（演習3-1）

- 空欄をうめてtf_{ij}の表を作成してください
  - D1. 私は、シャンパンが好きですが、フランスのブランデーも好きです。
  - D2. お酒といったらシャンパンとワインが美味しいです。
  - D3. フランスでシャンパンを飲みました。
  - D4. イギリスといえばウィスキー、フランスといえばブランデーですね。
  - D5. ウィスキーとブランデーなら、ウィスキーが好きです。
  - D6. イギリスとフランスに行ってきました。

$$\mathrm{idf}_j = \log \frac{N}{n_j}$$

i

Document *Di* Term *tj*	D_1	D_2	D_3	D_4	D_5	D_6	t_jを含む文書数n_j	idf_j
イギリス	0	0	0	1	0	1	2	log(6/2)
ウィスキー	0	0	0	1	2	0	2	log(6/2)
シャンパン	**(1)**	**(1)**	**(1)**	**(0)**	**(0)**	**(0)**	**(3)**	**log(6/3)**
ワイン	**(0)**	**(1)**	**(0)**	**(0)**	**(0)**	**(0)**	**(1)**	**log(6/1)**
フランス	**(1)**	**(0)**	**(1)**	**(1)**	**(0)**	**(1)**	**(4)**	**log(6/4)**
ブランデー	1	0	0	1	1	0	3	log(6/3)

j

tf_{ij}

（演習3-2）

- 空欄を埋めてwijの表を作成してください

$$w_{ij} = tf_{ij} \times idf_j$$

i

Document *Di* Term *tj*	D_1	D_2	D_3	D_4	D_5	D_6	idf_j
イギリス	0	0	0	0.477	0	0.477	0.477
ウィスキー	**(0)**	**(0)**	**(0)**	**(0.477)**	**(0.954)**	**(0)**	0.477
シャンパン	**(0.301)**	**(0.301)**	**(0.301)**	**(0)**	**(0)**	**(0)**	0.301
ワイン	**(0)**	**(0.778)**	**(0)**	**(0)**	**(0)**	**(0)**	0.778
フランス	**(0.176)**	**(0)**	**(0.176)**	**(0.176)**	**(0)**	**(0.176)**	0.176
ブランデー	0.301	0	0	0.301	0.301	0	0.301

j

w_{ij}

（演習3-3）

- 文書D4とD5の類似度（Cos(D4,D5)）を計算してください。

$$Cos(x,y) = \frac{\sum x_i y_i}{\sqrt{\sum {x_i}^2}\sqrt{\sum {y_i}^2}} = \frac{x_1y_1 + x_2y_2 \ldots + x_ny_n}{\sqrt{{x_1}^2 + {x_2}^2 \ldots + {x_n}^2}\sqrt{{y_1}^2 + {y_2}^2 \ldots + {y_n}^2}}$$

（例）

Cos(D1, D2) = {Σ($D1_i$*$D2_i$) / (√Σ$D1_i^2$)*(√Σ$D2_i^2$)}
= {(0*0)+(0*0)+(0.301*0.301)+(0*0.778)+(0.176*0)+(0.301*0)}
/ {(0^2+0^2+0.301^2+0^2+0.176^2+0.301^2)*(0^2+0^2+0.301^2+0.778^2+0^2+0^2)}
= **0.2358**

	D_1	D_2	D_3	D_4	D_5	D_6
イギリス	0	0	0	0.477	0	0.477
ウィスキー	0	0	0	0.477	0.954	0.477
シャンパン	0.301	0.301	**(0.301)**	**(0)**	**(0)**	**(0)**
ワイン	0	0.778	**(0)**	**(0)**	**(0)**	**(0)**
フランス	0.176	0	**(0.176)**	**(0.176)**	**(0)**	**(0.176)**
ブランデー	0.301	0	0	0.301	0.301	0
√ΣDx²	0.461	0.834	0.349	0.759	1.000	0.508

i

Cos(D4, D5) = **0.719**

（演習3-4）

- 次の4つの文書について、D2と最も類似度が高い文書はD1,D3,D4のどれですか？

D1. gold silver truck.
D2. Shipment of gold damaged in a fire.
D3. Delivery of silver arrived in a silver truck.
D4. Shipment of gold arrived in a truck.

