

[자료2] 경제·인문사회연구소소속10개정부출연연구기관의문서의분류

- 정민지(2017). A Study on Clustering Methods for Proximity Data in Text Mining.
- 정호영(2018). Comparison of Term Weighting Techniques for Text Classification.
- 이수진(2021). Document classification using a deep neural network in text mining

[텍스트 군집/분류 제안 프로세스]

[1단계]	비정형 자료 정제화 문장부호, 특수문자, 불용어 제거	117,820 terms from 343 pdf files ↓ 26,915 terms
[2단계]	문서-용어 빈도행렬 생성(document-term frequency matrix) 텍스트 자료 정형화	
[3단계]	문서-용어 가중행렬 생성(document-term weighted matrix) 용어 중요도의 수치화를 위한 가중치 부여	TF-IDF Algorithm
[4단계]	문서-핵심어 가중행렬 생성(document-keyword weighted matrix) 불필요한 용어 필터링	1,158 keywords
[5단계]	군집화 MDS와 k-means CA	분류 SVM , ANN, DNN

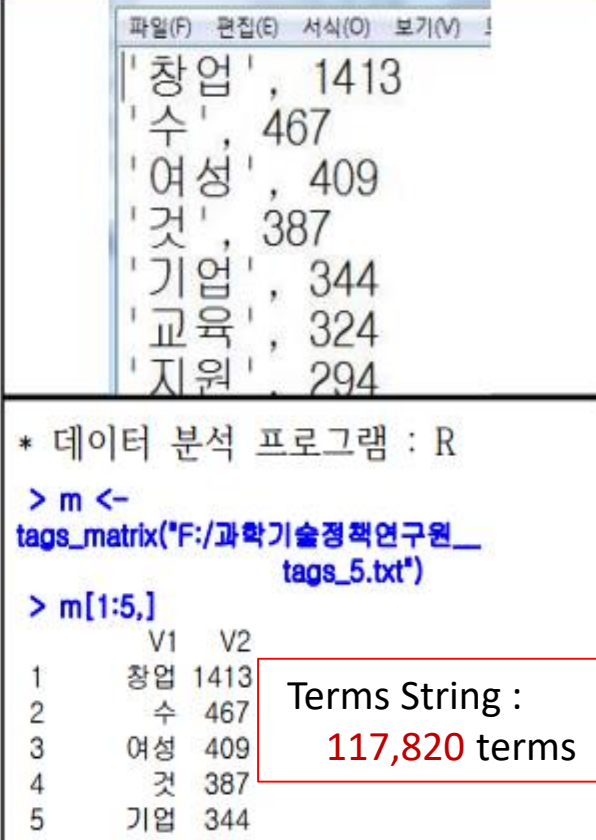
[1단계] 비정형 자료의 정제화

: Korean Natural Language Processing

Crawling

Scraping

Morphological Analysis

크롤링	스크래핑	형태소 분석
암호화되지 않은 PDF 파일	PDF 파일의 문자열 추출	명사 용어 추출
	Letters String 	

R: X-pdf => 파이썬(Python): KoNLPy(한국어자연어처리),

꼬꼬마(Kkma) 형태소 분석기

Terms String :
117,820 terms

[2단계] 문서-용어 빈도행렬 생성(document-term frequency matrix)

Terms : 343 X 117,820

Filtering and Cleaning

343 X 26,915

2016년도 경제인문사회 연구소 출연 연구기관 문서-용어 빈도행렬

개체 : 경제·인문사회 연구소 소속 10개 정부출연연구기관

문서 : 기관에서 2016년 한 해 동안 발간한 정기간행물 343개 PDF 파일

		terms										
개 체	ID	용어	가가	가계	가격	가격대	가공	가능	...	휴양객	흡냄새	
		문서										
A	과학기술정책연구원	1	A01	0	0	0	0	4	3	...	0	0
B	대외경제정책연구원	2	A02	0	0	0	0	0	2	...	0	0
C	산업연구원	3	A03	0	0	3	0	0	5	...	0	0
	육아정책연구소	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮
	정보통신정책연구원	23	A23	0	0	1	0	2	8	...	0	0
⋮	통일연구원	24	B01	0	0	0	0	0	3	...	0	0
	한국개발연구원	25	B02	0	0	3	0	0	3	...	0	0
	한국교육개발원	26	B03	0	0	2	0	0	12	...	0	0
	한국교육과정평가원	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮
J	한국교통연구원	155	B132	0	0	0	0	0	2	...	0	0
		156	C01	0	0	0	0	0	1	...	0	0
		157	C02	0	0	0	0	0	2	...	0	0
		158	C03	1	0	0	0	0	0	...	0	0
		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮

