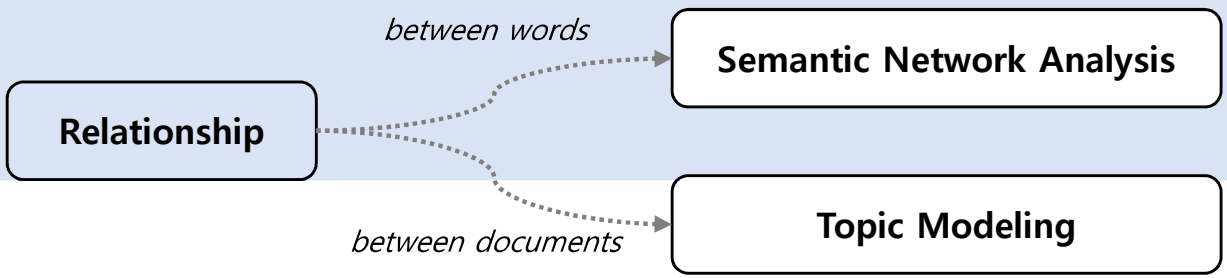
Keywords and Network Analysis

Useful Information?

Text Analysis

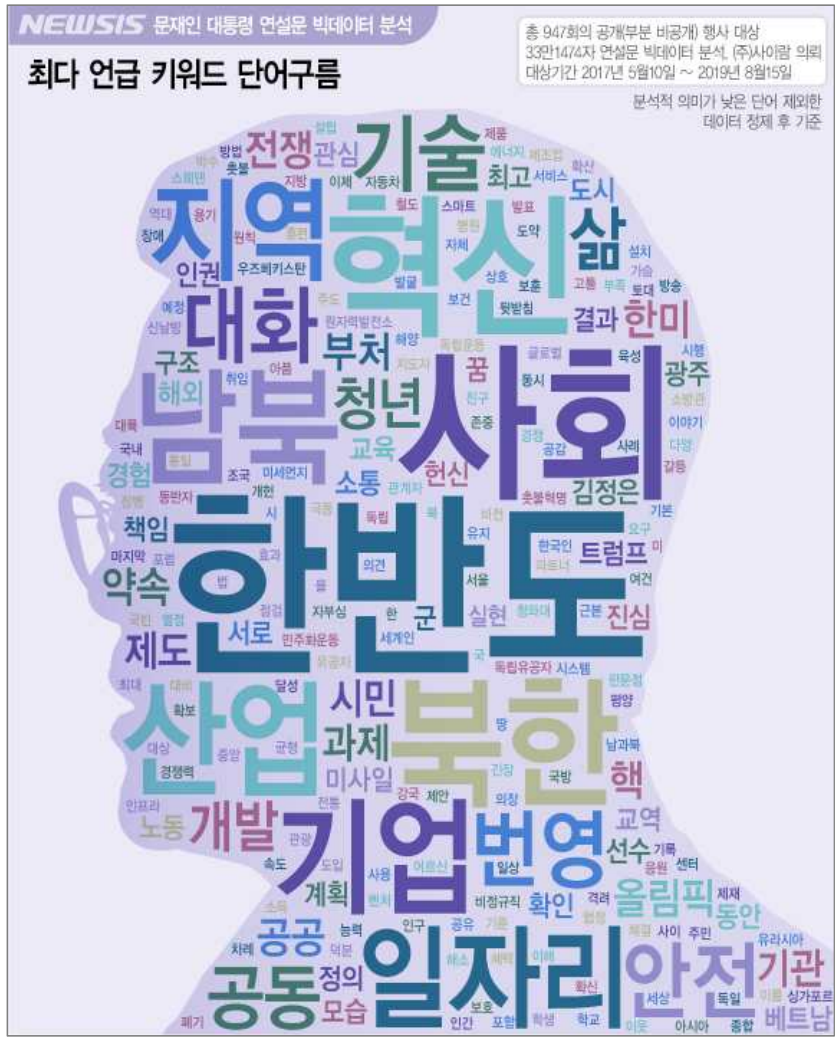


Keyword Analysis

Keyword Extraction

- To extract and visualize the most important words and expressions in a text

  - ***Summarizing***



문재인 대통령 연설문 키워드 분석
→ 경제와 평화 키워드로 수렴

(Source: https://www.newsis.com/view.html?ar_id=NISX20190918_0000772783#_enliple)