（ヒント）以下の空欄を埋めてD2と他の文書の類似度を計算しましょう

tj＼Di	*D1*	*D2*	*D3*	*D4*	tjを含む文書数*nj*	*idfj*
a	0	**1**	1	1	3	=log(4/3) = 0.125
arrived	0	**0**	1	1	2	=log(4/2) = 0.301
damaged	0	**1**	0	0	1	=log(4/1) = 0.602
delivery	0	**0**	1	0	1	=log(4/1) = 0.602
fire	0	**1**	0	0	1	=log(4/1) = 0.602
gold	1	**1**	0	1	3	=log(4/3) = 0.125
in	0	**1**	1	1	3	=log(4/3) = 0.125
of	0	**1**	1	1	3	=log(4/3) = 0.125
silver	1	**0**	2	0	2	=log(4/2) = 0.301
shipment	0	**1**	0	1	2	=log(4/2) = 0.301
truck	1	**0**	1	1	3	=log(4/3) = 0.125

tf・idf

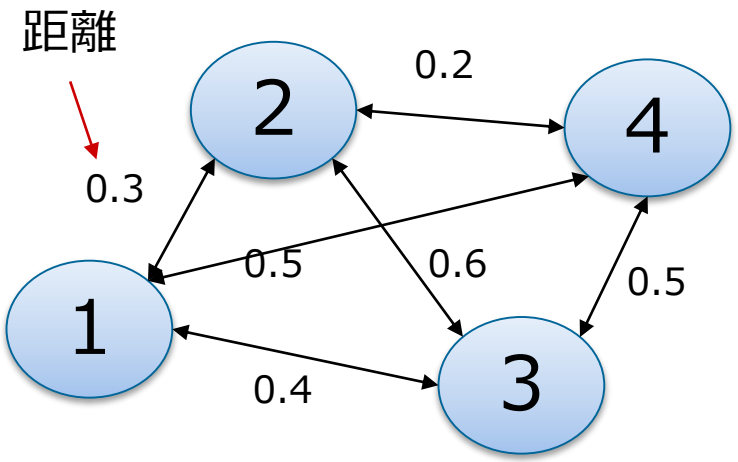
tj＼Di	*D1*	*D2*	*D3*	*D4*
a	0	**0.125**	0.125	0.125
arrived	0	**0**	0.301	0.301
damaged	0	**0.602**	0	0
delivery	0	**0**	0.602	0
fire	0	**0.602**	0	0
gold	0.125	**0.125**	0	0.125
in	0	**0.125**	0.125	0.125
of	0	**0.125**	**0.125**	**0.125**
silver	0.301	**0**	0.602	0
shipment	0	**0.301**	0	0.301
truck	0.125	**0**	0.125	0.125
$\sqrt{\Sigma Dx^2}$	0.349	0.937	0.937	0.509

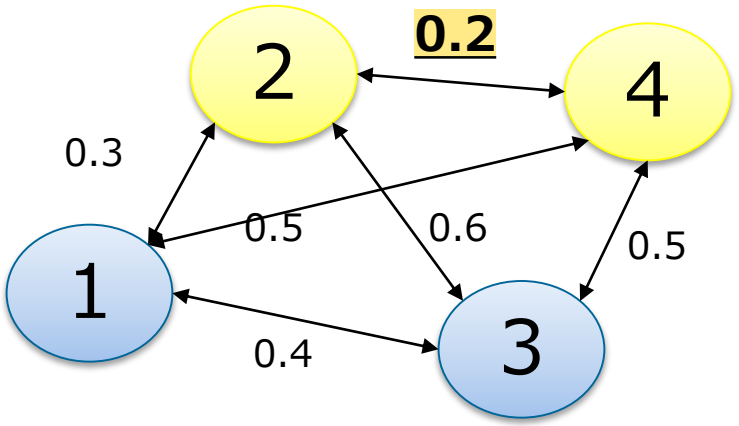
類似度 Cos(D2, Dy)