[3단계] 문서-용어 가중행렬 생성(document-term weighted matrix)

343 X 26,915

TF-IDF(Term Frequency-Inverse Document Frequency)

- 대표적 용어 가중치 기법 알고리즘 $N = \text{Total \# of documents}$
- 특정 문서 내에서 용어 빈도가 높을수록, 전체 문서들에는 그 용어를 포함한 문서가 적을수록 TF-IDF 값이 높아짐

$$w_{x,y} = tf_{x,y} \times \log \left(\frac{N}{df_x} \right)$$

- $tf_{x,y}$: y 번째 문서에서 x 번째 용어에 대한 빈도 = Freq. of x th term in y th document
- df_x : x 번째 용어가 포함된 문서의 수, N : 총 문서의 수 = Total # of documents including x th term

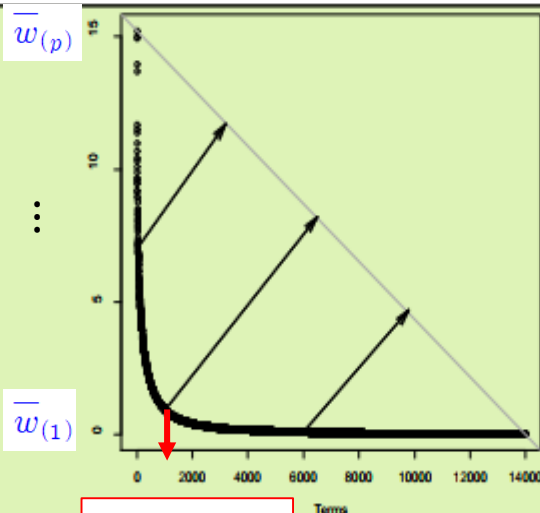
ID	문서	용어	가가	가게	가격	가격대	가공	가능	...	휴양객	흙냄새
1	A01		0	0	0	0	8.041	0.573	...	0	0
2	A02		0	0	0	0	0	0.383	...	0	0
3	A03		0	0	3.951	0	0	0.956	...	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮
23	A23		0	0	1.317	0	4.021	1.530	...	0	0
24	B01		0	0	0	0	0	0.573	...	0	0
25	B02		0	0	3.951	0	0	0.573	...	0	0
26	B03		0	0	2.634	0	0	2.295	...	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮
155	B132		0	0	0	0	0	0.383	...	0	0
156	C01		0	0	0	0	0	0.191	...	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮

```
makeTFIDF <- function(data){
  n=nrow(data)
  p=ncol(data)
  TFIDF <- matrix(NA, n, p)
  rownames(TFIDF) <- rownames(data)
  colnames(TFIDF) <- colnames(data)
  for(i in 1:n){
    for(j in 1:p){
      TFIDF[i,j] <- data[i,j]*(log(nrow(data)/sum(data[,j]!=0)))
    }
  }
  TFIDF[is.nan(TFIDF)] <- 0
  return(TFIDF)
}
```

$$\overline{w_1} \quad \overline{w_2} \quad \dots \quad \overline{w_p} \quad \longrightarrow \quad \overline{w_{(1)}} \geq \overline{w_{(2)}} \geq \dots \geq \overline{w_{(p)}}$$

[4단계] 문서-핵심어 가중행렬 생성(document-keyword weighted matrix)

343 X 1,158



$q = 1,158$

[용어 필터링 방법]

[1단계] 각 p 개 용어들의 평균 가중점수 계산 $\bar{w} = (\bar{w}_1, \bar{w}_2, \dots, \bar{w}_p)^T = n^{-1} W^T \mathbf{1}_n$

