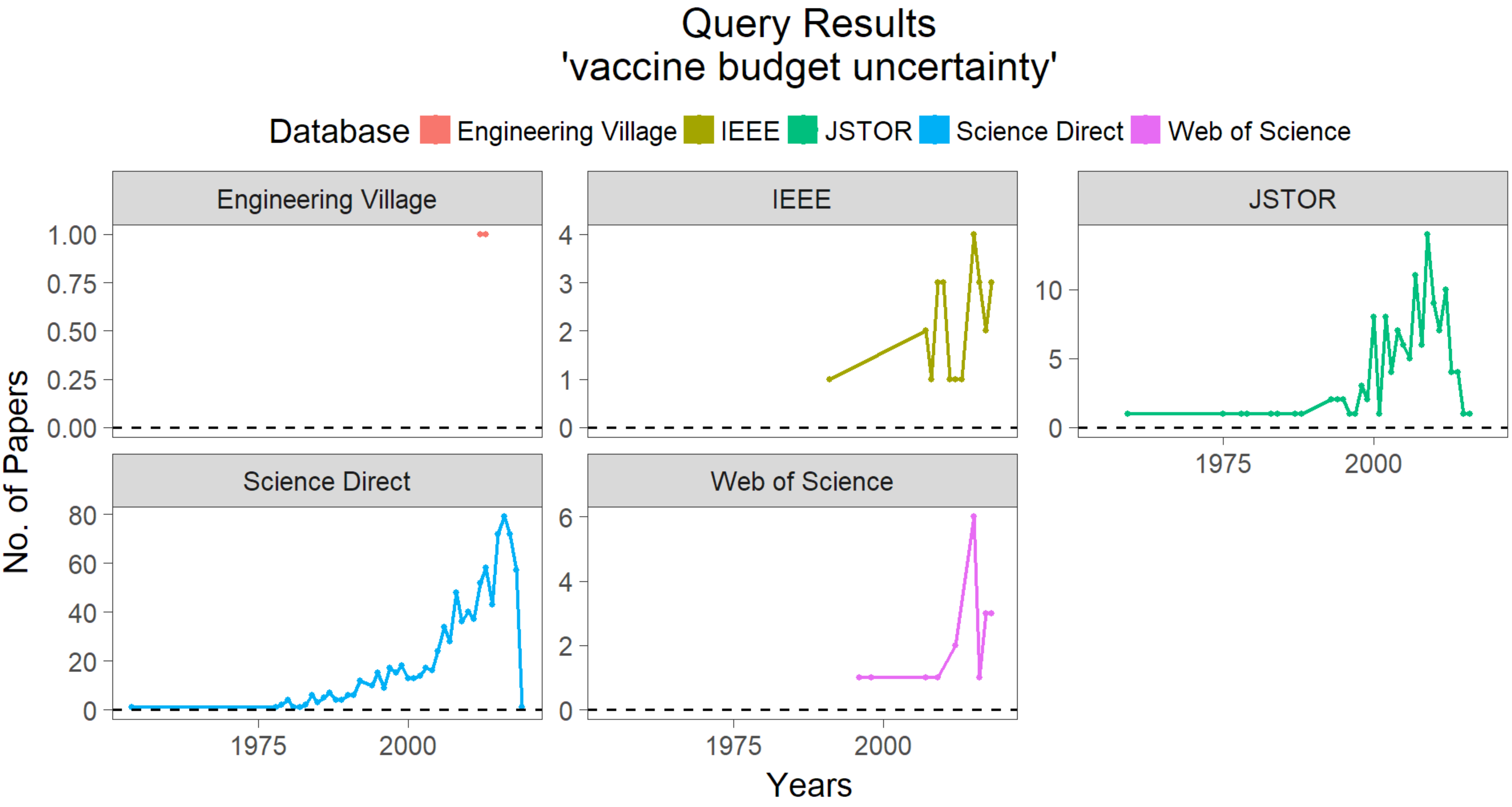
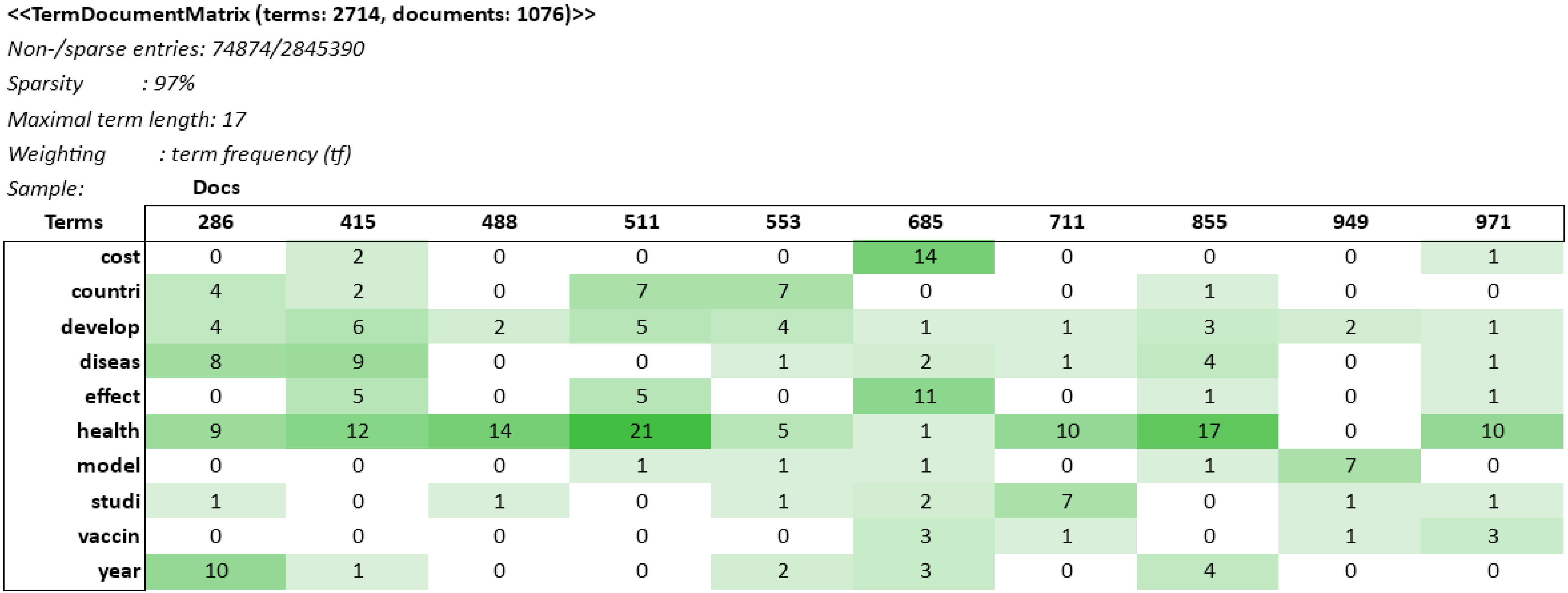