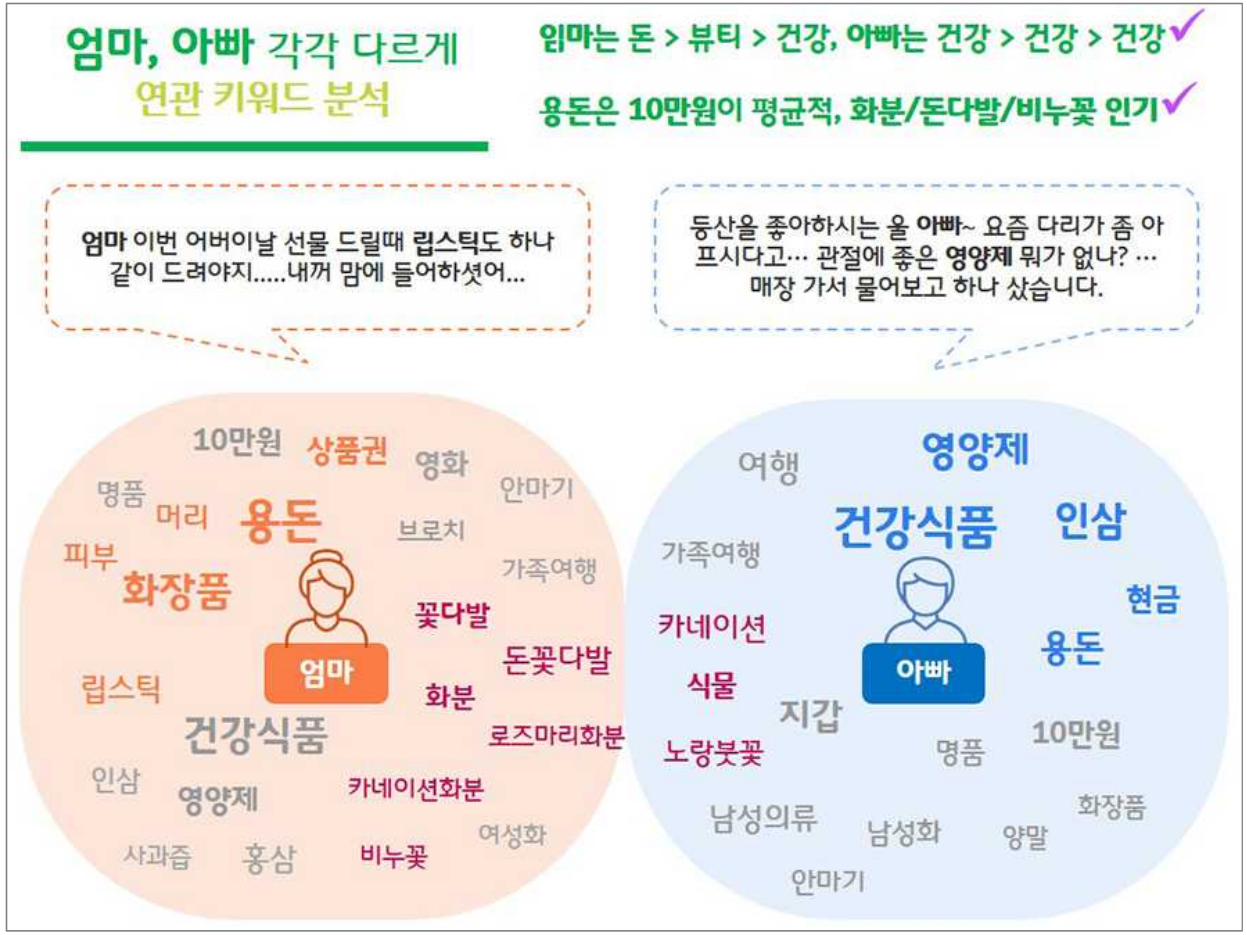
Keyword Analysis

Keyword Extraction

- To extract and visualize the most important words and expressions in a text

  - ***Understanding*** and ***Comparing***

2016년 1월~2018년 4월 15일 SNS(인터넷 뉴스, 블로그, 게시판 등)에서 발생한 57,186건의 데이터 대상



어버이날 관련 SNS 키워드 분석
→ 부모님 성별(엄마, 아빠)에 따른 연관 키워드 비교

(Source: https://www.sktinsight.com/103390)

Keyword Analysis

Basic Process

- Data Collection
- Text Preprocessing
  - To extract words or expressions
- ***Keyword Weighting***
  - To calculate the importance of words or expressions
  - e.g., Term Frequency (TF), Term Frequency – Inverse Document Frequency (TF-IDF)
- ***Visualization***
  - To represent keywords considering the importance of each keyword
  - e.g., Word Cloud, Word Network

Keyword Analysis

Keyword Weighting: Term Frequency (TF)

