

ADVANCED TEXT MINING

by FINGEREDMAN (fingeredman@gmail.com)

WEEK 11

Network Analysis



단어 가중치: 네트워크 중심성

Review

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

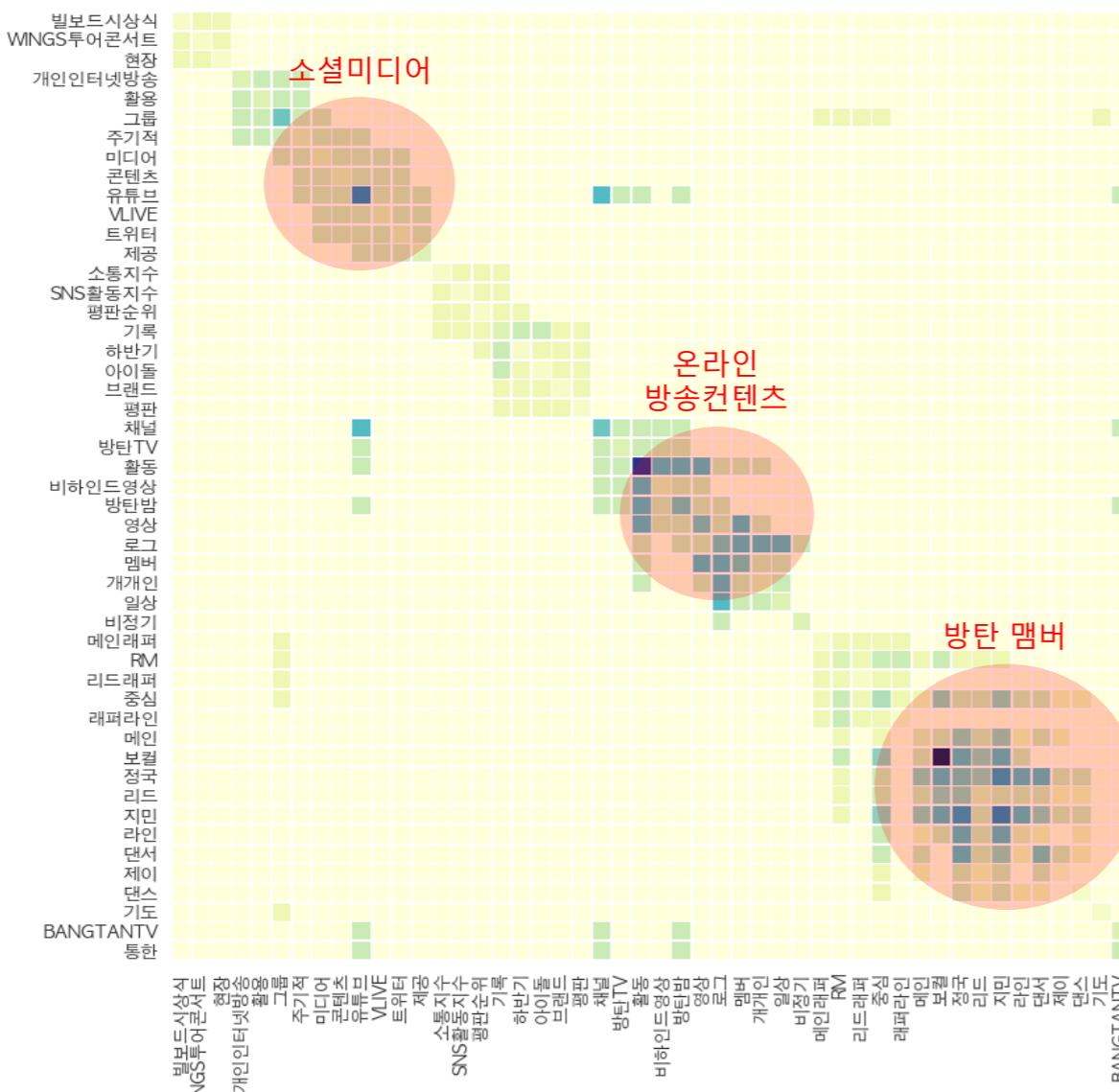


표1 '아쿠르트 아줌마' 연관어 변화

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%						

상승 키워드 하락 키워드 신규 키워드

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

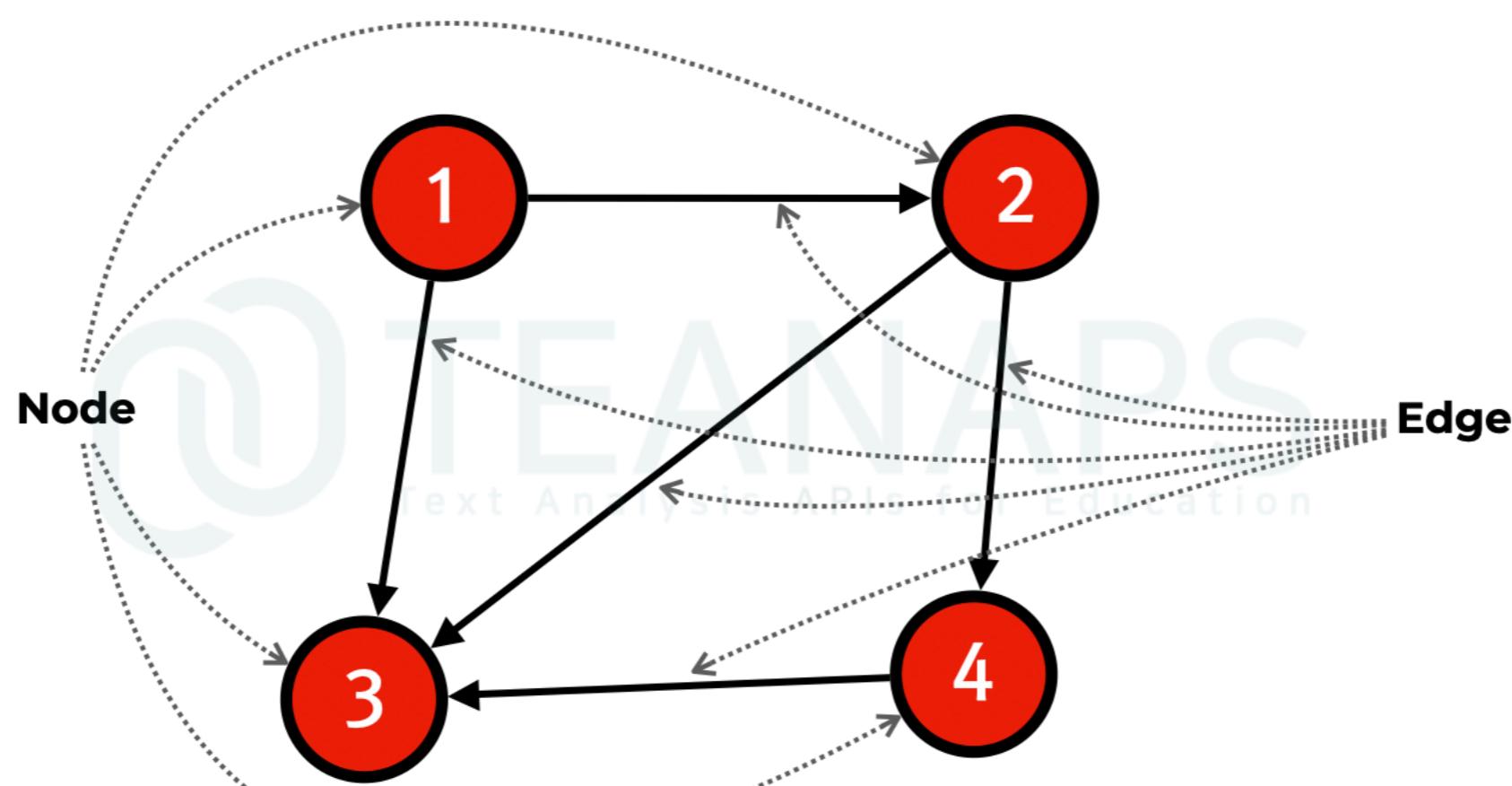
** 백경혜(DBR), "매력을 소비하는 나는 덕후! 즐거움을 위해 기꺼이 지갑을 연다", 2017.1., http://dbr.donga.com/article/view/1203/article_no/7935/.

