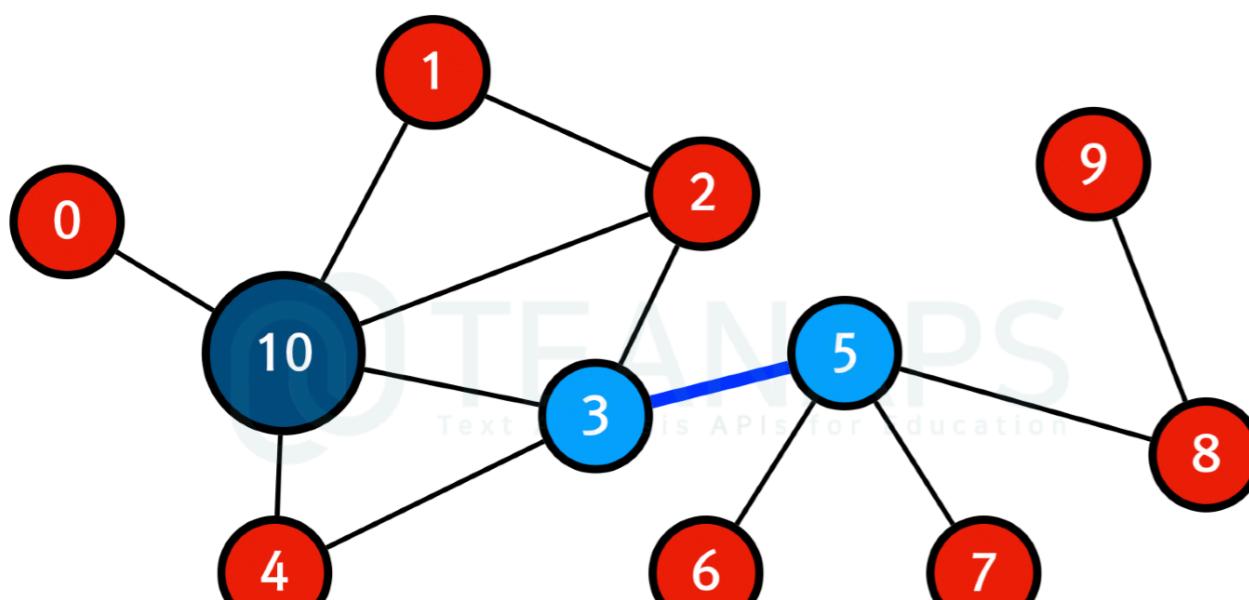
| 네트워크 중심성 척도의 활용

- 분석의 목적에 따라 척도를 다르게 적용하여 분석에 활용 (중심 노드를 예고네트워크 또는 연결된 N개의 노드 단위로 고려해도 됨)
 - 1) 노드 10은 연결 중심성 측면에서 가장 중심에 있음
 - 2) 노드 3과 노드 5는 매개 중심성 측면에서 노드 10 보다 더 중심에 있음
 - 3) 또한 노드 3과 노드 5 사이의 관계는 네트워크가 분리될 수 있는 중요한 연결로 볼 수 있음
 - 4) 다른 조건들이 동일할 때, 3과 5는 10보다 네트워크의 중심에 있음



단어 가중치: 네트워크 중심성

네트워크 중심성 계산



Sample Graph

노드리스트 (node list)

Node	Degree Centrality	Betweenness Centrality	Closeness Centrality
0			
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			

텍스트 데이터 시각화 (Visualization)

테이블 (Table)