Query Results
'vaccine budget uncertainty'
Database
Engineering Village
IEEE
JSTOR
Science Direct
Web of Science
Engineering Village
1.00
0.75
0.50
0.25
0.00
IEEE
4
3
2
1
0
JSTOR
10
5
0
1975
2000
Science Direct
80
60
40
20
0
1975
2000
Web of Science
6
4
2
0
1975
2000
No. of Papers
Years

Term-Document Matrix

Word-Abstract Matrix



<<TermDocumentMatrix (terms: 2714, documents: 1076)>>

Non-/sparse entries: 74874/2845390

Sparsity : 97%

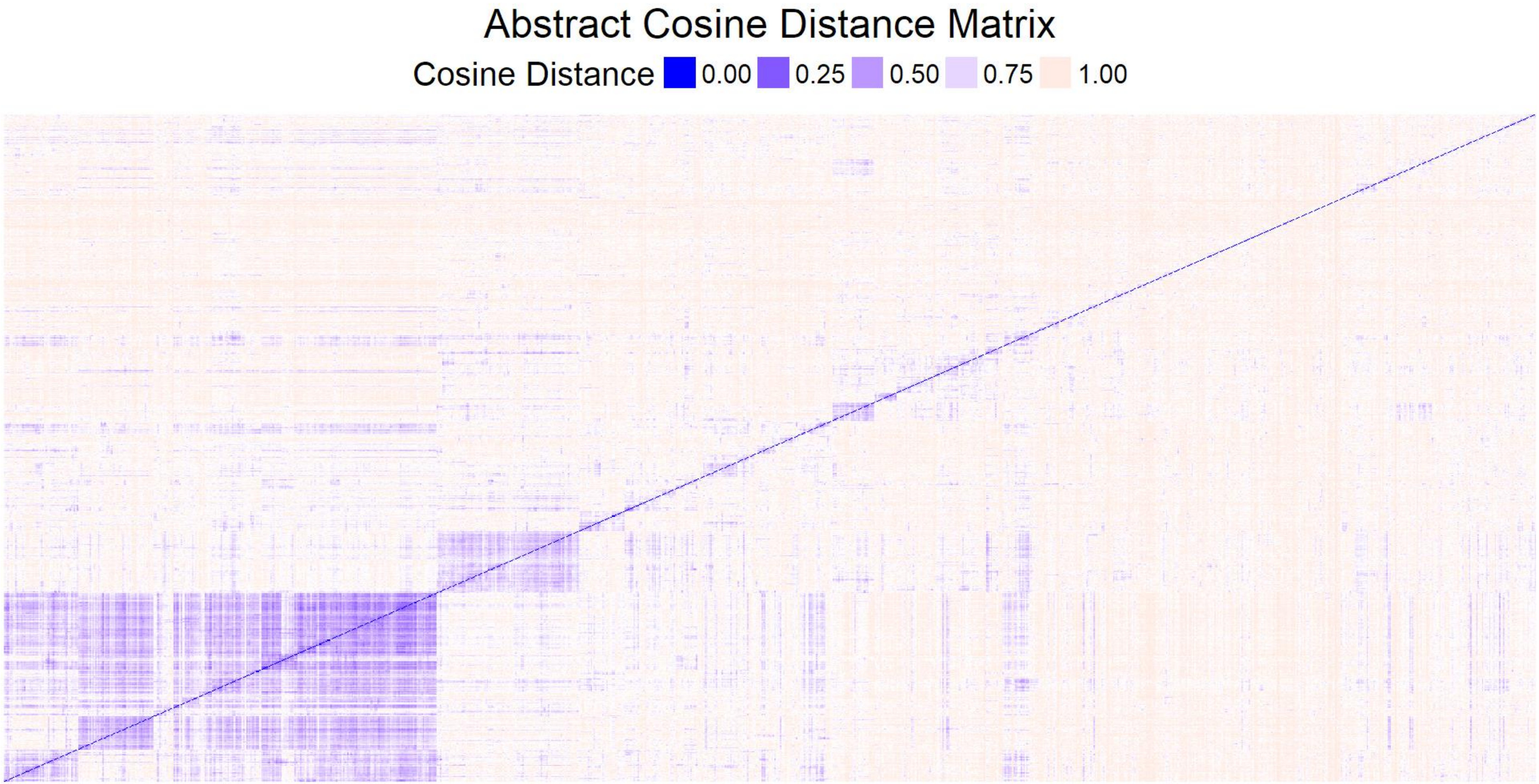