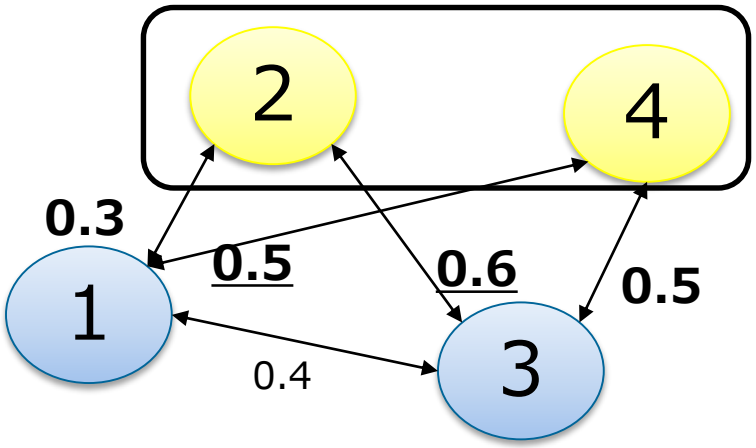
	D2
D1	0.048
D2	1
D3	**0.053**
D4	**0.321**

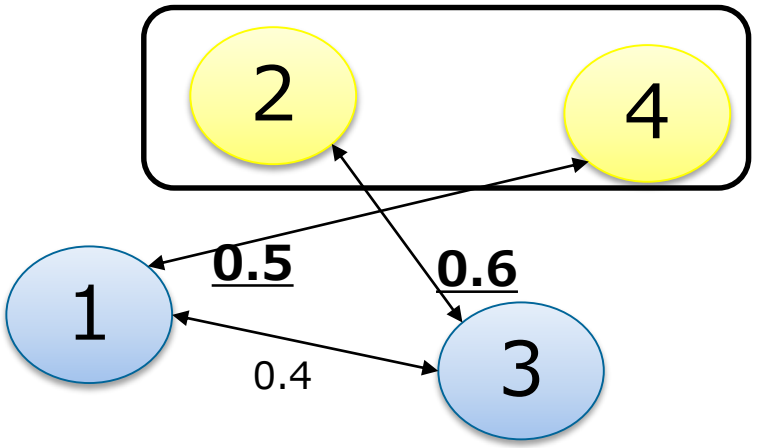
D2と類似度が最も高い文書は **D4**

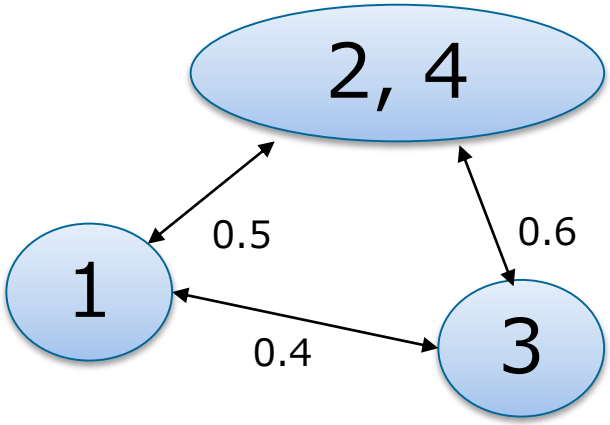
（演習3-5）最長距離法