[2단계] $\bar{w}$ 의 원소에서 팔꿈치 지점(elbow point) 찾기(Satopaa 등, 2011)

[3단계] 평균 가중점수가 높은 상위 용어 q 개를 핵심어로 선정

[4단계] 문서-핵심어 가중행렬 $\mathbf{X} = (x_{rit})$, $t = 1, \dots, q$ 생성

ID	문서 \ 용어	가격	가스	가족	가치	감사	감소	...	유기성	사측
1	A01	0	0	0	5.692	0	0.336	...	0	0
2	A02	0	0	0	1.897	0	2.018	...	0	0
3	A03	3.951	0	0	10.120	0	5.047	...	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮
23	A23	1.317	3.548	0	1.265	0	0.672	...	0	0
24	B01	0	0	5.195	0.632	0	0.672	...	0	0
25	B02	3.951	0	0	2.530	2.467	1.345	...	0	0
26	B03	2.634	0	0	13.283	0	2.672	...	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮
155	B132	0	0	0	1.265	0	3.364	...	0	0
156	C01	0	0	4.546	1.265	0	2.018	...	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮

$= \mathbf{X}$

[5단계] 개체 군집화 : Hierarchical CA(Ward Linkage)/Metric MDS (최종 용어 117,820 -> 1,158)

> m3Data_CD

Grouped Distance Matrix

	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	B01	B02	B03	B04	B05	B06
A01	0.000000	3.206905	3.446756	3.135831	3.112932	3.425774	3.164416	3.239197	3.120057	3.624077	3.458488	3.243560	5.951523	6.060868	4.809896	6.102394	5.427850	6.851673
A02	3.206905	0.000000	3.447395	3.136805	3.249183	3.372051	3.121716	3.138945	3.078453	3.590831	3.373510	3.224884	5.888912	5.973620	4.765283	6.103320	5.337483	6.658822
A03	3.446756	3.447395	0.000000	3.235597	3.445325	3.479209	3.208754	3.371477	3.282020	3.629488	3.534736	3.306393	6.007731	6.026458	4.875460	6.174279	5.452535	6.920198
A04	3.135831	3.136805	3.235597	0.000000	3.150504	3.212510	2.943465	3.014069	2.964491	3.429519	3.248745	3.072847	5.837448	5.951662	4.716090	6.067601	5.327308	6.807162
A05	3.112932	3.249183	3.445325	3.150504	0.000000	3.330516	3.093575	3.207166	3.151760	3.604940	3.474772	3.318196	5.925739	6.015682	4.795459	6.002871	5.379061	6.837892
A06	3.425774	3.372051	3.479209	3.212510	3.330516	0.000000	3.148407	3.304958	3.220470	3.636024	3.513455	3.370602	5.984741	6.015938	4.869386	6.094700	5.445845	6.894796
