

Data Analytics

Assignment -Network Analysis-

윤장혁 교수님

산업공학과

201811527

이영은

Week4

■ 음식 - 재료 간의 관계

- 이태원에 위치한 꿀밤 포차가 있습니다.

현재 대표메뉴는 (순두부찌개, 제육볶음, 김치찌개, 육전, 두부김치, 숙주 삼겹 볶음) 6가지 입니다.

최근 코로나-19로 인하여 가게에 손님이 급격하게 줄어들었고, 재료가 많이 남게 되어 손해가 심각합니다. 따라서 겹치는 재료들이 많은 대표메뉴 3가지를 찾아서 대표메뉴 3가지만 판매하려고 합니다.

대표메뉴 6가지와 사용되는 재료들은 다음 matrix와 같습니다.

■ 6가지 대표메뉴와 재료

	두부	고기	양파	파	김치	계란	부침가루	숙주
순두부찌개	1	1	1	1	1	1	0	0
제육볶음	0	1	1	1	1	0	0	0
김치찌개	1	1	1	1	1	0	0	0
육전	0	1	0	0	0	1	1	0
두부김치	1	0	1	0	1	0	0	0
숙주 <u>삼겹</u> 볶음	0	1	1	1	0	0	0	1

각각의 메뉴에 필요한 재료이면 1을, 필요하지 않는 재료는 0의 관계를 갖습니다.

각 메뉴들 사이에 공통으로 필요한 재료들이 몇 개가 있는지 알아보기 위하여 Jaccard index를 Similarity measures로 사용하려고 합니다.

■ Jaccard index

두 가지 메뉴에 필요한 각각의 유한집합 사이의 유사도를 통하여 공통적으로 필요한 재료를 구합니다. 이를 통하여 두 가지 메뉴 사이의 similarity를 구하게 됩니다.

구하는 과정은 다음과 같습니다.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cup B|}{|A| + |B| - |A \cap B|}$$

■ Jaccard Index 계산과정

	순두부 찌개	제육 볶음	김치 찌개	육전	두부 김치	숙주 삼겹 볶음
순두부 찌개	1	0.6667	0.8333	0.2857	0.5	0.4286
제육 볶음	0.6667	1	0.8	0.1667	0.4	0.6
김치 찌개	0.8333	0.8	1	0.1429	0.6	0.5
육전	0.2857	0.1667	0.1429	1	0	0.1667
두부 김치	0.5	0.4	0.6	0	1	0.1667
숙주 삼겹 볶음	0.4286	0.6	0.5	0.1667	0.1667	1

편의를 위하여 순두부찌개 = A, 제육볶음 = B, 김치찌개 = C, 육전 = D, 두부김치 E, 숙주 삼겹볶음 = F 로 표기하였습니다.