- 분석결과를 테이블 형태로 행과 열을 구분하여 표현하는 방법

표1 '아쿠르트 아줌마' 연관어 변화											
2013년			2014년			2015년			2016년		
No.	연관어	언급 비중	No.	연관어	언급 비중	No.	연관어	언급 비중	No.	연관어	언급 비중
1	아쿠르트	21.3%	1	아쿠르트	26.3%	1	아쿠르트	26.6%	1	아쿠르트	13.1%
2	먹다	4.9%	2	건강	4.5%	2	집	4.7%	2	콜드브루	8.2%
3	아침	4.4%	3	아침	4.0%	3	아침	4.4%	3	커피	7.4%
4	엄마	4.2%	4	집	3.6%	4	맛	3.9%	4	맛	6.6%
5	집	3.5%	5	제품	3.4%	5	먹다	3.4%	5	끼리	5.7%
6	오다	2.8%	6	엄마	3.3%	6	사다	2.8%	6	치즈	5.3%
7	사다	2.7%	7	맛	2.7%	7	주다	2.8%	7	과자	5.0%
8	주다	2.5%	8	같다	2.6%	8	다니다	2.7%	8	아메리카노	4.1%
9	구입하다	2.4%	9	우유	2.6%	9	엄마	2.6%	9	먹다	3.3%
10	아이	2.4%	10	주다	2.2%	10	우유	2.1%	10	크림치즈	3.1%
11	아쿠르트 주다	2.3%	11	먹다	2.2%	11	만나다	2.1%	11	라떼	2.8%
12	배달하다	2.3%	12	만나다	2.0%	12	제품	2.0%	12	만나다	2.7%
13	수입	2.3%	13	사다	1.9%	13	사진	2.0%	13	가격	2.4%
14	다니다	2.1%	14	알다	1.9%	14	나오다	2.0%	14	찾다	1.9%
15	얼려먹다	2.0%	15	배달하다	1.8%	15	팔다	1.9%	15	아침	1.8%
16	살다	2.0%	16	다니다	1.8%	16	지나가다	1.8%	16	10일	1.6%
17	제품	2.0%	17	하루야채	1.7%	17	하나	1.7%	17	엄마	1.5%
18	세븐	1.8%	18	나누다	1.7%	18	판매	1.7%	18	우유	1.4%
19	가다	1.8%	19	지나가다	1.6%	19	일하다	1.6%	19	팔다	1.3%
20	자녀	1.8%	20	세븐	1.5%	20	오다	1.6%	20	발견하다	1.3%
21	만나다	1.8%	21	수입	1.5%	21	찾다	1.6%	21	사다	1.2%
22	마시다	1.7%	22	찾다	2.3%	22	음료	1.5%	22	인기	1.2%
23	유산균	1.7%	23	노인	1.4%	23	마시다	1.4%	23	편의점	1.2%
24	일하다	1.7%	24	마시다	1.4%	24	길	1.4%	24	끼리딥앤크런치	1.1%
...				
29	팔다	1.4%	29	묻다	1.3%	29	배달하다	1.3%	29	구입하다	1.0%

■ 상승 키워드 ■ 하락 키워드 ■ 신규 키워드

<표 6> 불행요인 세부 토픽 모델링 결과

#	토 픽	키 워 드
1	가정 불화	불행, 사랑, 가족, 집, 아버지, 가정, 부모
2	가난	분배, 돈, 소득, 빈곤, 경제, 가난
3	자녀 문제	학교, 위험, 아이, 행동, 상황
4	부정적 인생관	불행, 사람, 인생, 마음, 성공, 공통점
5	인간관계 문제	불행, 자신, 관계, 마음, 생각, 환경, 상황
6	직업 불만족	불행, 사람, 생각, 인생, 직업, 친구
7	건강 문제	불행, 건강, 수명, 질병, 생명, 병, 사고
8	미 취업	오늘, 운세, 불행, 건강, 취업, 뱀띠, 금전
9	부정적 마음가짐	불행, 사람, 마음, 생각, 이기심, 자만심, 피해의식
10	-	예수, 교회, 신앙, 설교, 말씀, 축복

Table 10. Top Seller Characteristics of Rescator

#	Top key words	Interpretation
5	shop, wmz, icq, webmoney, price, dump,	Product: CCs, dumps (valid, verified);
6	валид (valid), чекер (checker), карты (cards), баланс (balance), карт (cards)	Payment: wmz, webmoney, bitcoin, lesspay;
8	shop, good, CCs, bases, update, cards, bitcoin, webmoney, validity, lesspay	Contact: shop, register, deposit, e-mail, icq, jabber
11	dollars, dumps, deposit, payment, sell, online, verified	
16	e-mail, shop, register, icq, account, jabber,	

텍스트 데이터 시각화 (Visualization)

워드클라우드 (Wordcloud)

- 단어의 가중치 (TF, TF-IDF, 중심성 등)를 단어의 크기로 반영하여 그 분포를 아름답게 표현하는 방법
- 가중치를 비롯해 단어의 색깔, 배치 등을 통해 더 많은 정보를 표현할 수 있음



* NÉSTOR CORREA, Cómo implementar el Big Data en tu empresa, 2017., <http://bluelight.tistory.com/298/>.