Maximal term length: 17

Weighting : term frequency (tf)

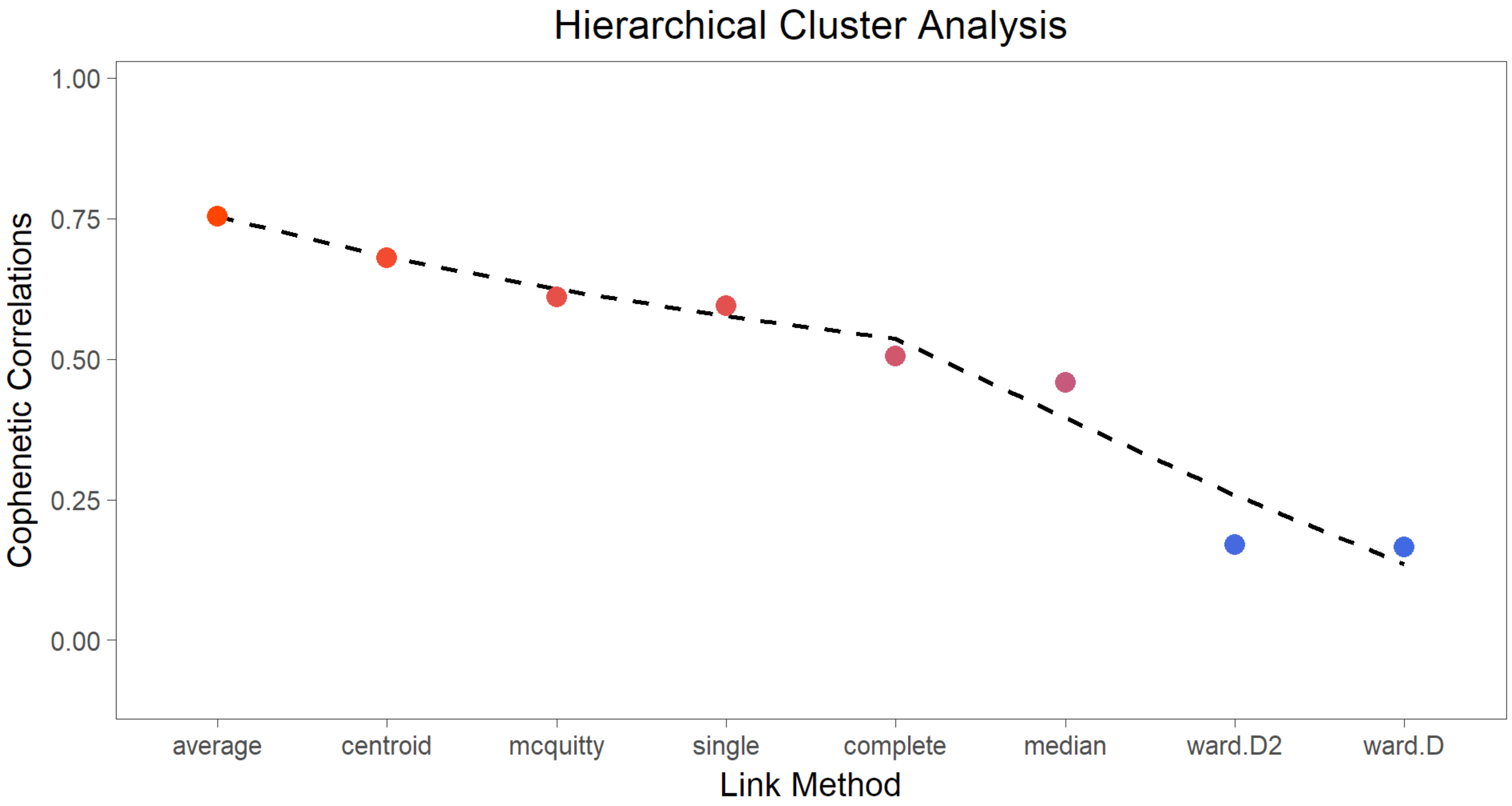
Sample: **Docs**

Terms	**286**	**415**	**488**	**511**	**553**	**685**	**711**	**855**	**949**	**971**
cost	0	2	0	0	0	14	0	0	0	1
countri	4	2	0	7	7	0	0	1	0	0
develop	4	6	2	5	4	1	1	3	2	1
diseas	8	9	0	0	1	2	1	4	0	1
effect	0	5	0	5	0	11	0	1	0	1
health	9	12	14	21	5	1	10	17	0	10
model	0	0	0	1	1	1	0	1	7	0
studi	1	0	1	0	1	2	7	0	1	1
vaccin	0	0	0	0	0	3	1	0	1	3
year	10	1	0	0	2	3	0	4	0	0