A07	3.164416	3.121716	3.208754	2.943465	3.093575	3.148407	0.000000	3.041731	2.938896	3.408736	3.221334	3.112927	5.833370	5.936283	4.673042	6.040274	5.309467	6.765517
A08	3.239197	3.138945	3.371477	3.014069	3.207166	3.304958	3.041731	0.000000	3.029635	3.520874	3.372225	3.181459	5.840920	5.963889	4.647631	6.089577	5.389903	6.832247
A09	3.120057	3.078453	3.282020	2.964491	3.151760	3.220470	2.938896	3.029635	0.000000	3.390363	3.224326	3.002332	5.826901	5.899427	4.664486	6.042246	5.285818	6.761967
A10	3.624077	3.590831	3.629488	3.429519	3.604940	3.636024	3.408736	3.520874	3.390363	0.000000	3.671097	3.573772	6.115112	6.240692	4.957642	6.304374	5.650009	7.017427
A11	3.458488	3.373510	3.534736	3.248745	3.474772	3.513455	3.221334	3.372225	3.224326	3.671097	0.000000	3.407937	6.020923	6.114658	4.902031	6.188610	5.500348	6.898208
A12	3.243560	3.224884	3.306393	3.072847	3.318196	3.370602	3.112927	3.181459	3.002332	3.573772	3.407937	0.000000	5.959055	6.041648	4.763086	6.119602	5.441414	6.852891
B01	5.951523	5.888912	6.007731	5.837448	5.925739	5.984741	5.833370	5.840920	5.826901	6.115112	6.020923	5.959055	0.000000	7.623281	6.747794	7.731151	7.163997	8.347855
B02	6.060868	5.973620	6.026458	5.951662	6.015682	6.015938	5.936283	5.963889	5.899427	6.240692	6.114658	6.041648	7.623281	0.000000	6.777784	7.627627	6.545571	8.325572
B03	4.809896	4.765283	4.875460	4.716090	4.785269	4.869386	4.673042	4.647631	4.664486	4.957642	4.902031	4.763086	6.747794	6.777784	0.000000	6.711388	6.295994	7.588896
B04	6.102394	6.103320	6.174279	6.067601	6.002871	6.094700	6.040274	6.089577	6.042246	6.304374	6.188610	6.119602	7.731151	7.627627	6.711388	0.000000	7.068456	8.491877
B05	5.427850	5.337483	5.452535	5.327308	5.379061	5.445845	5.309467	5.389903	5.285818	5.650009	5.500348	5.441414	7.163997	6.545571	6.295994	7.068456	0.000000	7.975529
B06	6.851673	6.658822	6.920198	6.807162	6.837892	6.894796	6.765517	6.832247	6.761967	7.017427	6.898208	6.852891	8.347855	8.325572	7.588896	8.491877	7.975529	0.000000