- Basic frequency: most popular in keyword analysis

  : $TF(t,d) = f(t,d) = count(t)\ in\ d$ (t: $term$, d: $document\ where\ the\ term\ t\ appears$)

- Boolean frequency

  : $TF(t,d) = 1\ if\ t\ occurs\ in\ d\ and\ 0\ otherwise$;

- Logarithmically scaled frequency

  : $TF(t,d) = \log(1 + f(t,d))$

- Augmented frequency, to prevent a bias towards longer documents

  : $TF(t,d) = 0.5 + \frac{0.5 \times f(t,d)}{\max\{f(w,d), w \in d\}}$

Keyword Analysis

Keyword Weighting: Term Frequency – Inverse Document Frequency (TF-IDF)

- To normalize the TF considering the number of documents where a word appears
- $TFIDF(t,d,D) = TF(t,d) \times IDF(t,D) = TF(t,d) \times \log\left(\frac{|D|}{|\{d \in D: t \in d\}|+\alpha}\right)$

  |D|: total number of documents,

  |{d ∈ D: t ∈ d}|: number of documents where the t appears (Document Frequency, DF)

#1. 발판 탈락으로 추락사고 발생
#2. 추락사고로 골절사고 발생
#3. 가시설 탈락으로 손목골절 발생
#4. 가시설 이동 중 낙상 사고 발생

Text Data
(4 documents and 13 words)

Rank	Word	TF
1	사고	4
1	발생	4
3	으로	3
4	탈락	2
4	가시설	2
.....		

TF Rank

Rank	Word	TF-IDF	TF	DF	IDF
1	탈락	0.60	2	2	0.30
1	가시설	0.60	2	2	0.30
.....					
11	사고	0.50	4	3	0.12
12	으로	0.37	3	3	0.12
13	발생	0.00	4	4	0.00

TF-IDF Rank

Visualization: Word Cloud (Tag Cloud)

- The simplest and most common tool for text visualization
  - To depict words arranged in space varied in size, color, and position based on word frequency, categorization, or significance

Size → Importance (e.g., TF, TF-IDF)
Color → Categorization
Position → Significance

Word Cloud from the READMEs of the Top 2,000 GitHub Repositories

(Source: https://itnext.io/basics-of-text-analysis-visualization-1978de48af47)

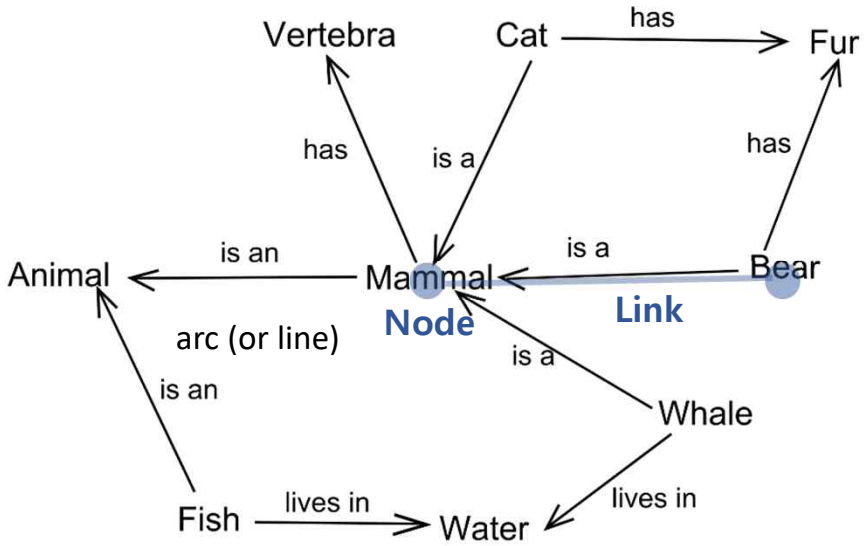
Semantic Network Analysis

Definition

- Automatic exploration and visualization of semantic networks based on unstructured data

  - ***Semantic network***: A knowledge base that represents semantic relations between concepts in network

  - Components: ***node*** (word, person, concept, or event), ***link or line*** (semantic relationships between nodes)



Example of a Semantic Network Graph

(Source: https://en.wikipedia.org/wiki/Semantic_network)

Semantic Network Analysis

Basic Concepts

- ***1) Degree***: the number of connections that a node has
  - ***In-degree***; is the measure of popularity / ***Out-degree***; is the measure of influence
  - Undirected Network / Directed Network / Bi-directed Network
- ***2) Density***: the number of connections a node has, divided by the total possible connections a node could have