*** references

단어 가중치: 네트워크 중심성

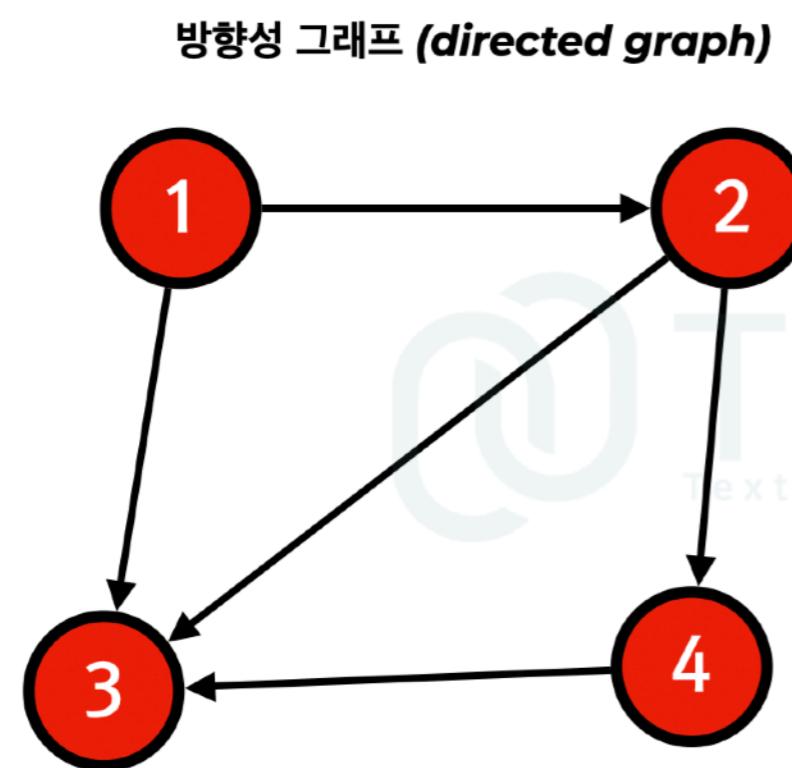
그래프 (Graph) 기본개념

- 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



엣지리스트 (edge list)

Vertex Vertex

1	2
1	3
2	3
2	4
3	4

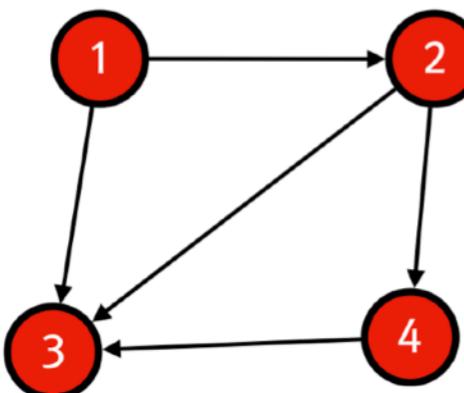
인접행렬 (adjacency matrix)

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*)

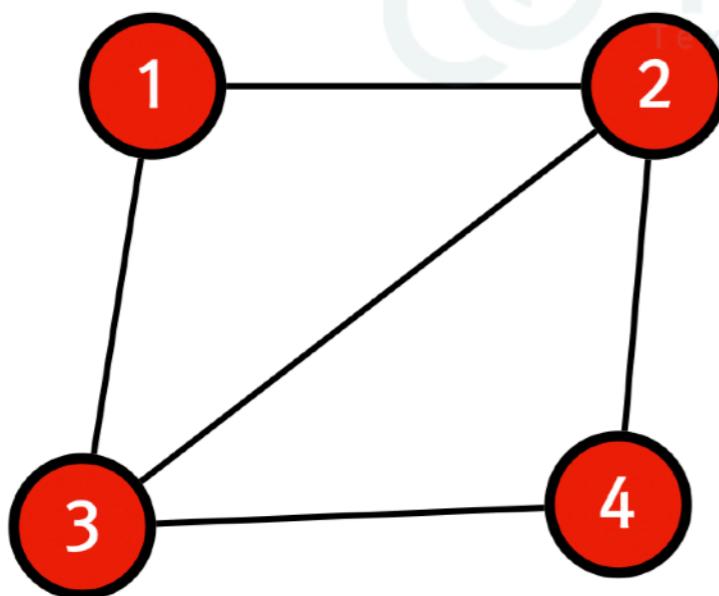


엣지리스트 (*edge list*)

Vertex Vertex

1	2
1	3
2	3
2	4
3	4

비방향성 그래프 (*undirected graph*)