$$J(A, B) = \frac{4}{6}$$

$$J(A, C) = \frac{5}{6}$$

$$J(A, D) = \frac{2}{7}$$

$$J(A, E) = \frac{3}{6} = \frac{1}{2}$$

$$J(A, F) = \frac{3}{7}$$

$$J(B, C) = \frac{4}{5}$$

$$J(B, D) = \frac{1}{6}$$

$$J(B, E) = \frac{2}{5}$$

$$J(B, F) = \frac{3}{5}$$

$$J(C,D) = \frac{1}{7}$$

$$J(C,E) = \frac{3}{5}$$

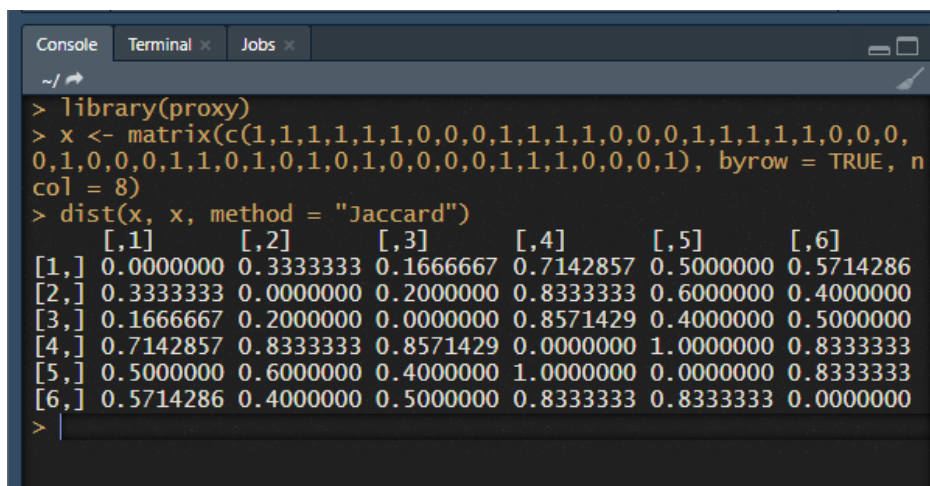
$$J(C,F) = \frac{3}{6}$$

$$J(E,F) = \frac{1}{6}$$

로 각각의 Jaccard index는 위의 과정과 같고, Jaccard distance는 1에서 각각의 index를 뺀 값입니다.

R Studio 에서 실행 한 결과는 다음과 같습니다.

■ R Studio 를 통한 Jaccard distance



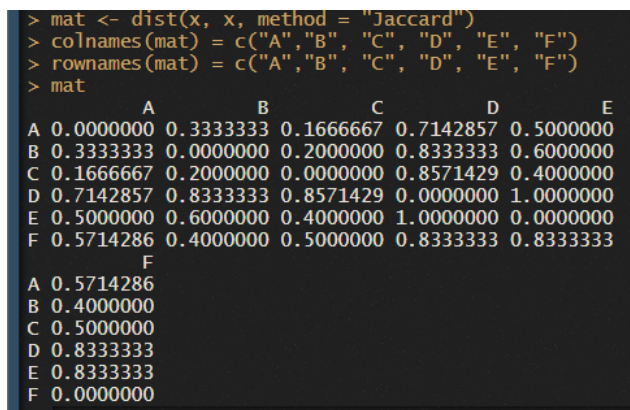
```

> library(proxy)
> x <- matrix(c(1,1,1,1,1,1,0,0,0,1,1,1,1,0,0,0,1,1,1,1,0,0,0,0,
0,1,0,0,0,1,1,0,1,0,1,0,1,0,0,0,0,1,1,1,0,0,0,1), byrow = TRUE, n
col = 8)
> dist(x, x, method = "Jaccard")
      [,1] [,2] [,3] [,4] [,5] [,6]
[1,] 0.0000000 0.3333333 0.1666667 0.7142857 0.5000000 0.5714286
[2,] 0.3333333 0.0000000 0.2000000 0.8333333 0.6000000 0.4000000
[3,] 0.1666667 0.2000000 0.0000000 0.8571429 0.4000000 0.5000000
[4,] 0.7142857 0.8333333 0.8571429 0.0000000 1.0000000 0.8333333
[5,] 0.5000000 0.6000000 0.4000000 1.0000000 0.0000000 0.8333333
[6,] 0.5714286 0.4000000 0.5000000 0.8333333 0.8333333 0.0000000
>

```

행과 열 이름을 A-F로 설정하여 구한 Jaccard distance는 다음과 같습니다.

Jaccard distance가 1에 가까울수록 겹치는 재료가 없음을 알 수 있습니다.

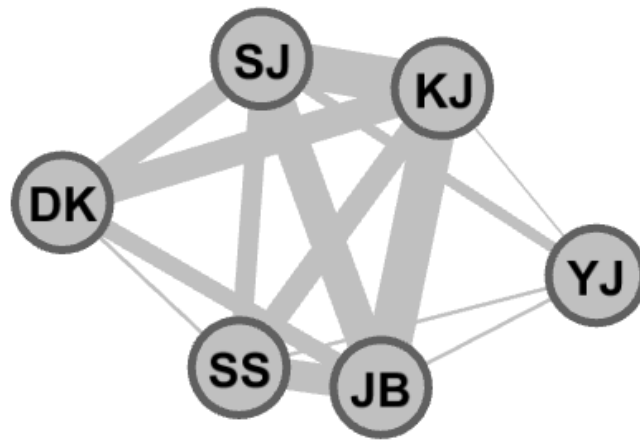


```

> mat <- dist(x, x, method = "Jaccard")
> colnames(mat) = c("A", "B", "C", "D", "E", "F")
> rownames(mat) = c("A", "B", "C", "D", "E", "F")
> mat
      A      B      C      D      E
A 0.0000000 0.3333333 0.1666667 0.7142857 0.5000000
B 0.3333333 0.0000000 0.2000000 0.8333333 0.6000000
C 0.1666667 0.2000000 0.0000000 0.8571429 0.4000000
D 0.7142857 0.8333333 0.8571429 0.0000000 1.0000000
E 0.5000000 0.6000000 0.4000000 1.0000000 0.0000000
F 0.5714286 0.4000000 0.5000000 0.8333333 0.8333333
      F
A 0.5714286
B 0.4000000
C 0.5000000
D 0.8333333
E 0.8333333
F 0.0000000
>

```

■ Jaccard index를 통한 gephi 로 네트워크 시각화



Gephi 프로그램을 통한 네트워크 시각화 결과는 위의 네트워크와 같습니다.