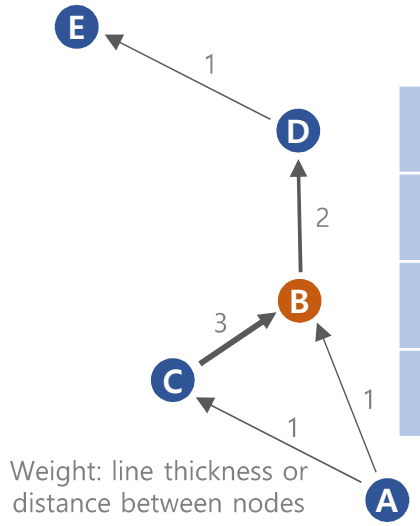
Example 1. Directed Network

Source	Target	Weight
A	B	1
C	B	3
A	C	1
B	D	2
D	E	1

Network Data (List)

	A	B	C	D	E
A	-	1	1	0	0
B	0	-	0	2	0
C	0	3	-	0	0
D	0	0	0	-	1
E	0	0	0	0	-

Matrix



Network Graph

<Node B>

Degree (Weighted)	3 (6)
In-degree (Weighted)	2 (3+1=4)
Out-degree (Weighted)	1 (2)
Density (Weighted)	3/4=75% (6/8=75%)

Basic Concepts

- ***3) Centrality***: an indicator that represents the extent to which an node interacts with other nodes in the network

  - To identify most critical nodes in the network (e.g., critical keywords, influential person)

Type	Definition	The Examples of Application
Degree Centrality	The counts of how many connections a node has	Popularity or influence of a node (e.g., word, person) in the network
Betweenness Centrality	The extent to which a node lies on the shortest paths between other nodes	Central city or infrastructure in urban network
Closeness Centrality	The average length of the shortest path from a node to other nodes	
Eigenvector Centrality	The relative scores based on the centrality of other nodes to which a node has connections	Analysis of Influencer in social network or influential Web Pages (e.g., Google's Page Rank)
Katz Centrality		
PageRank		
……		

(Source: https://digitaluncovered.com/use-social-network-analysis/)

Semantic Network Analysis

Basic Concepts

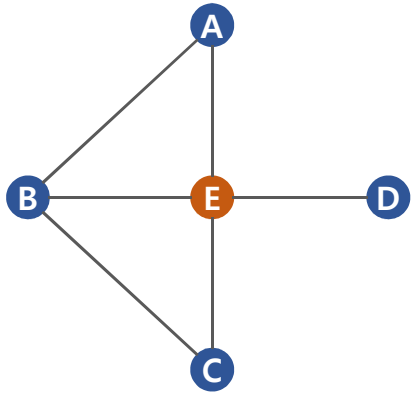
- ***3) Centrality***: an indicator that represents the extent to which an node interacts with other nodes in the network

  - To identify most critical nodes in the network (e.g., critical keywords, influential person)

Example 2. Undirected Network

	A	B	C	D	E
A	-	1	0	0	1
B	1	-	1	0	1
C	0	1	-	0	1
D	0	0	0	-	1
E	1	1	1	1	-

Matrix



Network Graph

The shortest path between two nodes
: A-B[A-B], A-C[A-B-C or A-E-C], A-D[A-E-D], A-E[A-E], B-C[B-C], B-D[B-E-D], B-E[B-E], C-D[C-E-D], D-E[D-E]

<Degree Centrality>

A	2
B	3
C	2
D	1
E	4

<Closeness Centrality>

A	4/(1+2+2+1)=0.67
B	4/(1+1+2+1)=0.8
C	4/(2+1+2+1)=0.67
D	4/(2+2+2+1)=0.57
E	4/(1+1+1+1)=1

<Betweenness Centrality>

A	0
B	0.5(A-C)
C	0
D	0
E	0.5(A-C)+1(A-D)+ 1(B-D)+1(C-D)=3.5

(Source: https://bab2min.tistory.com/554)

Semantic Network Analysis

Application in Text Data: Word Network Analysis

- Basic Approach: ***Co-occurrence Analysis***

  - To identify relationships between keywords based on the co-occurrence of two words in the same document

  - Undirected Network

Example 3. Word Network based on Co-occurrence Analysis

#11. 유성 온천교교량 신축이음이 파손되고 있음
#12. 구서역 인근 금정교 쪽에 안전 난간이 파손되어 있음
#13. … 보행자보호용 안전 난간이 부분 파손되어 보행자에게 위험…
#14. 반포대교 남단에 시선유도봉이 일부 파손되어 있습니다
……

Preprocessed Data
(4 documents and 5 words)