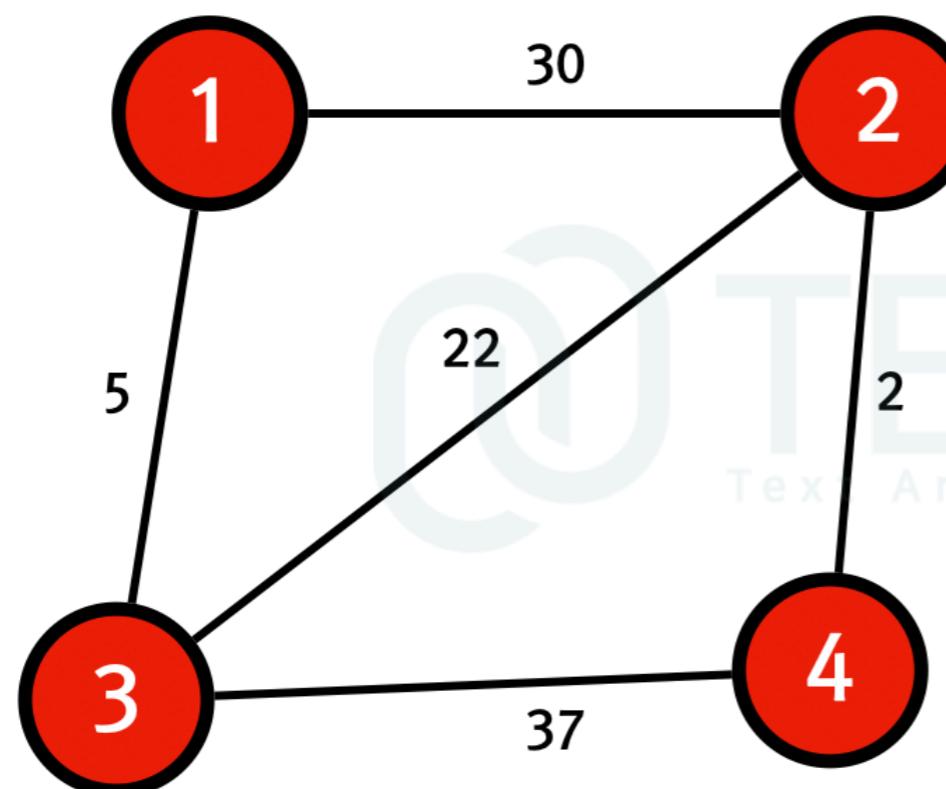
인접행렬 (*adjacency matrix*)

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



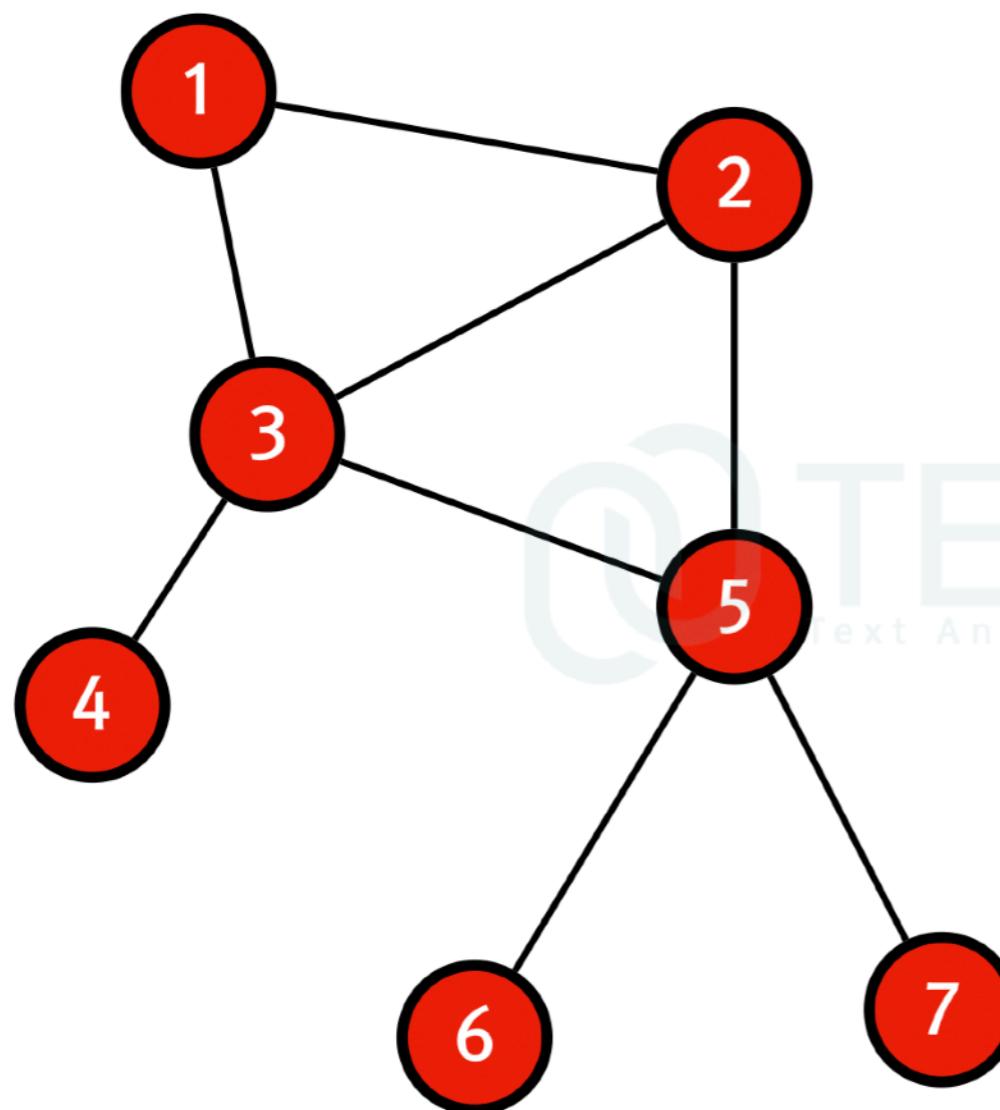
엣지리스트 (edge list)