간행물 간의 유사성 : 유클리드거리, 가중유클리드거리

10 X 10

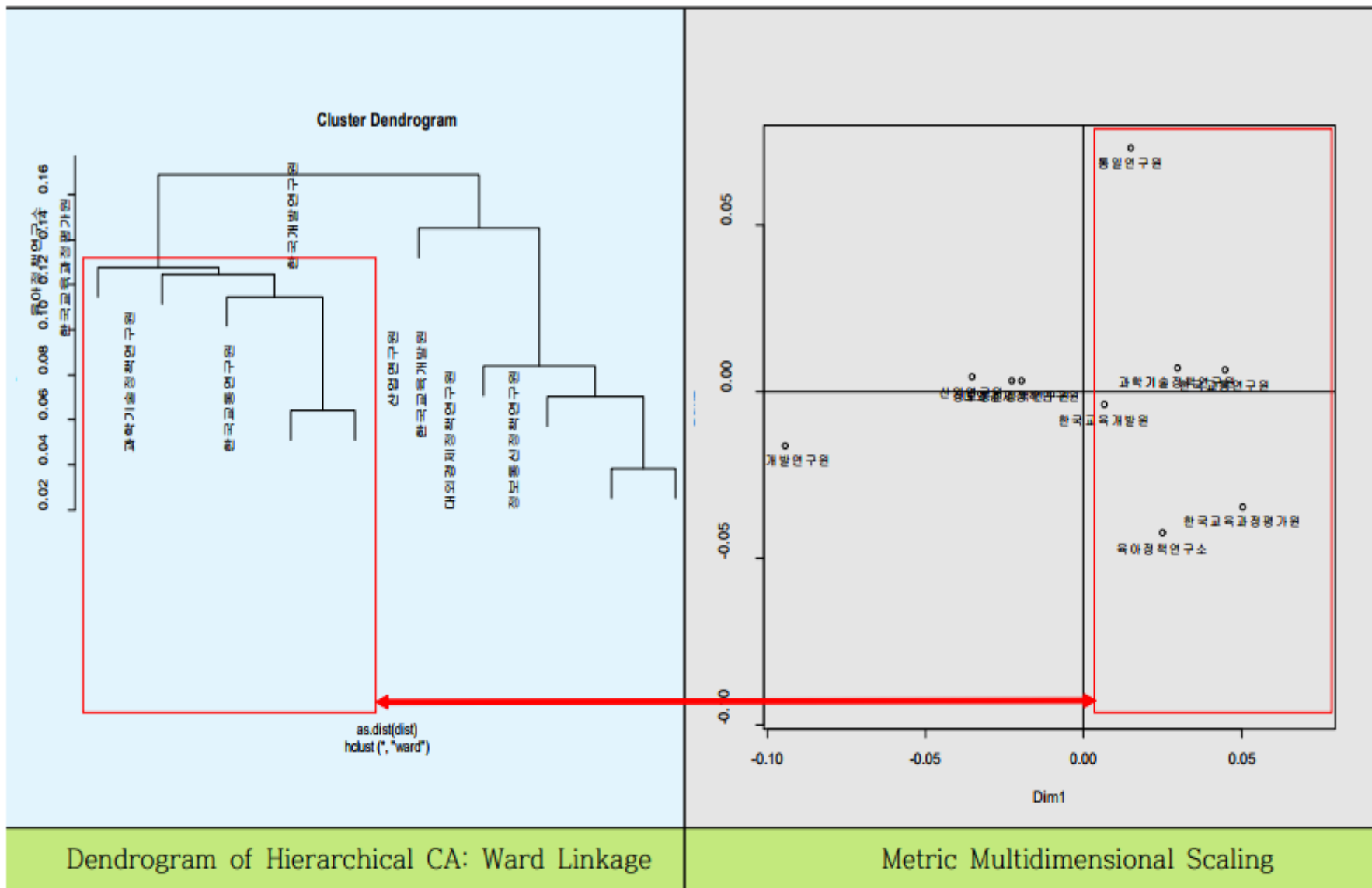
= D Between Group Distance Matrix

343 X 343

> round(dist,3)

	과학기술정책연구원	대외경제정책연구원	산업연구원	육아정책연구소	정보통신정책연구원	통일연구원	한국개발연구원	한국교육개발원	한국교육과정평가원	한국교통연구원
과학기술정책연구원	0.000	0.077	0.100	0.102	0.084	0.099	0.140	0.066	0.099	0.064
대외경제정책연구원	0.077	0.000	0.063	0.101	0.038	0.097	0.107	0.061	0.107	0.097
산업연구원	0.100	0.063	0.000	0.122	0.068	0.121	0.121	0.090	0.132	0.118
육아정책연구소	0.102	0.101	0.122	0.000	0.105	0.130	0.152	0.086	0.125	0.111
정보통신정책연구원	0.084	0.038	0.068	0.105	0.000	0.103	0.111	0.064	0.114	0.103
통일연구원	0.099	0.097	0.121	0.130	0.103	0.000	0.150	0.095	0.132	0.110
한국개발연구원	0.140	0.107	0.121	0.152	0.111	0.150	0.000	0.120	0.161	0.152
한국교육개발원	0.066	0.061	0.090	0.086	0.064	0.095	0.120	0.000	0.100	0.081
한국교육과정평가원	0.099	0.107	0.132	0.125	0.114	0.132	0.161	0.100	0.000	0.105
한국교통연구원	0.064	0.097	0.118	0.111	0.103	0.110	0.152	0.081	0.105	0.000