Word	Word	Co-occurrence
신축이음	파손	1
난간	파손	2
난간	보행자	1
시선유도봉	파손	1
신축이음	난간	0
보행자	신축이음	0
신축이음	시선유도봉	0
파손	보행자	1
난간	시선유도봉	0
보행자	시선유도봉	0

Co-occurrence Network (List)

Semantic Network Analysis

Application in Text Data: Word Network Analysis

- Basic Approach: ***Co-occurrence Analysis***

  - To identify relationships between keywords based on the co-occurrence of two words in the same document

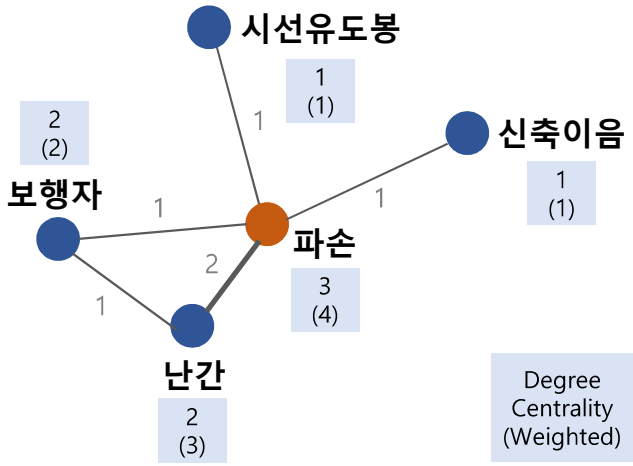
  - Undirected Network

Example 3. Word Network based on Co-occurrence Analysis

(Co-occurrence = Weitght)

	신축이음	파손	난간	보행자	시선유도봉
신축이음	-	1	0	0	0
파손	1	-	2	1	1
난간	0	2	-	1	0
보행자	0	1	1	-	0
시선유도봉	0	1	0	0	-