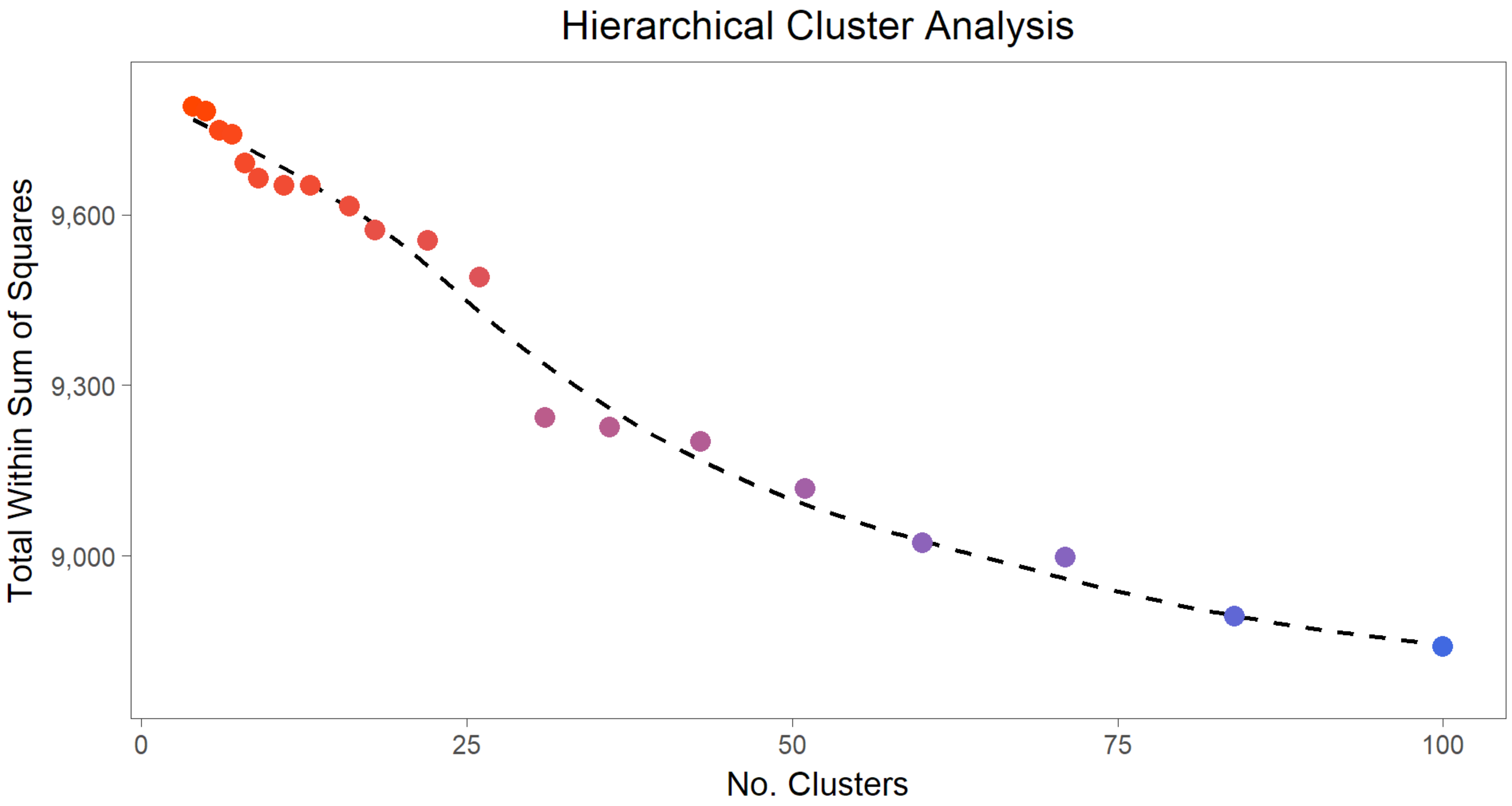
Term-Document Matrix →
Document Cosine Distance Matrix