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

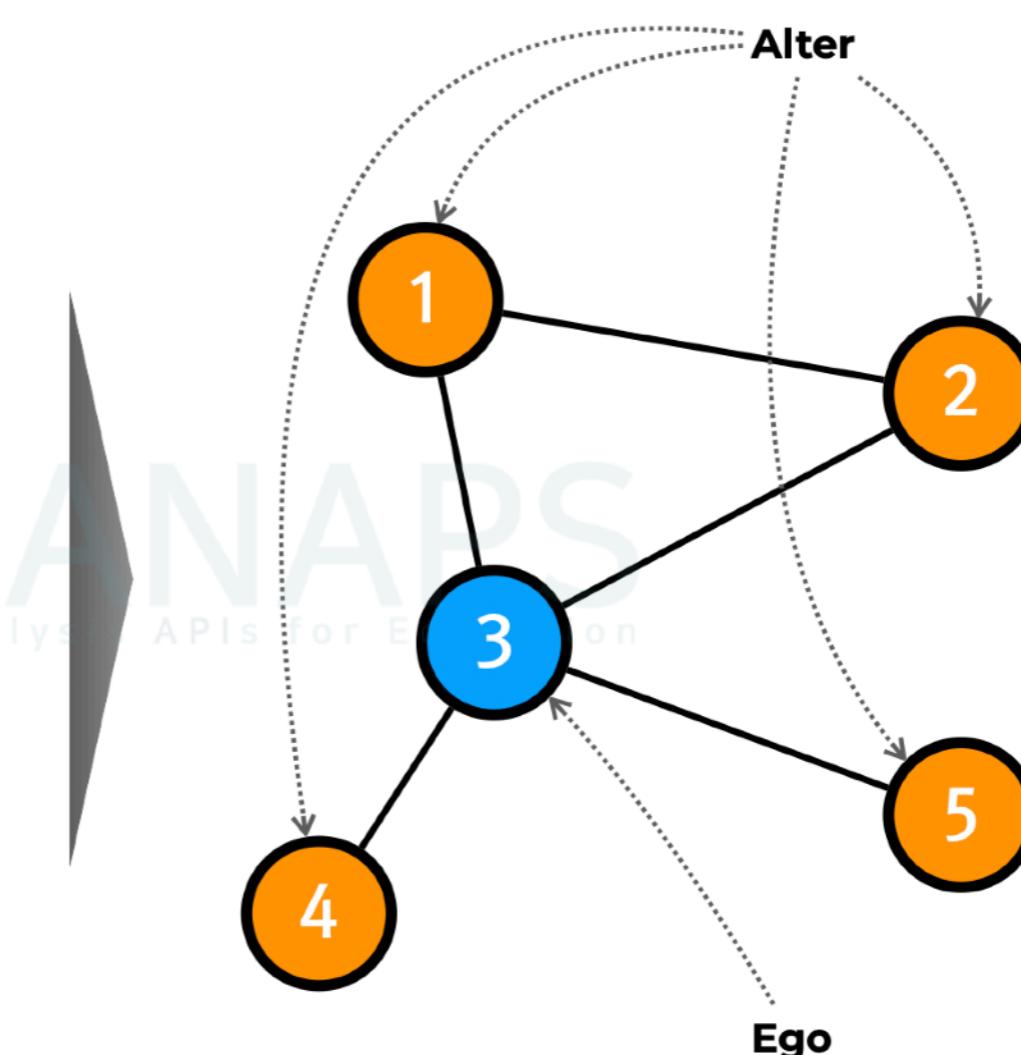
단어 가중치: 네트워크 중심성

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: 동시출현 빈도

동시출현 분석 (Co-word/Co-occurrence Analysis)

- 문서에 서로다른 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

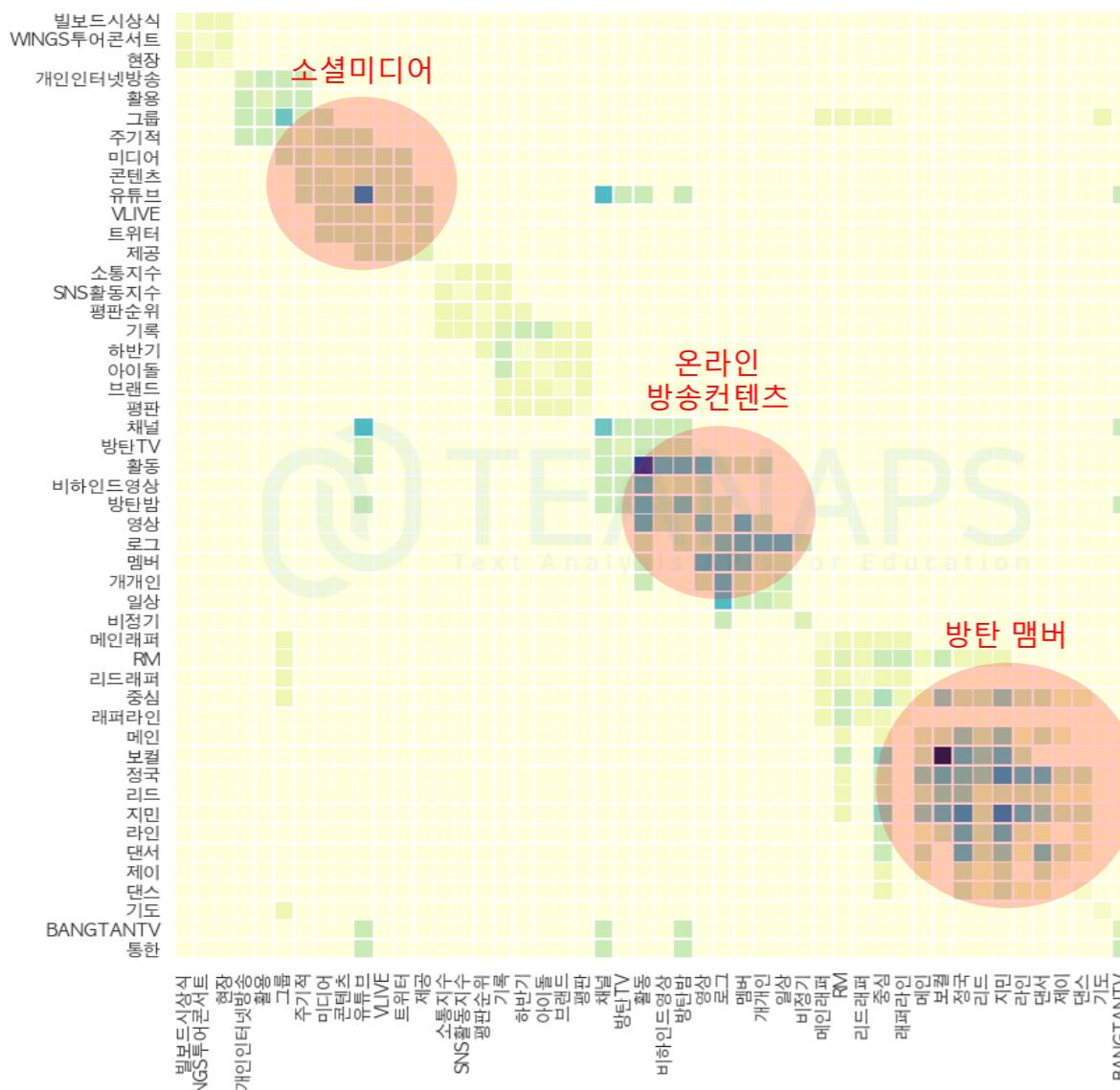


표1 '아쿠르트 아줌마' 연관어 변화

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.