Co-occurrence Matrix



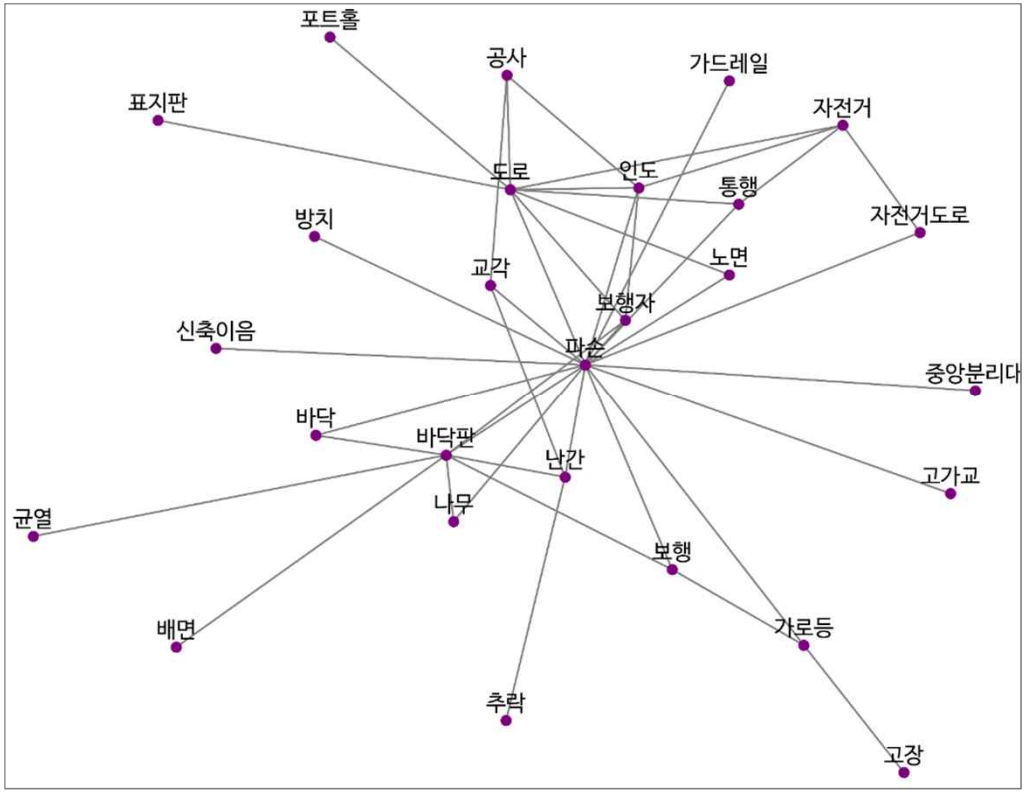
Word Network Graph

Semantic Network Analysis

가까이 있으면 weight가 크다
(물론 굵기로도 코딩하면 가능)
Centrality가 크면 Node의 크기가 크다

Application in Text Data: Word Network Analysis

2017년 1월~2018년 12월 발생한 2,945건의 교량 관련 안전신문고 민원데이터 대상



교량 관련 민원데이터 단어 네트워크
→ '파손' 및 '도로'를 중심으로 연관 단어 파악

2019년 1월 1일~2019년 8월 15일 대통령 연설문, 담화문, 회의록 대상

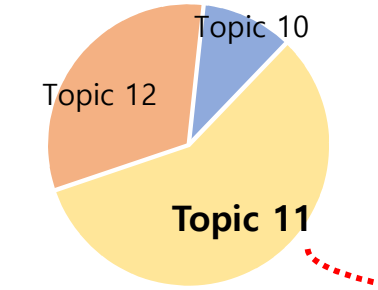
2019년 대통령 연설문 단어 네트워크
→ '혁신', '일자리'와의 연관 단어 파악 및 핵심주제 군집화

(Source: https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&oid=003&aid=0009465748&sid1=001)

Topic Modeling

Definition

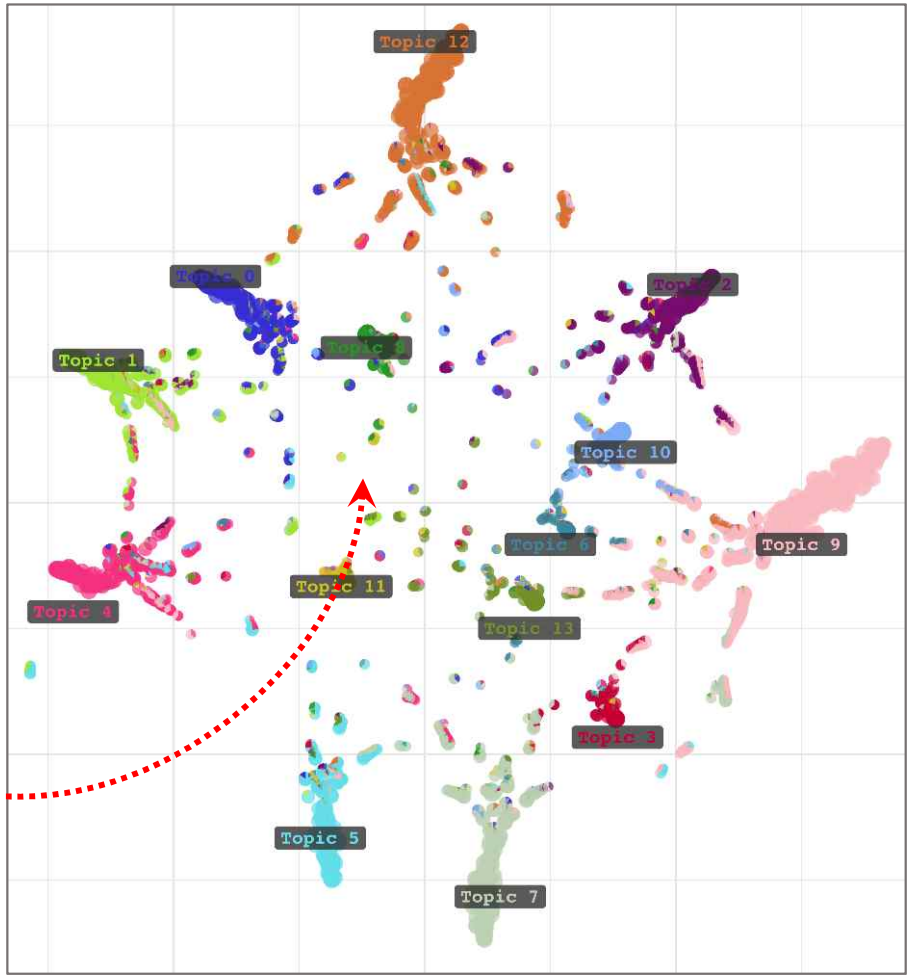
- One of text mining techniques that is to ***arrange a large volume of documents*** based on their topics
- To generate a representation for documents in the topic space by ***providing topics*** presented in each document



Topics of a Document

Topic 11

점 하나가 문서 하나



(Source: https://towardsdatascience.com/visualizing-topic-models-with-scatterpies-and-t-sne-f21f228f7b02)