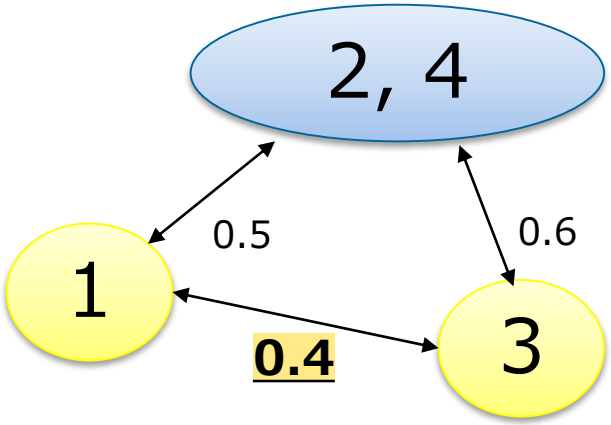

距離
2
4
1
3
0.2
0.3
0.5
0.6
0.5
0.4

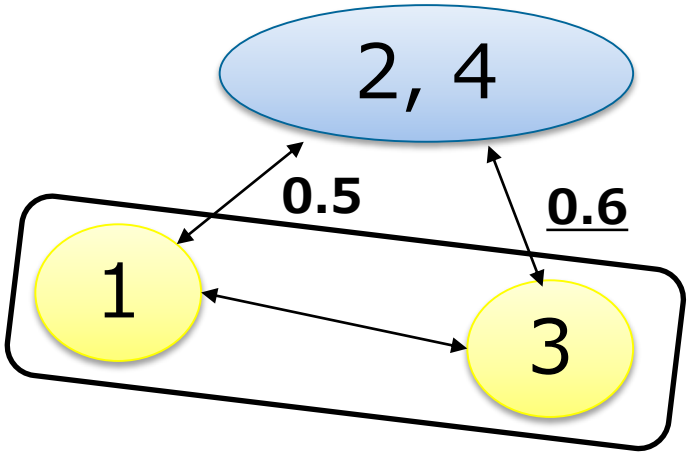

2
4
1
3
0.2
0.3
0.5
0.6
0.5
0.4

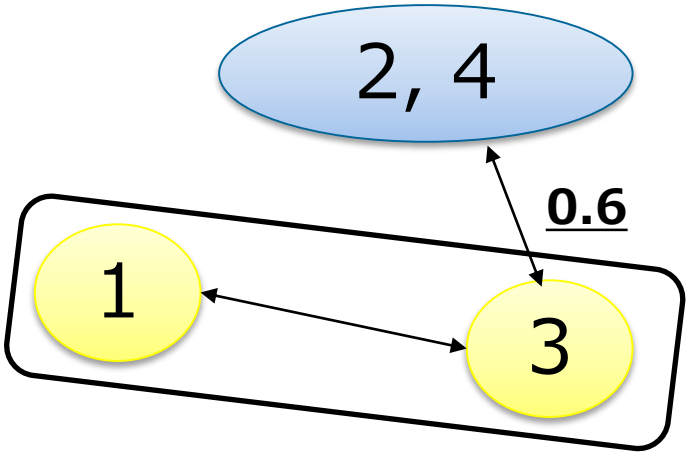

2
4
1
3
0.3
0.5
0.6
0.5
0.4