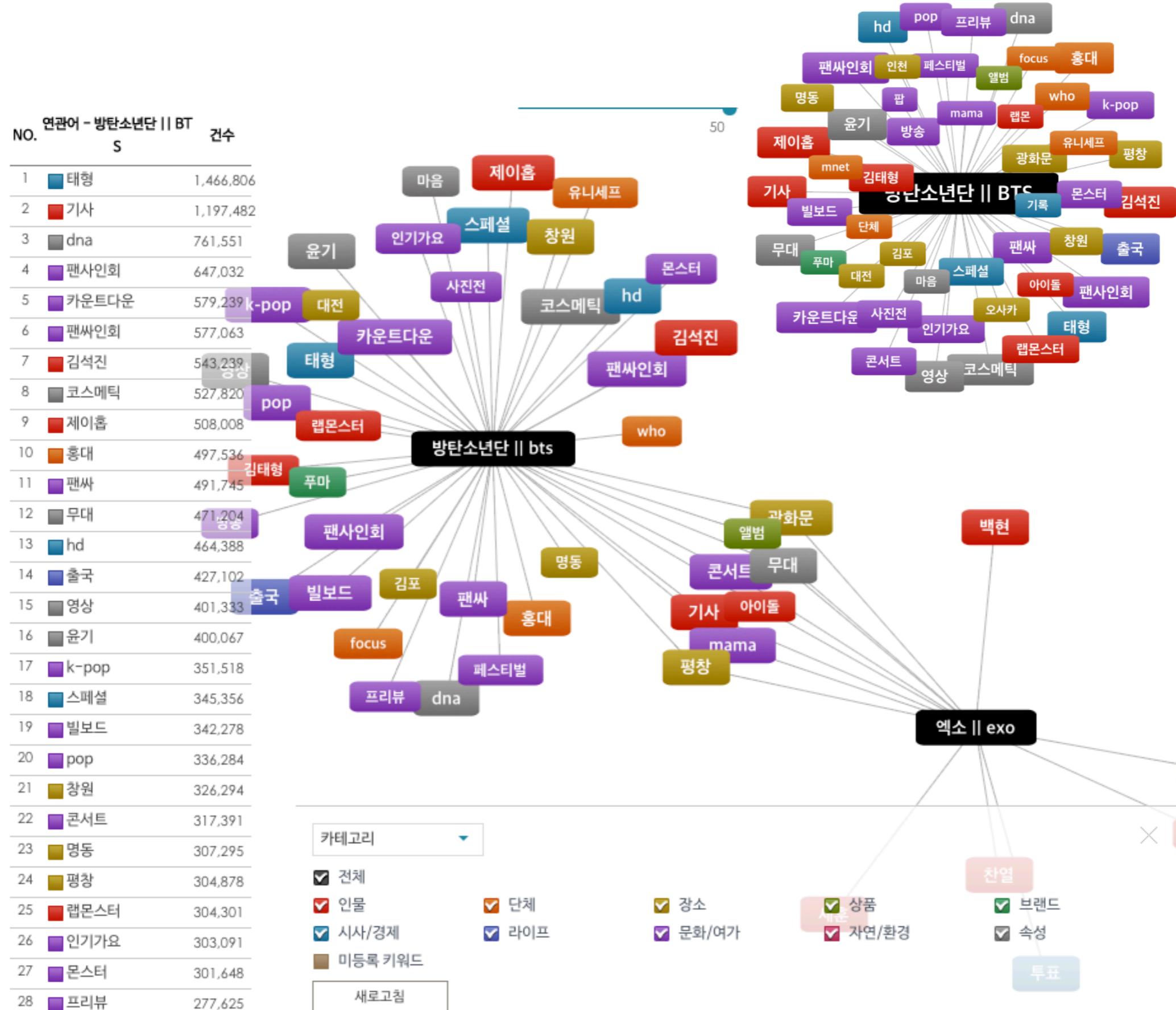
2013년			2014년			2015년			2016년		
No.	연관어	언급 비중	No.	연관어	언급 비중	No.	연관어	언급 비중	No.	연관어	언급 비중
1	아쿠르트	21.3%	1	아쿠르트	26.3%	1	아쿠르트	26.6%	1	아쿠르트	13.1%
2	먹다	4.9%	2	건강	4.5%	2	집	4.7%	2	콜드브루	8.2%
3	아침	4.4%	3	아침	4.0%	3	아침	4.4%	3	커피	7.4%
4	엄마	4.2%	4	집	3.6%	4	맛	3.9%	4	맛	6.6%
5	집	3.5%	5	제품	3.4%	5	먹다	3.4%	5	끼리	5.7%
6	오다	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

* 저번주, 시하은해 파이썬으로 시작하는 데이터분석: 텍스트 마이닝 기초, 2018.12.1

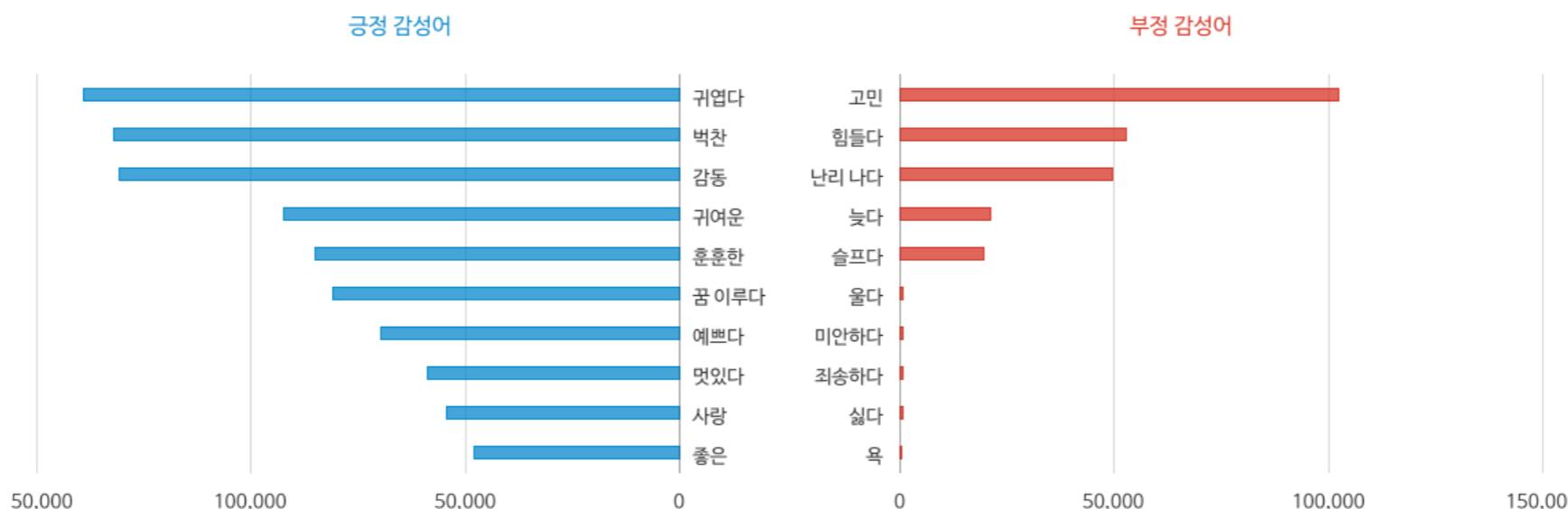
****간경환, 간경환은 쌍피아그룹도 시국인은 대나무나무다**. 쿠데타는 미안이었지만, 2016.12.12.

*** 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	■ 시우민	98,662
24	■ 기사	98,416
25	■ 아이돌	95,535
26	■ 고척스카이돔	88,950
27	■ music	88,537
28	■ 올림픽	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

* 소셜메트릭스, 2017.11.3., <http://www.socialmetrics.co.kr/>.