** 워드클라우드.kr, 열정 긍정적 - 워드클라우드, 2017.11.5., <http://wordcloud.kr/1295/>.

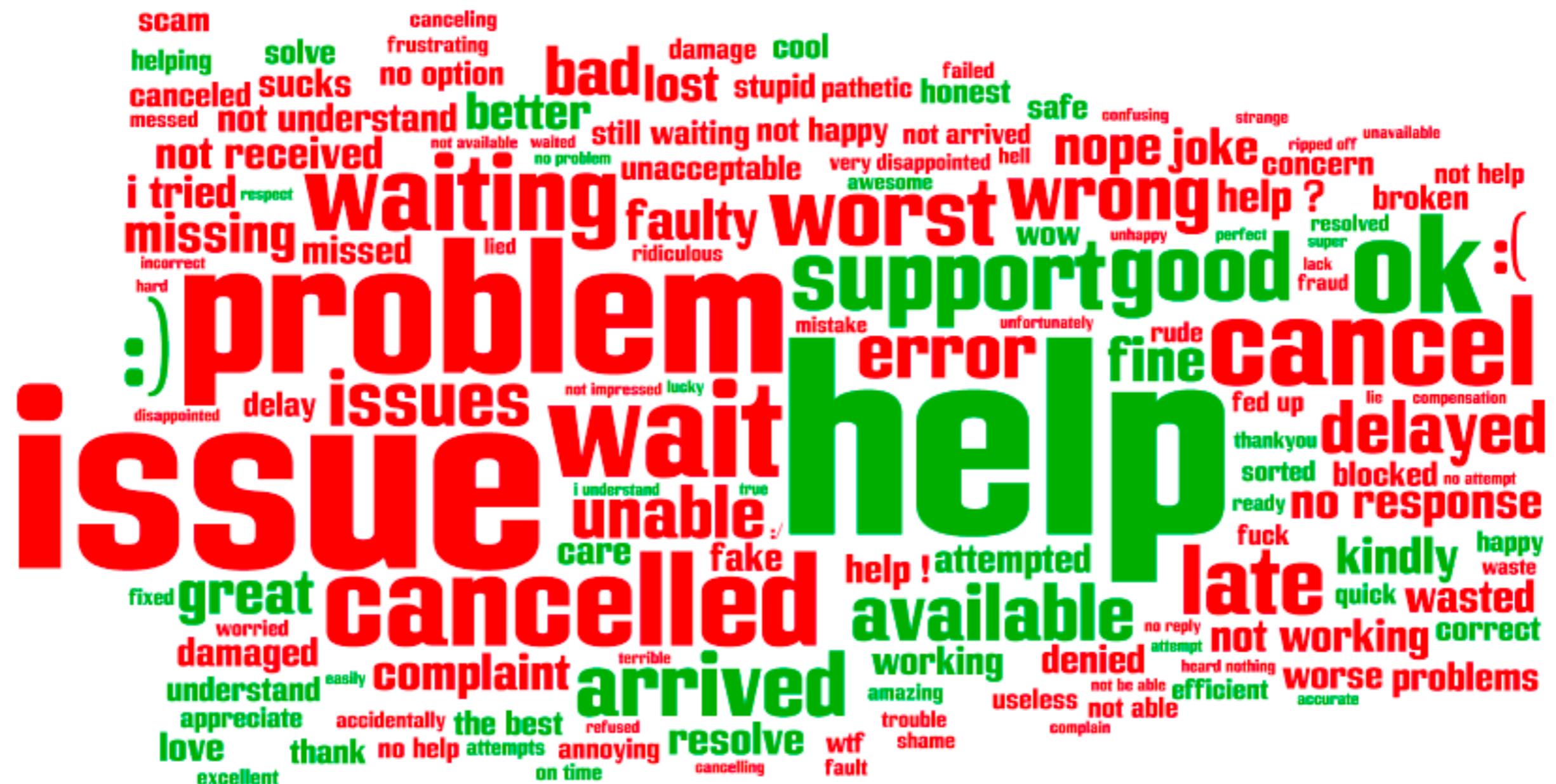
*** Kumo - Java Word Cloud, <http://kennycason.com/posts/2014-07-03-kumo-wordcloud.html/>.