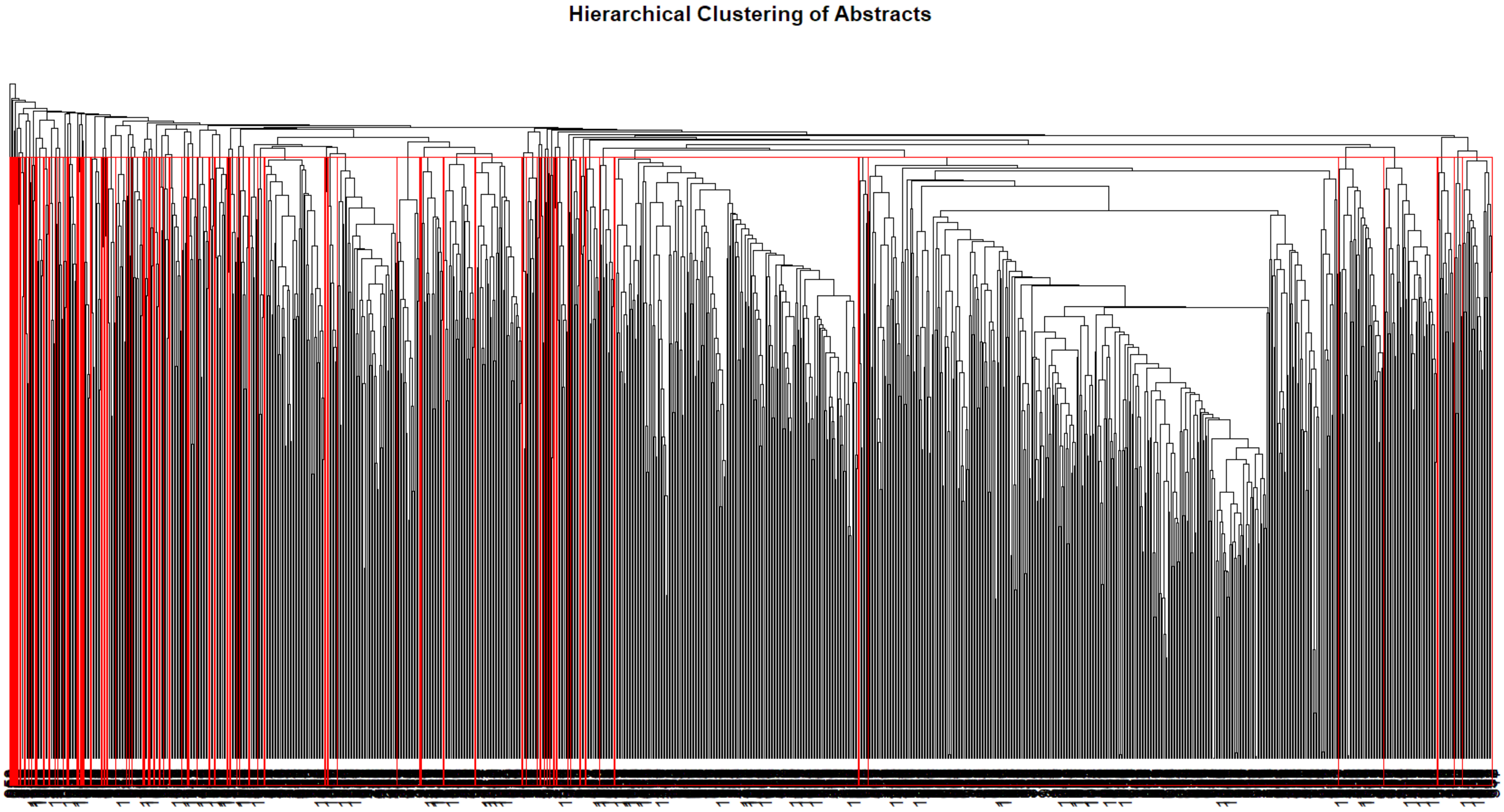
Clustering the Document Cosine Distance Matrix