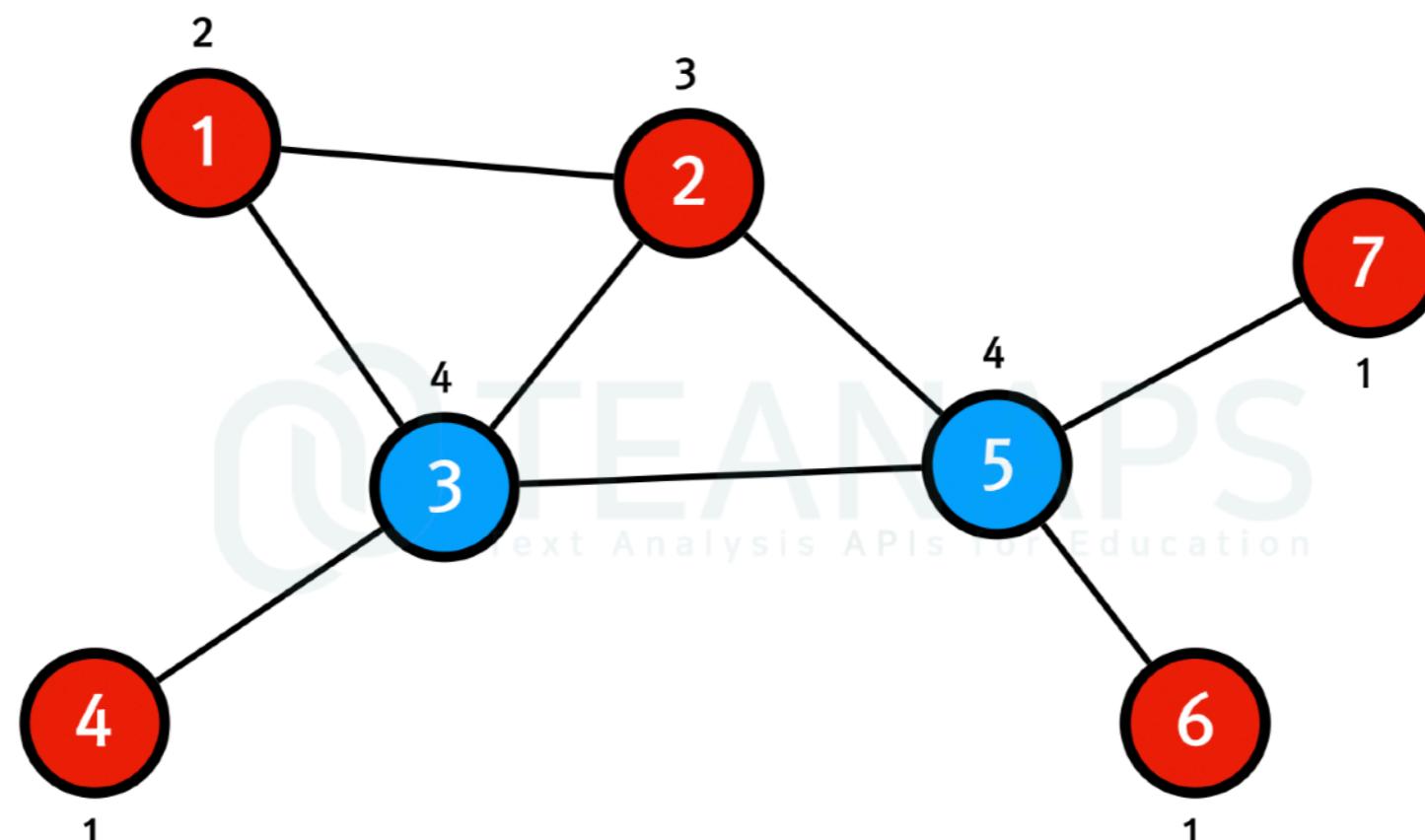
** references

*** references

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



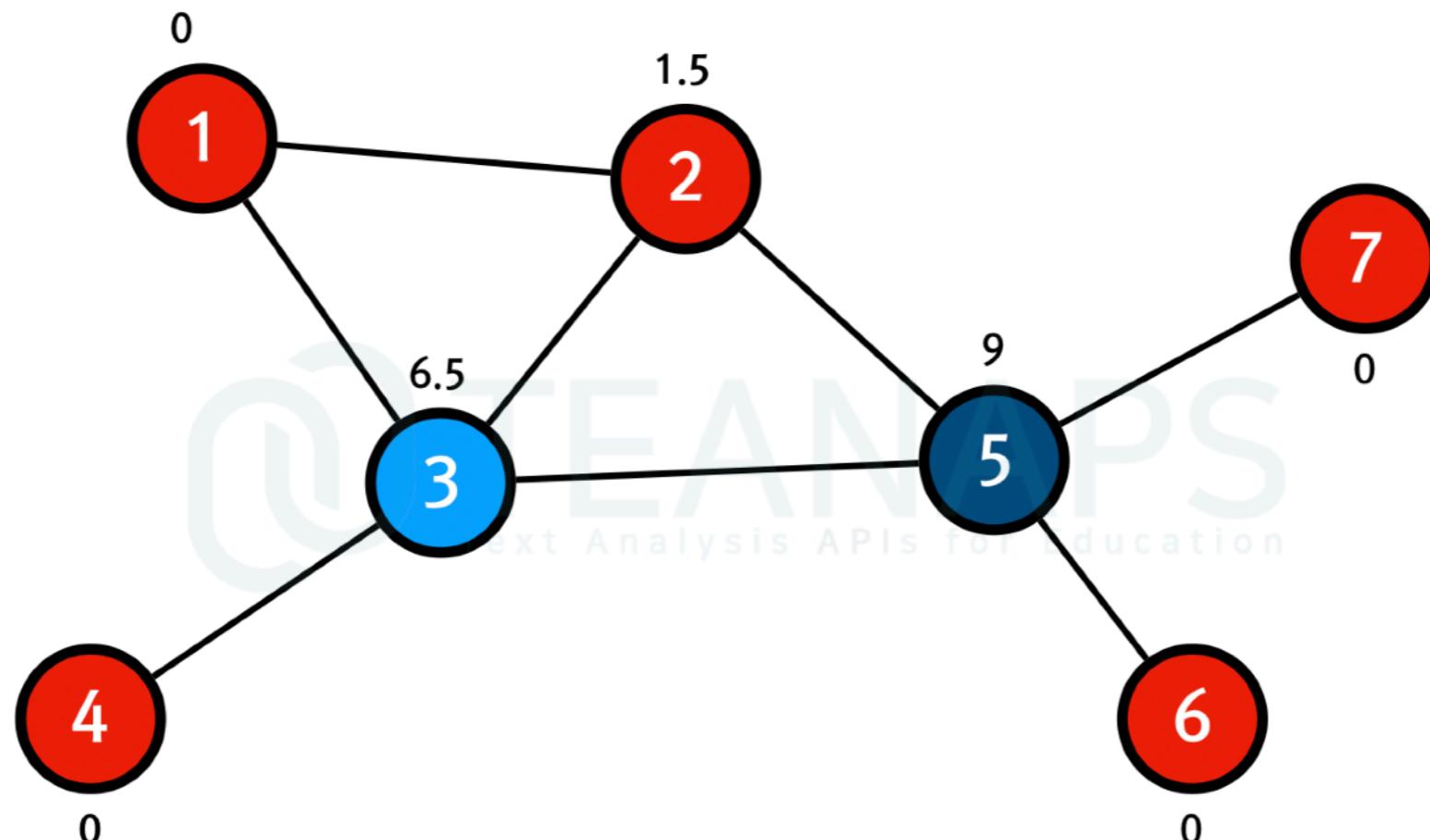
연결 중심성 (Degree Centrality)