Topic Modeling

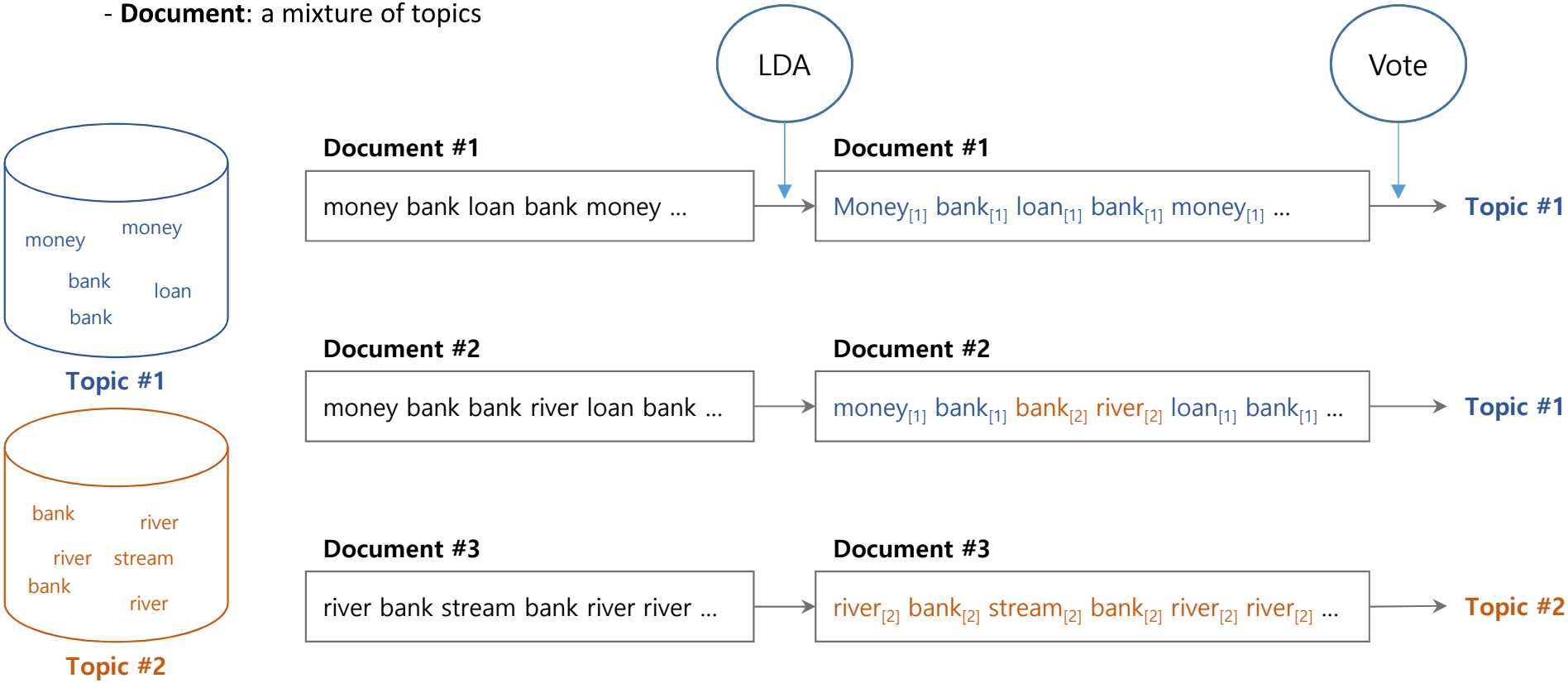
같은 단어가 여러 문서에 등장할 수 있지만 그것이 어떤 토픽을 대표하는지는 확률적으로 계산

Basic Concept

- Assign the ***most probable topic*** for each document based on ***which topic the words came from***