해당 네트워크 시각화를 하기위해 node list와 edge list를 csv 파일로 저장하고, Jaccard index를 통하여 나온 유사성을 weight로 두어 edge의 두께로 표현하였습니다.

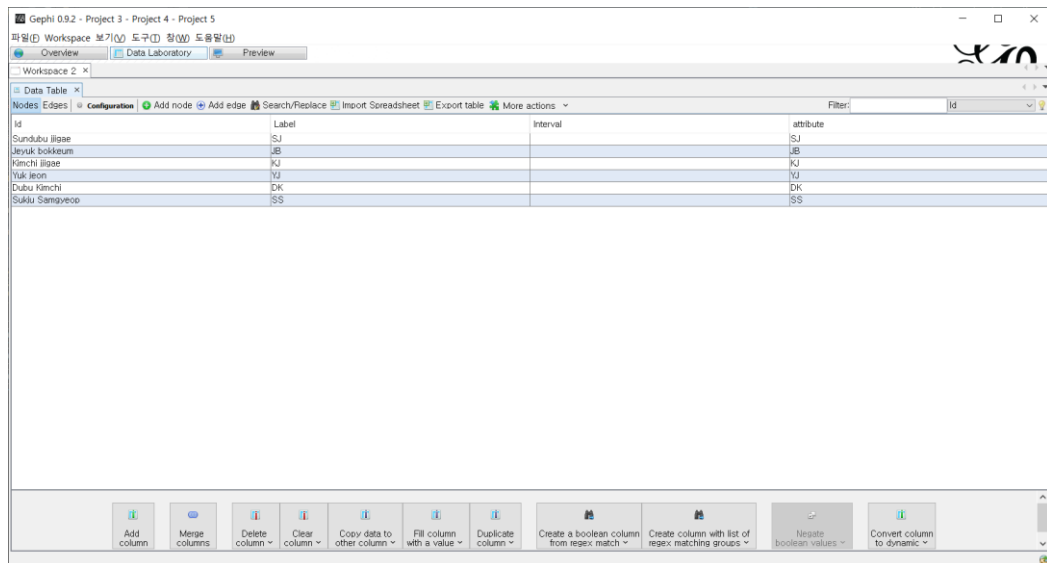
■ Edgelist .csv파일

edgelist.csv [C:\...]				
파일 편집 보기 입력 서식 수식 데이터				
오려 두기 복사하기 붙이기 모양 복사 셀 서식 일반 % , .00 →.0				
A1 X f00 Source				
	A	B	C	D
1	Source	Target	Weight	
2	Sundubu jigae	Jeyuk bokkeum	667	
3	Sundubu jigae	Kimchi jigae	833	
4	Sundubu jigae	Yuk jeon	286	
5	Sundubu jigae	Dubu Kimchi	500	
6	Sundubu jigae	Sukju Samgyeo	429	
7	Jeyuk bokkeum	Kimchi jigae	800	
8	Jeyuk bokkeum	Yuk jeon	167	
9	Jeyuk bokkeum	Dubu Kimchi	400	
10	Jeyuk bokkeum	Sukju Samgyeo	600	
11	Kimchi jigae	Yuk jeon	143	
12	Kimchi jigae	Dubu Kimchi	600	
13	Kimchi jigae	Sukju Samgyeo	500	
14	Yuk jeon	Dubu Kimchi	0	
15	Yuk jeon	Sukju Samgyeo	167	
16	Dubu Kimchi	Sukju Samgyeo	167	
17				