**** CX DATA SCIENCE, SIMPLY SENTIMENT 2 SUPPORT, https://www.cxdatascience.com/ssv2_support

***** 네이버블로그(지그드시), WordCloud(워드클라우드) 만들어 주는 사이트 사용 방법 안내 - taxedo, 2016.2.3., [https://blog.naver.com/liberty264/220616866103/](https://blog.naver.com/liberty264/220616866103).

텍스트 데이터 시각화 (Visualization)

워드클라우드 (Wordcloud)



텍스트 데이터 시각화 (Visualization)

워드클라우드: 무엇이 잘못되었을까요?



* How we build MyRealTrip. 마이리얼트립 여행 후기 데이터 분석

<https://medium.com/myrealtrip-product/%EB%A7%88%EC%9D%B4%EB%A6%AC%EC%96%BC%ED%8A%B8%EB%A6%BD-%EC%97%AC%ED%96%89-%ED%9B%84%EA%B8%BO-%EB%8D%BO%EC%9D%B4%ED%84%BO-%EB%B6%84%EC%84%9D-be3f6c557ca2/>

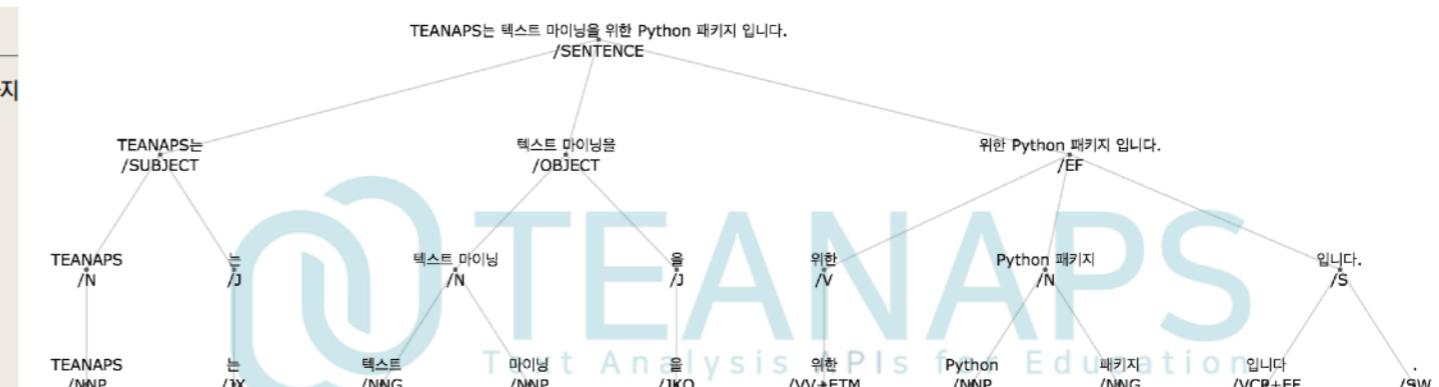
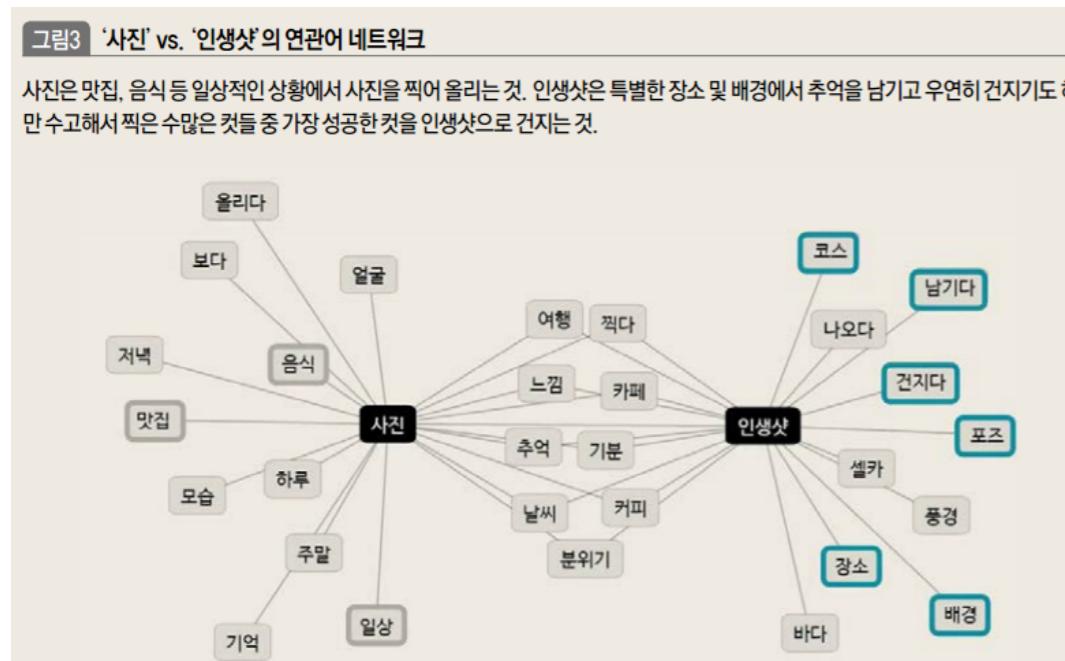
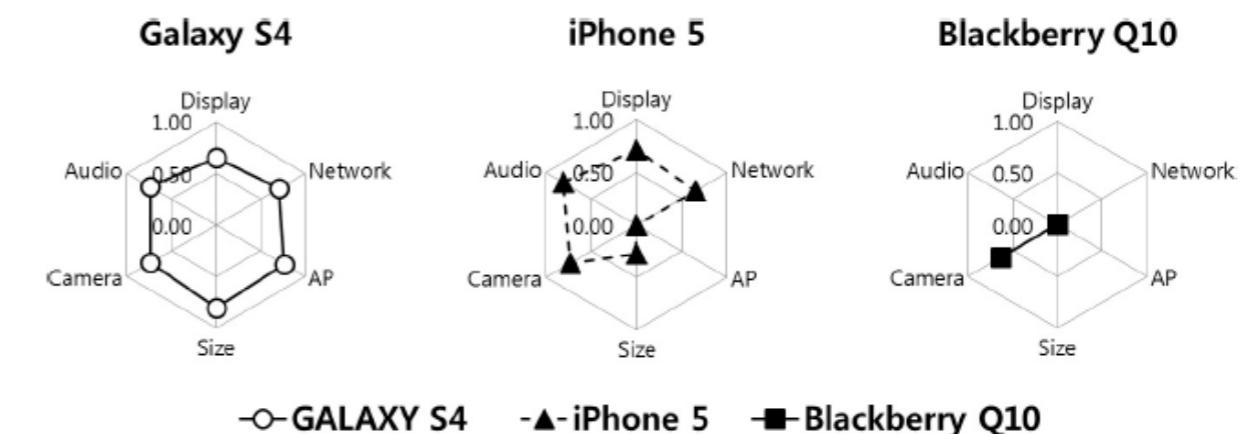
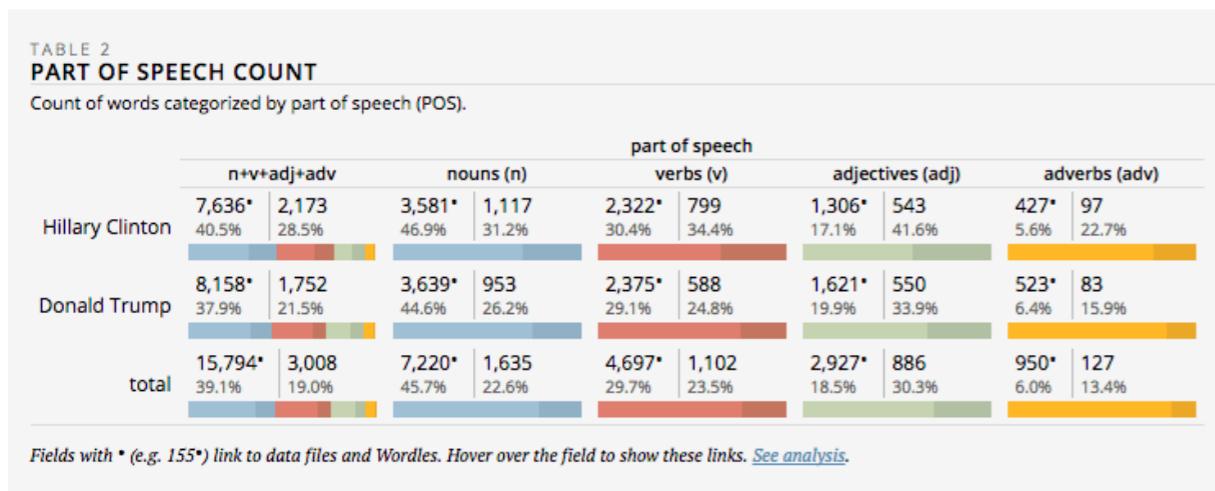
https://www.donga.com/graphics/qdbr_no/695