  - **Topic**: a probability distribution of words

  - **Document**: a mixture of topics

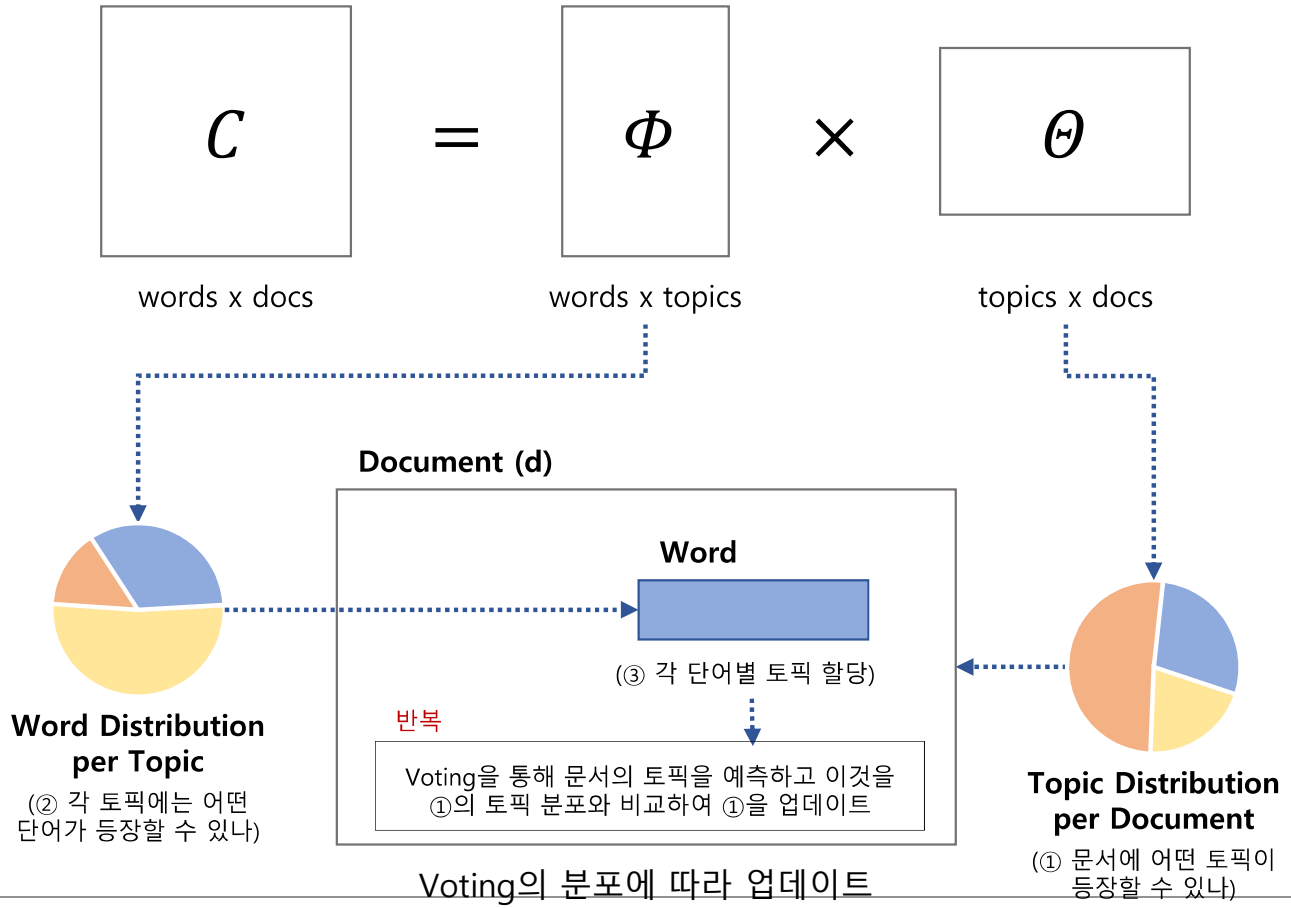


Topic Modeling

토픽별 단어분포와 문서별 토픽분포를 모두 고려

Latent Dirichlet Allocation

- A ***probabilistic model for document generation***, which is most commonly used for topic modeling
- Modeling the process of how each document was generated → Recognizing the latent topic structures



Latent Dirichlet Allocation – Example

Text Data	
Doc #1	The dog sat on the table
Doc #2	The cat sat on the table
Doc #3	The table is black

LDA Modeling

Topic-document Probability

	Topic #1	Topic #2	Total
Doc #1	0.80	0.20	1.00
Doc #2	0.75	0.25	1.00
Doc #3	0.13	0.87	1.00

Word-topic Probability

	Topic #1	Topic #2
the	0.05	0.04
dog	0.30	0.02
cat	0.25	0.01
sat	0.02	0.01
on	0.24	0.01
table	0.11	0.80
is	0.01	0.06
black	0.02	0.05
total	**1.00**	**1.00**

New Doc The black cat sat on the table

Topic #1

The = 0.8 x 0.05 vs 0.2 x 0.04

Topic Modeling

λ가 크면, 해당 토픽에서 자주 등장하는 단어들이 상위에 랭크(람다=1이면 TF, 단 다른 토픽에서도 많이 등장하는 단어들일 수 있음)
λ가 작으면, 빈도는 적어도 해당 토픽에서 중요한 역할을 하는 단어들이 상위에 랭크(다른 토픽 대비 해당 토픽에만 많이 등장하는 단어)

Application: Construction Complaints Analysis