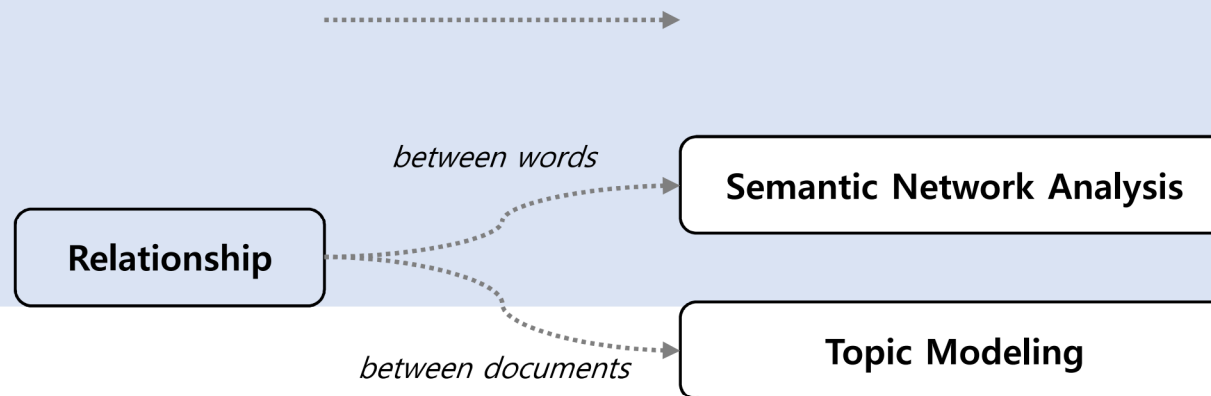


Keywords and Network Analysis

Useful Information?

Text Analysis



I 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

I Keyword Weighting: Term Frequency (TF)

- Basic frequency: most popular in keyword analysis

: $TF(t, d) = f(t, d) = \text{count}(t) \text{ in } d$ (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 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,

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

#1. 발판 **탈락**으로 추락**사고** **발생**
 #2. 추락**사고**로 골절**사고** **발생**
 #3. **가시설** **탈락**으로 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



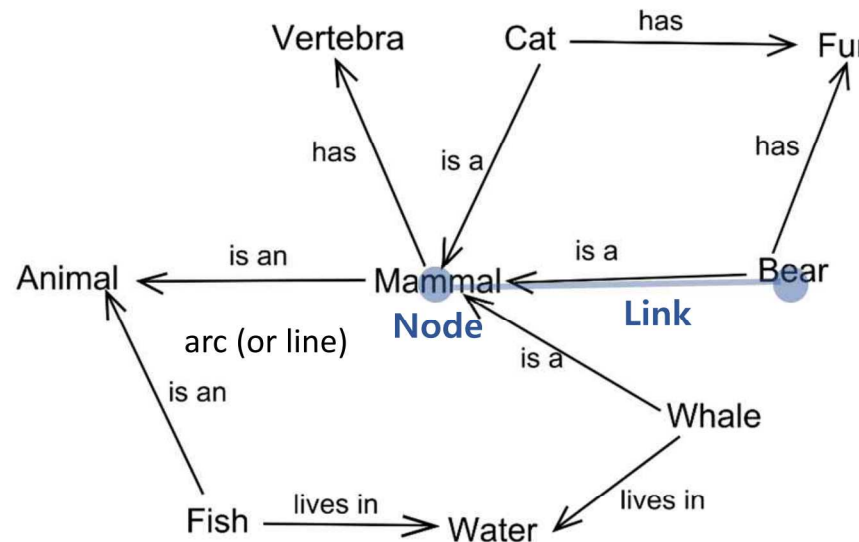
Size → Importance (e.g., TF, TF-IDF)
Color → Categorization
Position → Significance

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

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

I 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)

I 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

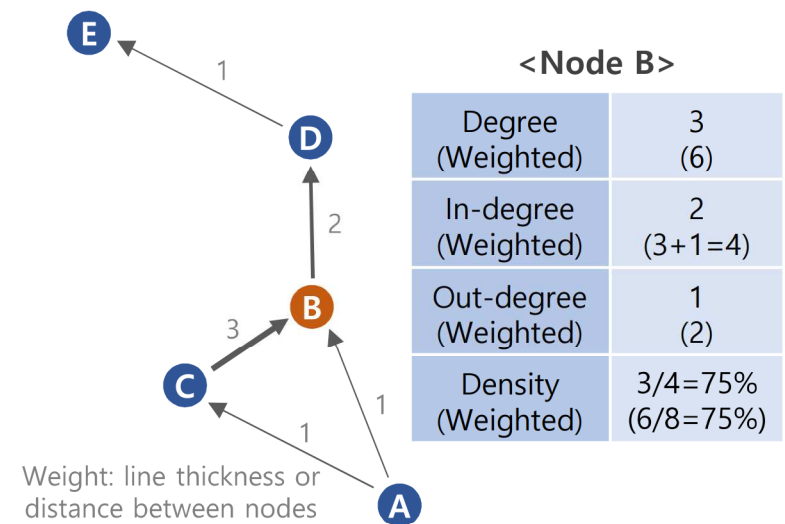
Source	Target	Weight
A	B	1
C	B	3
A	C	1
B	D	2
D	E	1

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%)

I 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 <u>centrality of other nodes</u> 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/>)

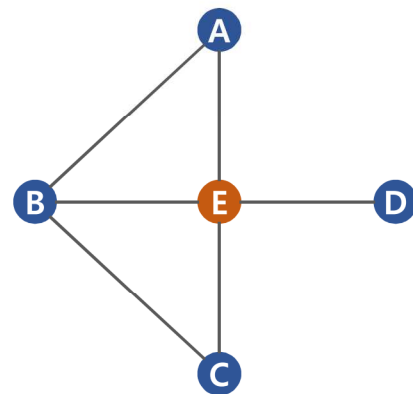
I Basic Concepts

- **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

<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	$0.5(A-C)+1(A-D)+1(B-D)+1(C-D)=3.5$

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]

I 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)

Latent Dirichlet Allocation – Example