Clustering the Document Cosine Distance Matrix

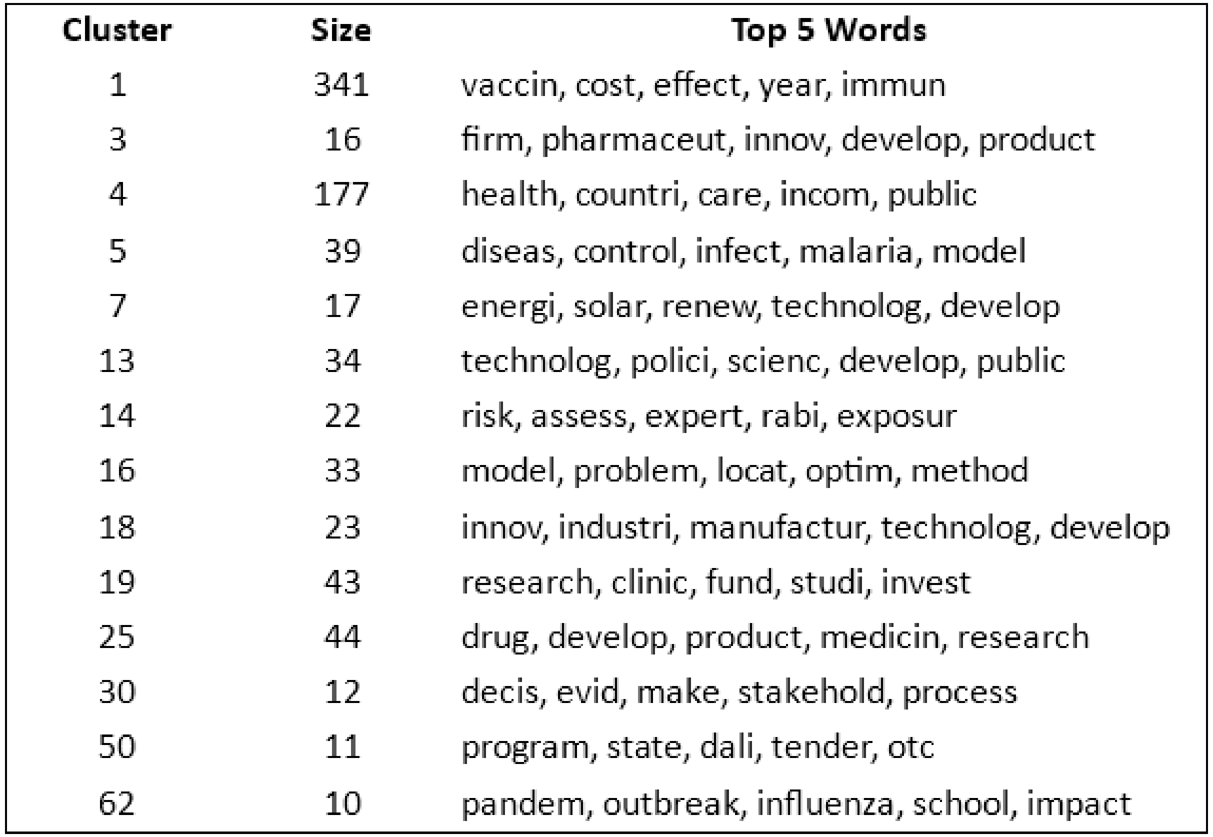


Clustering the Document Cosine Distance Matrix



Words in Abstract Clusters

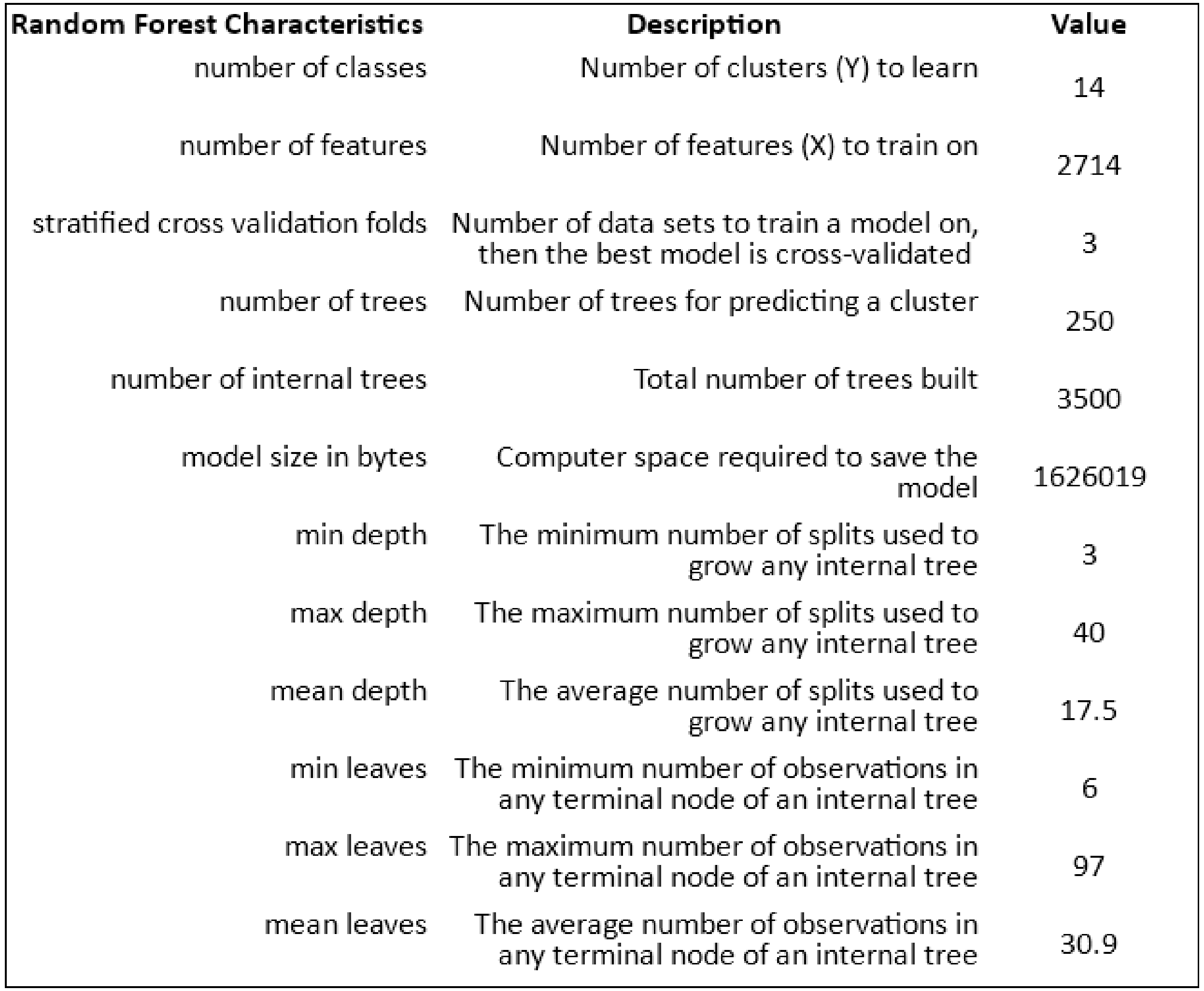
Clusters with less than 10 abstracts are ignored



Cluster	Size	Top 5 Words
1	341	vaccin, cost, effect, year, immun
3	16	firm, pharmaceut, innov, develop, product
4	177	health, countri, care, incom, public
5	39	diseas, control, infect, malaria, model
7	17	energi, solar, renew, technolog, develop
13	34	technolog, polici, scienc, develop, public
14	22	risk, assess, expert, rabi, exposur
16	33	model, problem, locat, optim, method
18	23	innov, industri, manufactur, technolog, develop
19	43	research, clinic, fund, studi, invest
25	44	drug, develop, product, medicin, research
30	12	decis, evid, make, stakehold, process
50	11	program, state, dali, tender, otc
62	10	pandem, outbreak, influenza, school, impact