*** references

텍스트 데이터 시각화 (Visualization)

그래프와 네트워크 (Graph & Network)

- 단어 사이의 관계와 그 강도를 선으로 연결하여 표현하는 방법
- **그래프 (graph)** : 문서 또는 단어의 정량화된 특징을 도표로 표현하는 방법
- **네트워크 (network)** : 단어를 노드, 단어들 사이의 관계를 엣지로 취급하여 네트워크를 표현하는 방법



* 백경혜(DBR), “매력을 소비하는 나는 덕후! 즐거움을 위해 기꺼이 지갑을 연다”, 2017.1., http://dbr.donga.com/article/view/1203/article_no/7935.

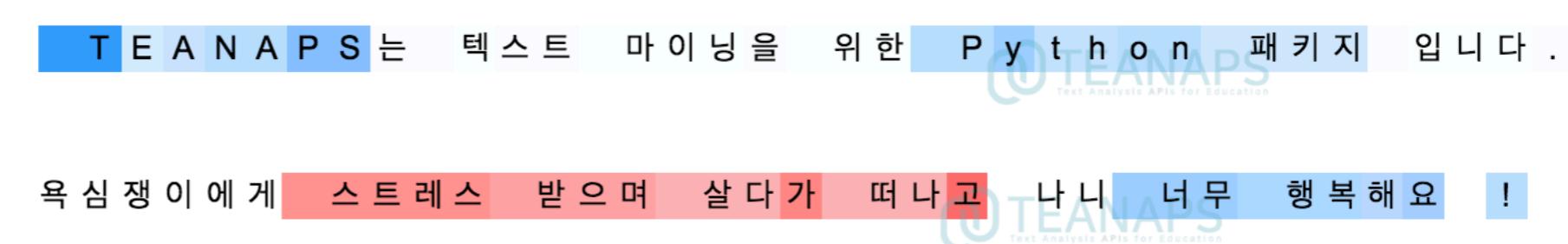
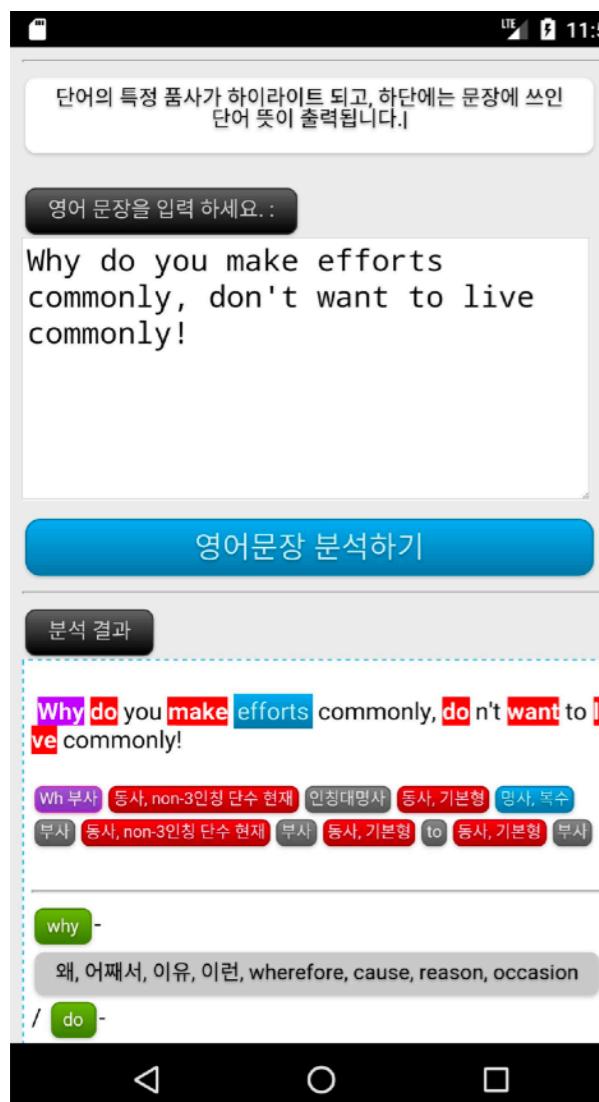
** 최홍규(슬로우뉴스), 2016 미국 대선을 보여주는 텍스트 마이닝 분석방법들, 2017.1.9., <http://slownews.kr/60919>.

*** Kim et al. (2014). Analysis on smartphone related twitter reviews by using opinion mining techniques. In Advanced Approaches to Intelligent Information and Database Systems (pp. 205-212).

텍스트 데이터 시각화 (Visualization)

강조표시 (Highlighting)

- 문서 또는 문장의 일부를 색상으로 강조하여 표현하는 방법
- 음절, 단어, 문장 단위로 강조범위를 지정할 수 있으며 색상에 따라 다양한 특징을 표현할 수 있음



where you remembered them . It was slow and quiet , like breathing . I watching the tree . Through in the kitchen , we could hear mum shouting Bits and pieces and starts and stops . Havering , just . Not like this . Not s looking down . Not straight ahead . Not at the telly . At us , sitting on the train . We looked at each other . Sandra was scared . ' Mum ! ' she called . ' Stop us right . ' I patted his knee to tell him not to be so daft . And what I ? You're just kidding . Just having us on . ' By the time Mum came through , for all the sign of it left there . ' What is it ? ' she said . We big , we opened the presents . They were all under the tree in their piles - I . We finished and sat around him , eating chocolate . There was a small bag in it . He sat there , fingering the parcel . The label said it was from Gran in it . He was happy just to have it sitting on his lap . He sorted the parcel , tweaked at the parcel a bit , pulled the sellotape off . He didn't mind as a bunnet . Soft and checked , and with the label inside still white and same as the last one , only new . But we wanted to see it on . So mum put the parcel in the chair . Uncle Tom who was about to go out - just stopped in his chair for a minute and said , ' Come on , Tom ! ' Mum ran and mum and Uncle Tom and the boys and us and the dog , even - and took a photo of everybody crouched together round the big , high chair . Then they went into the living room . " Quick ! Come and see this ! " We thought it was a niddle of the room , among the scrunched paper and the walnut shells , where he door on the way to his Christmas dinner , still with his bunnet on . Mum , holding his hands . He knew what to do . It was like dancing . He

E.O.D

Contact

-  <http://www.teanaps.com>
-  fingeredman@gmail.com