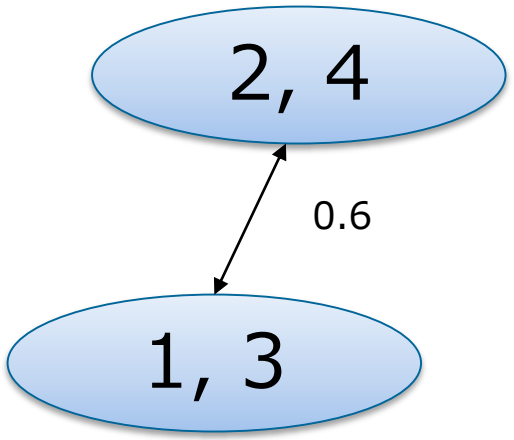

2
4
1
3
0.5
0.6
0.4

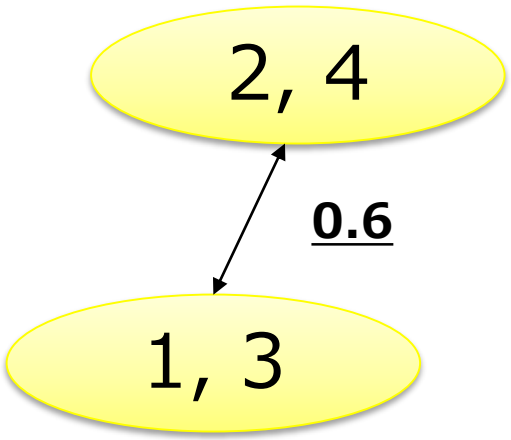

2, 4
0.5
0.6
1
0.4
3


2, 4
0.5
0.6
1
0.4
3


2, 4
0.5
0.6
1
3

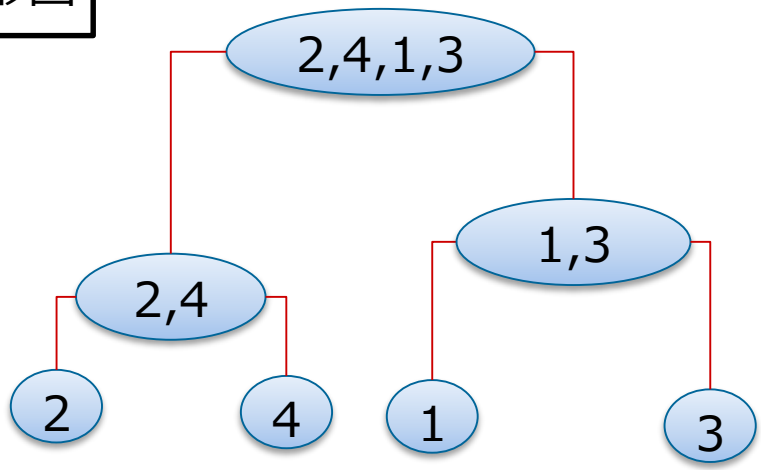

2, 4
0.6
1
3