Edge list csv 파일은 다음과 같고 Jaccard 유사도를 통하여 weight를 가중치로 두었습니다

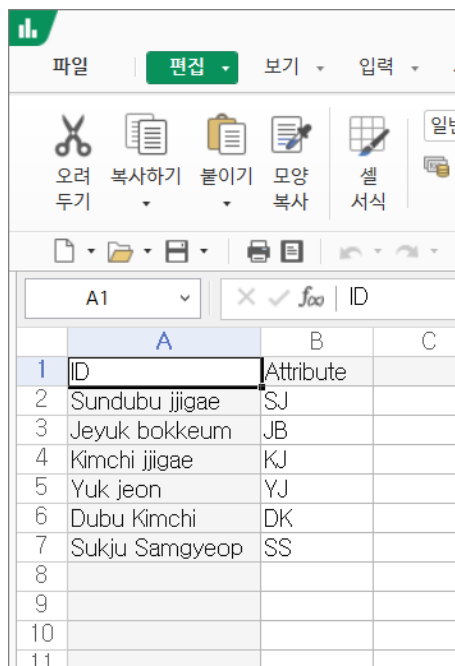
다.

■ Gephi에서 불러온 edge list



gephi에서 edgelist.csv 파일을 불러오고, attribute에 있던 속성을 label로 복사하여 그래프에서 label이 보일 수 있도록 설정하였습니다.

■ nodelist .csv파일



노드 리스트에는 ID에는 노드를 적고 Attribute에는 노드 약자를 적었습니다.

■ Gephi에서 불러온 node list

Gephi 0.9.2 - Project 3 - Project 4 - Project 5

파일 Workspace 보기 도구 창 도움말

Overview Data Laboratory Preview

Workspace 2

Data Table

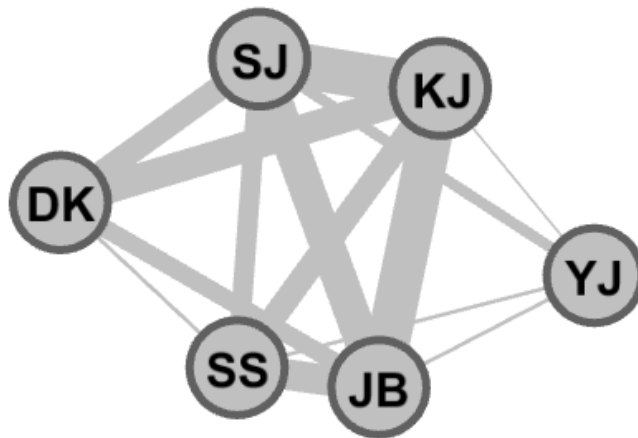
Source	Target	Type	Id	Label	Interval	Weight
Kimchi ilgae	Yuk ieon	Undirected	54			143.0
Ueyuk bokkeum	Yuk ieon	Undirected	51			167.0
Yuk ieon	Sukku Samgyeop	Undirected	58			167.0
Dubu Kimchi	Sukku Samgyeop	Undirected	59			167.0
Sundubu ilgae	Yuk ieon	Undirected	47			286.0
Ueyuk bokkeum	Dubu Kimchi	Undirected	52			400.0
Sundubu ilgae	Sukku Samgyeop	Undirected	43			429.0
Sundubu ilgae	Dubu Kimchi	Undirected	48			500.0
Kimchi ilgae	Sukku Samgyeop	Undirected	56			500.0
Ueyuk bokkeum	Sukku Samgyeop	Undirected	53			600.0
Kimchi ilgae	Dubu Kimchi	Undirected	55			600.0
Sundubu ilgae	Ueyuk bokkeum	Undirected	45			667.0
Ueyuk bokkeum	Kimchi ilgae	Undirected	50			800.0
Sundubu ilgae	Kimchi ilgae	Undirected	46			833.0

Add column Merge columns Delete column Clear column Copy data to other column Fill column with a value Duplicate column

Create a boolean column from regex match Create column with list of regex matching groups

Rescale boolean values Convert column to dynamic

■ Gephi network 시각화 분석



SJ = 순두부찌개, JB = 제육볶음, KJ = 김치찌개, YJ = 육전

DK = 두부김치, SS = 숙주 삼겹 볶음 Node로 나타냈습니다.

메인 메뉴 3개로 줄이기 결과는 다음과 같습니다.

- 결과

- SJ, KJ, JB

- 즉, 순두부찌개, 김치찌개, 제육볶음 3개 메뉴의 재료 유사성이 가장 크다는 것을 알게 되었습니다.
- 따라서 세개의 메뉴를 메인 메뉴로 운영하는 것이 재료 재고를 가장 적게 남기는 것에 도움이 될 것이라고 생각합니다.