기관별 유사성 거리 : 기관 내와 기관 간의 거리 평균

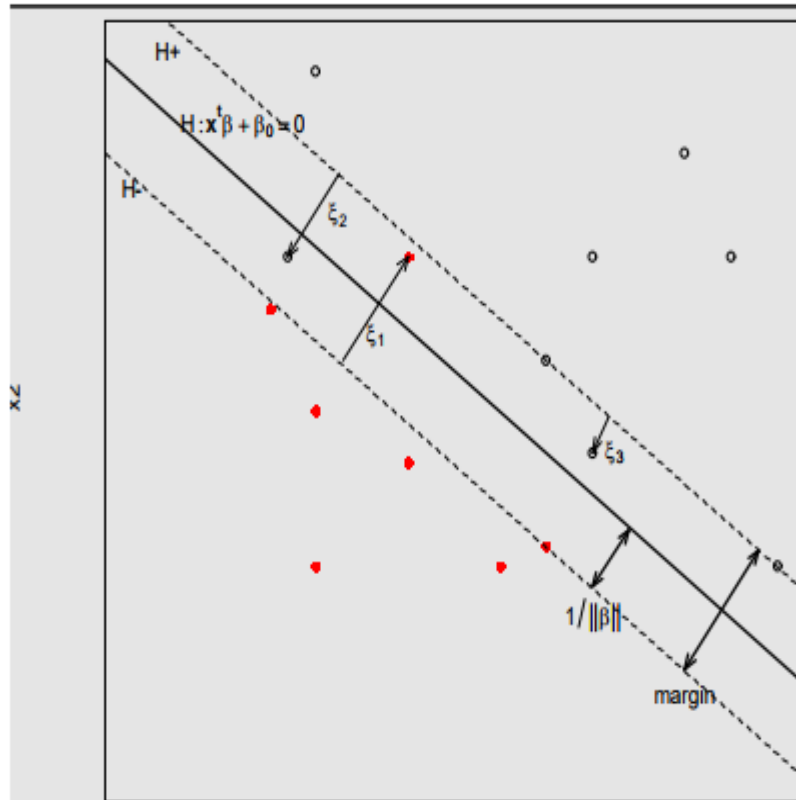


Dendrogram of Hierarchical CA: Ward Linkage

Metric Multidimensional Scaling

[5단계] 개체 분류 : SVM을 활용한 판별분석 (최종 용어 27,203 -> 13,980)

- 텍스트 분류 분석에서 자주 이용되는 machine learning 중 지도학습(supervised learning) 알고리즘 SVM
- 학습 시간이 데이터 개수와 무관, 기존의 분류 방법인 판별분석과 로지스틱 회귀분석보다 예측력이 높음
- 마진을 최대로 하는 초평면 H 을 찾아 범주를 나누는 이진 분류기



$$\mathcal{L} = \{(\mathbf{x}_i, y_i) : i = 1, \dots, n\}, \text{ s.t. } \mathbf{x}_i \in \mathbb{R}^p, y_i \in \{-1, 1\}$$

$$H = \{\mathbf{x} : f(\mathbf{x}) = \beta_0 + \mathbf{x}^t \boldsymbol{\beta}\} : \text{초평면}$$

$$\mathbf{x} = (x_1, \dots, x_p)^t : p\text{차원 변수벡터}$$

$$\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^t : \text{가중치 모수 벡터, } \beta_0 : \text{bias 항}$$

$$\text{분류기준: } \hat{f}(\mathbf{x}) > 0, y = +1; \hat{f}(\mathbf{x}) < 0, y = -1$$

$$\min_{\boldsymbol{\beta}, \beta_0} \frac{1}{2} \|\boldsymbol{\beta}\|^2 + C \sum_{i=1}^n \xi_i, \text{ s.t. } \xi_i \geq 0, y_i(\mathbf{x}^t \boldsymbol{\beta} + \beta_0) \geq 1 - \xi_i, i = 1, \dots, n$$

선형적으로 완전히 분리가 되지 않는 경우 : Linearly Non-separable SVM

[4 단계]의 문서-핵심어 가중행렬의 SVM을 이용한 10개 기관 문서의 텍스트 분류

- training : test = 80:20의 비율로 기관별 문서에 대해 층화추출

예측	실제									
	A	B	C	D	E	F	G	H	I	J
A	3	0	0	0	0	0	0	0	0	0
B	0	4	0	0	0	0	0	0	0	0
C	0	0	3	0	0	0	0	0	0	0
D	0	0	0	3	0	0	0	0	0	0
E	0	0	0	0	3	0	0	0	0	0
F	0	0	0	0	0	2	0	0	0	0
G	0	0	0	0	0	0	3	0	0	0
H	0	1	0	0	0	0	0	26	0	1
I	0	0	0	0	0	0	0	0	8	0
J	0	0	0	0	0	0	0	0	0	5