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

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

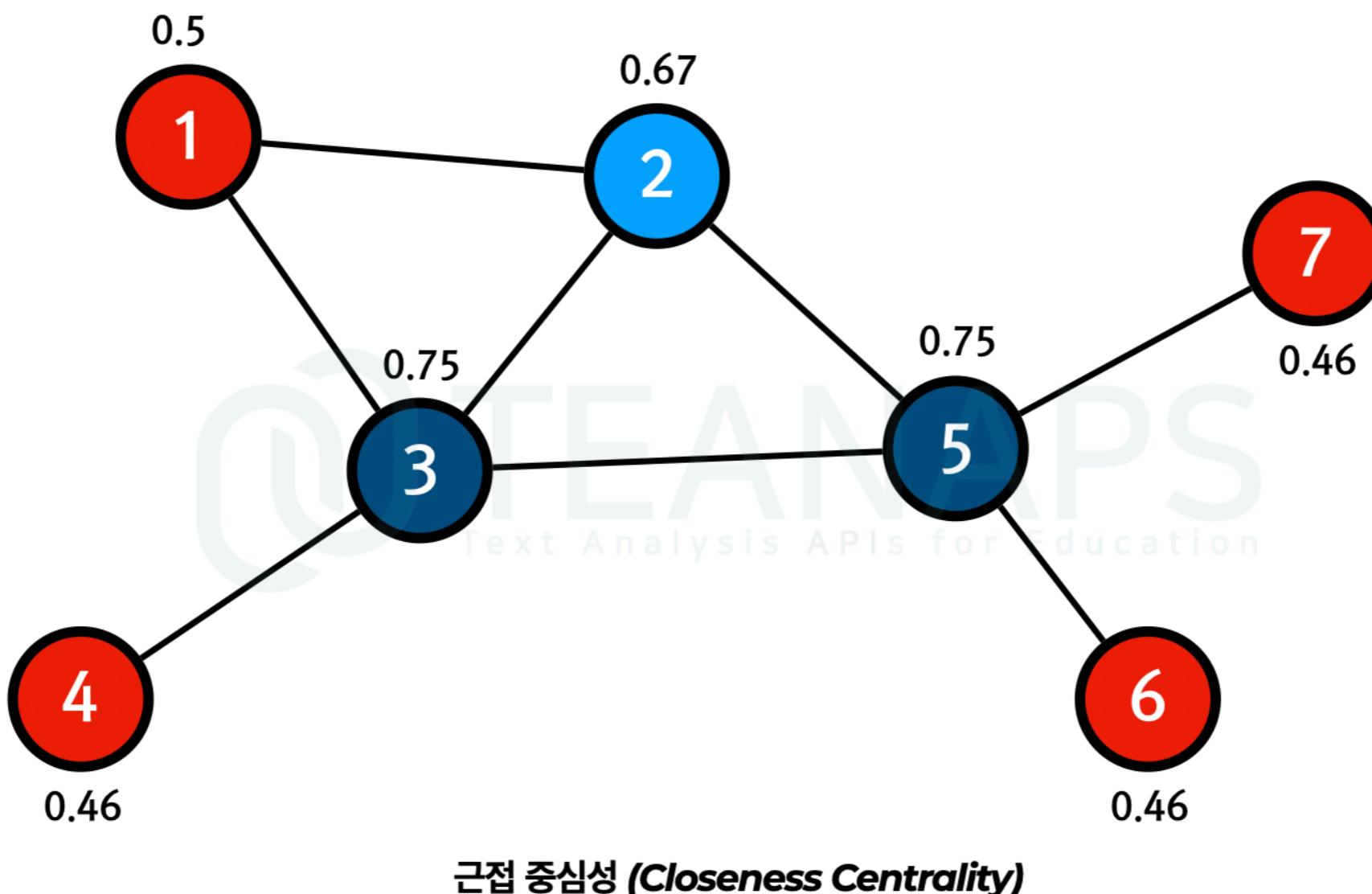


매개 중심성 (Betweenness Centrality)

단어 가중치: 네트워크 중심성

근접 중심성 (Closeness Centrality)

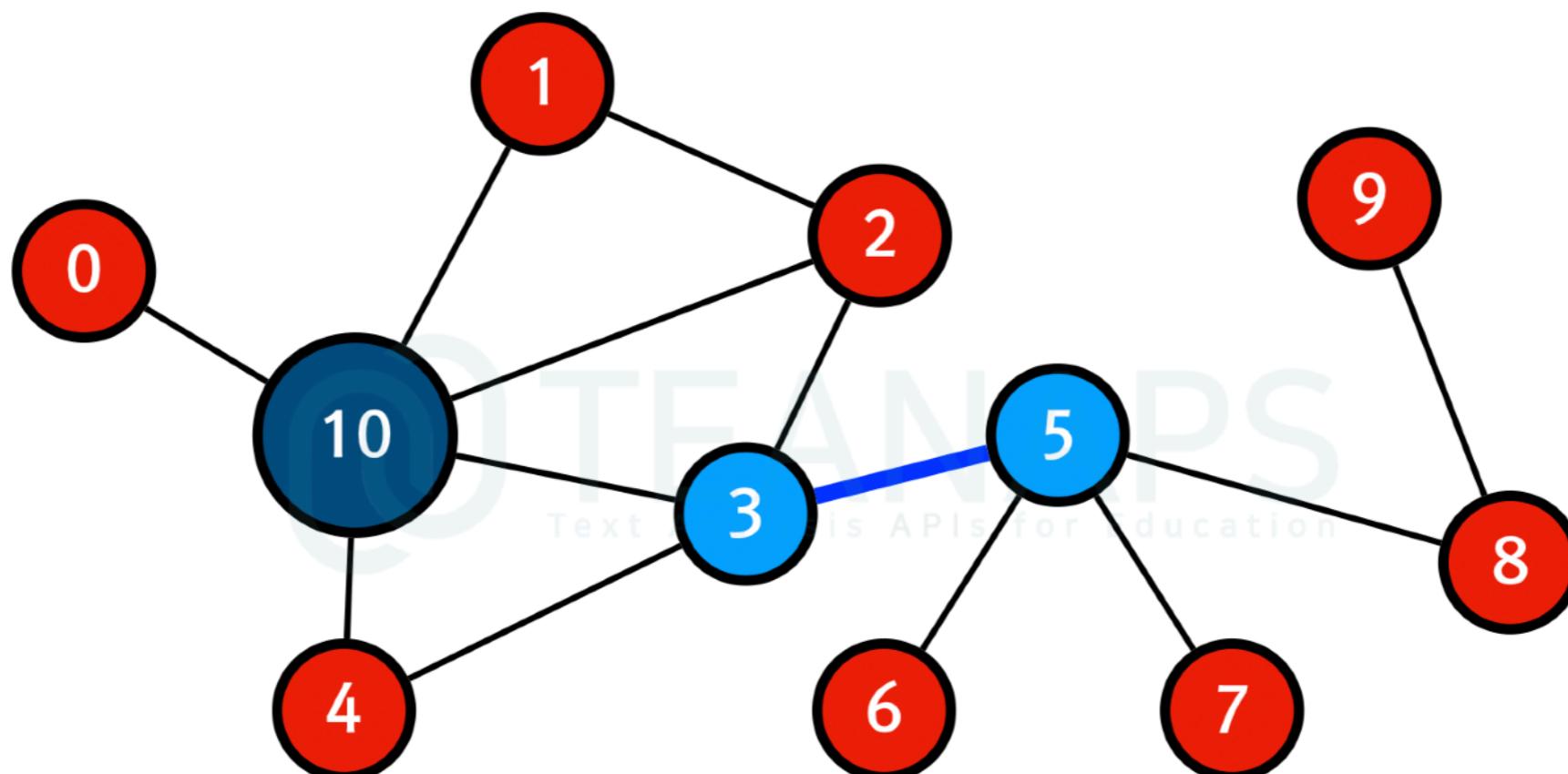
- 어떤 단어가 다른 단어들과의 가장 가까운 거리에 있는가에 대한 척도
- 한 노드에서 다른 모든 노드까지 모든 최단 경로의 평균 또는 이의 역수
- 모든 다른 노드에 도달하는데 까지 평균 소요 시간