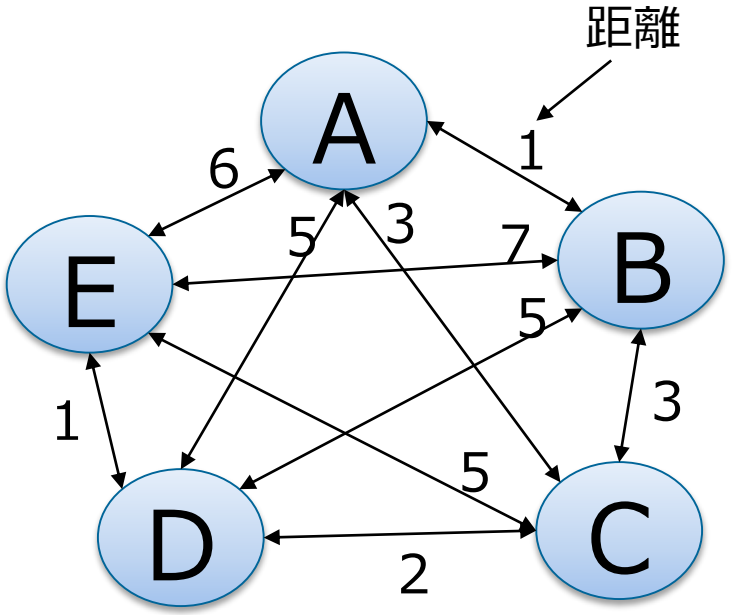
樹形図

2, 4, 1, 3

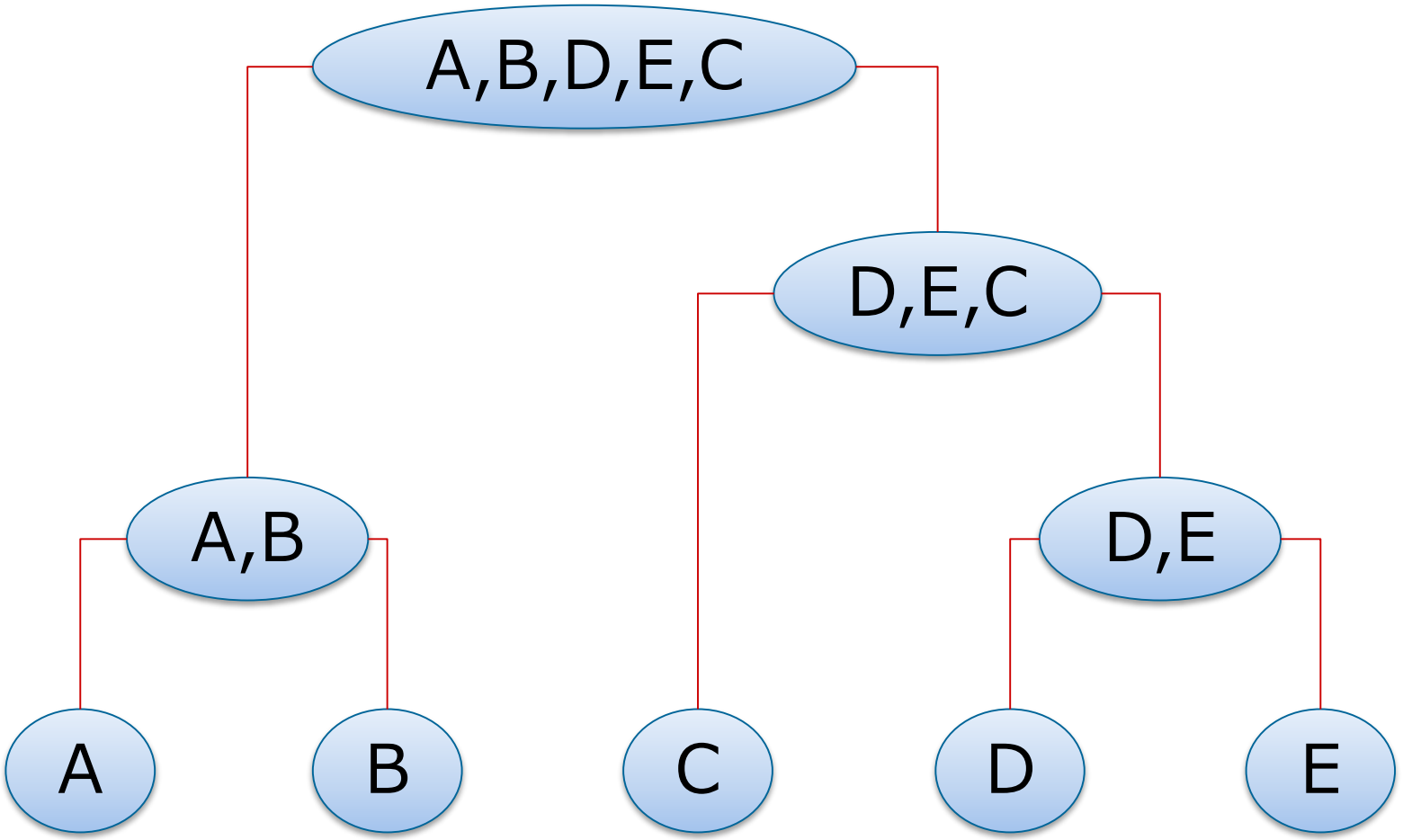


（演習3-6）クラスタリング

- 以下のデータがあったとき、~~どちらか~~両方の手法でクラスタリングを行って樹形図を作成してください
  - **「最短距離法」and 「最長距離法」**



樹形図（最短距離法）



樹形図（最長距離法）