Accuracy : $0.9677 = 60/62 \implies 96.77\%$

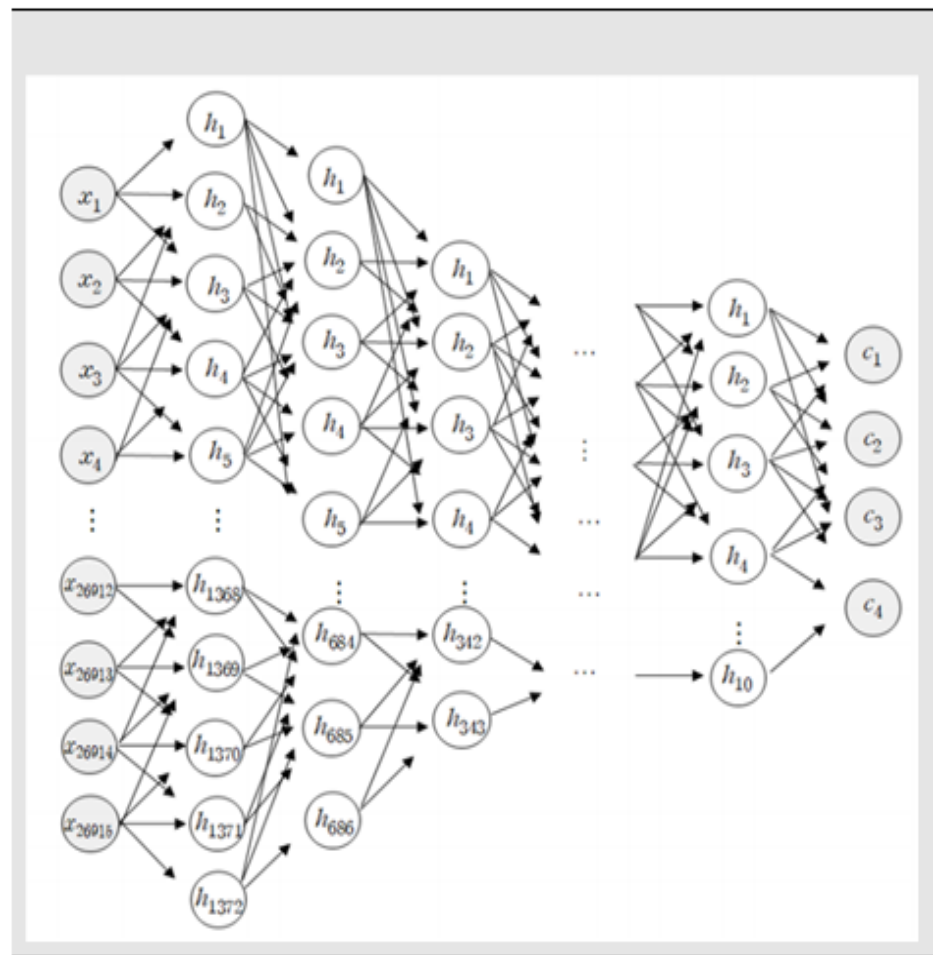
SVM에 의한 10개 기관 문서의 분류표

[5단계] 개체 분류 : DNN을 활용한 분류분석

- machine learning 중 DNN(Deep Neural Network)
- In general, accuracies of DDN are higher than SVM's.

IDF : Inverse Document Freq.
IGM : Inverse Gravity Moment

Document-term weighted function



Model	The number of hidden layers and hidden nodes	Accuracy	
		M1 (TF-IDF)	M2 (TF-IGM)
1	86-43-21-10	0.793	0.810
2	172-86-43-21-10	0.801	0.801
3	343-172-86-43-21-10	0.750	0.818
4	686-343-172-86-43-21-10	0.818	0.827
5	1372-686-343-172-86-43-21-10	0.827	0.853

Optimal best model

Research fields	SVM		DNN
Economic policy (C1)	0.940	<	1.000
Resources · Infrastructure (C2)	0.822	<	0.853
Public policy (C3)	0.862	<	0.991
Human resources (C4)	0.947	>	0.896

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

TP(true positive): 실제 참인 범주를 참으로 정확하게 예측한 빈도
FP(false positive): 실제 거짓인 범주를 참으로 예측한 빈도
FN(false negative): 실제 참인 범주를 거짓으로 예측한 빈도
TN(true negative): 실제 거짓인 범주를 거짓으로 예측한 빈도

입력층: 26,915 용어, 출력층: 4가지, 은닉층의 수= 8개, 은닉 수 1,372의 절반 씩 감소 시킨 구조