단어 가중치: 네트워크 중심성

| 네트워크 중심성 척도의 활용

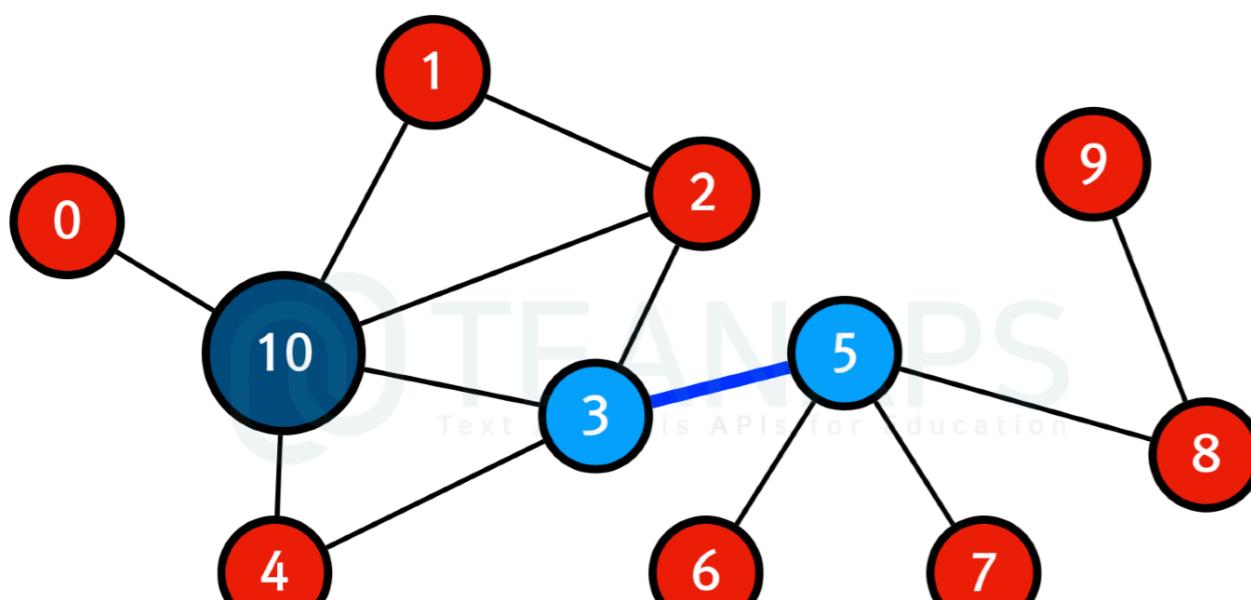
- 분석의 목적에 따라 척도를 다르게 적용하여 분석에 활용 (중심 노드를 예고네트워크 또는 연결된 N개의 노드 단위로 고려해도 됨)
 - 1) 노드 10은 연결 중심성 측면에서 가장 중심에 있음
 - 2) 노드 3과 노드 5는 매개 중심성 측면에서 노드 10 보다 더 중심에 있음
 - 3) 또한 노드 3과 노드 5 사이의 관계는 네트워크가 분리될 수 있는 중요한 연결로 볼 수 있음
 - 4) 다른 조건들이 동일할 때, 3과 5는 10보다 네트워크의 중심에 있음



Sample Graph

단어 가중치: 네트워크 중심성

네트워크 중심성 계산



Sample Graph

노드리스트 (*node list*)

Node	Degree Centrality	Betweenness Centrality	Closeness Centrality
0	3	0	0.8333
1	3	0	0.8333
2	4	0.5	0.6667
3	4	0.5	0.6667
4	4	0	0.6667
5	5	1.5	0.5
6	2	0	0.8333
7	2	0	0.8333
8	2	0	0.8333
9	2	0	0.8333
10	6	0	0.3333

HOW?

어떻게 해결하면 좋을까요?

NAVER TV

한혜진

인생술집 + 구독 방송 다시보기 구독 38,125명

장현성 ♦ 송영규 ♦ 송창이 ♦ 지일주
인생술집 ♦ 뜻밖의 별명 배틀(?) ♦

00:33 | 01:10

배우들의 유치한 과거 별명 **한혜진의 하드캐리ㅋㅋ**

▷ 7,797 | 등록 2019.03.01. 신고

#tvN #목요예능 #술예능 #인생술집 #신동엽 #한혜진 #김희철 #김준현 #별명 #지일주

모두가 한 번쯤은 가져본 유치한 별명들.
한혜진의 과거 별명을 듣고 모두 폭소!
역대급 별명 하드캐리ㅋㅋ

#인생술집 매주 목 밤 11시 tvN

한혜진 모델

출생 1983년 3월 23일
신체 177cm
소속 에스팀엔터테인먼트

한혜진 탤런트, 영화배우

신체 164cm, 46kg
소속 지킴엔터테인먼트

한혜진 가수

출생 1965년 10월 16일
신체 162cm, 48kg
소속 아랑엔터테인먼트

E.O.D

Contact

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