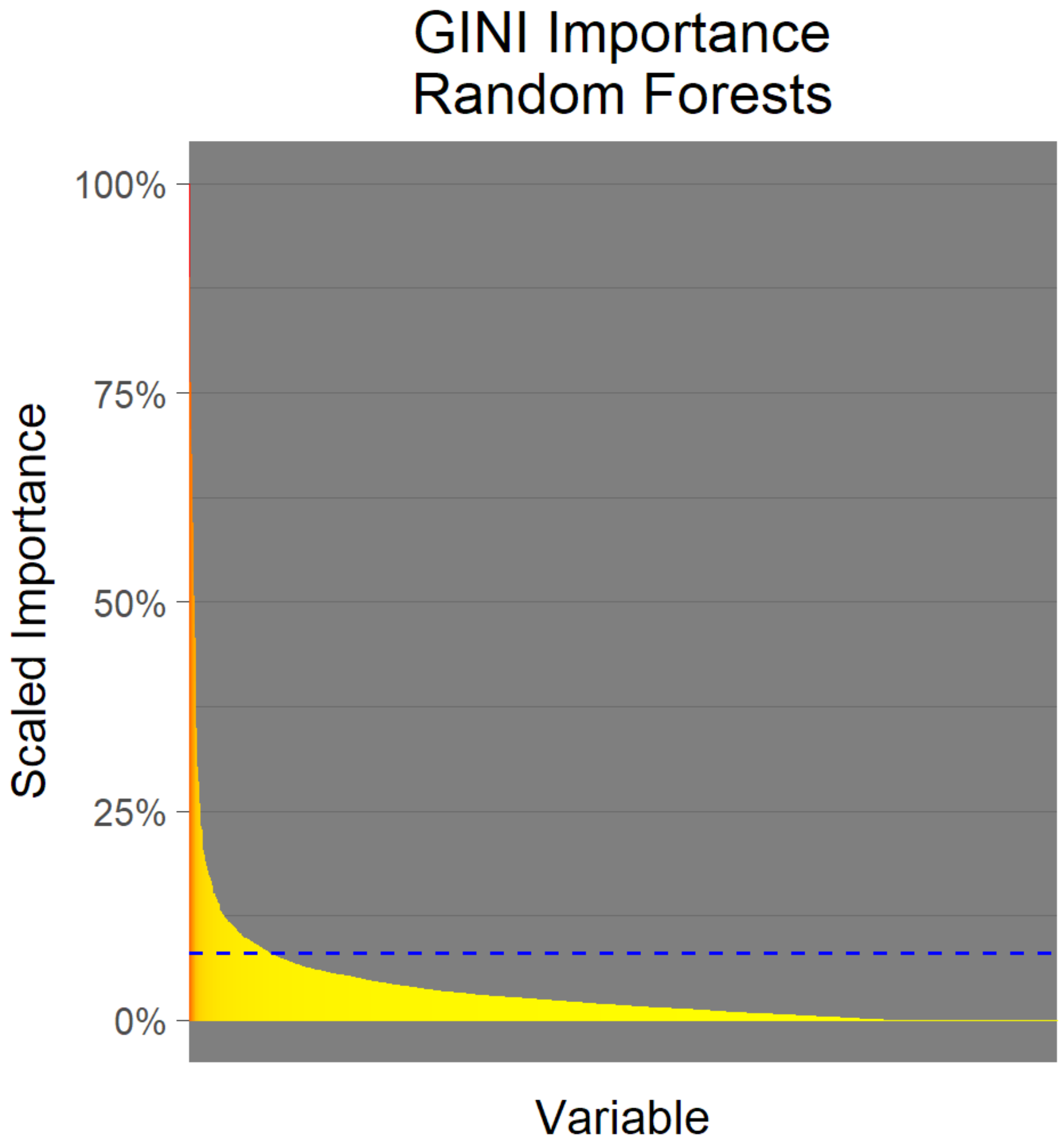
Ranking Words with Random Forests

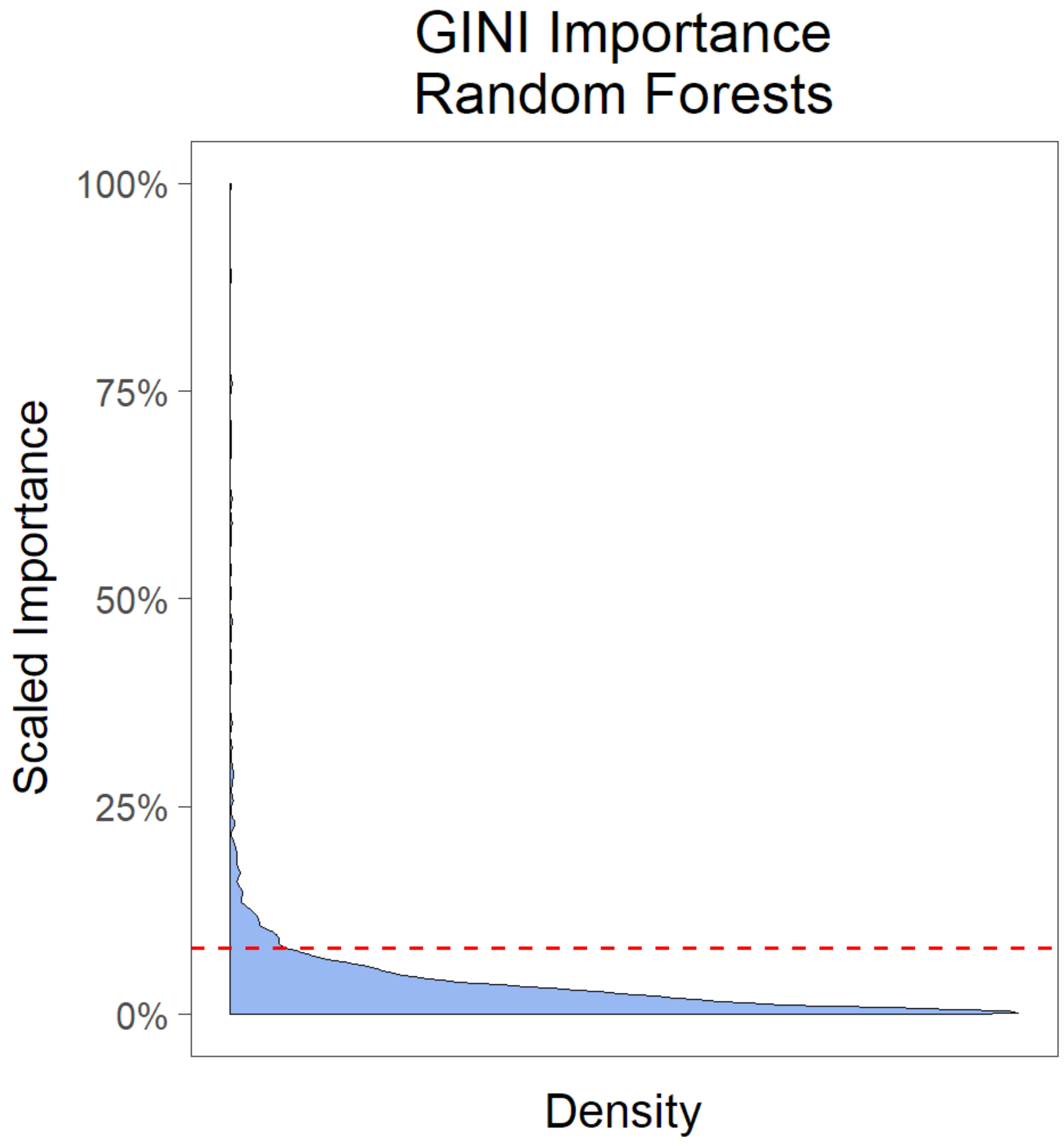
features: *term frequency-inverse document frequency (tfidf) matrix*



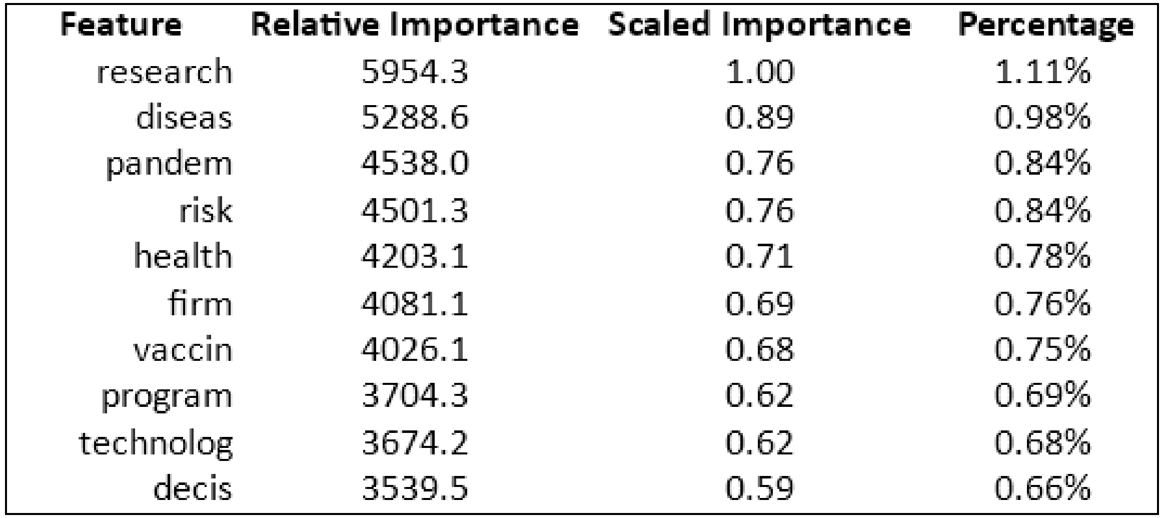
Random Forest Characteristics	Description	Value
number of classes	Number of clusters (Y) to learn	14
number of features	Number of features (X) to train on	2714
stratified cross validation folds	Number of data sets to train a model on, then the best model is cross-validated	3
number of trees	Number of trees for predicting a cluster	250
number of internal trees	Total number of trees built	3500
model size in bytes	Computer space required to save the model	1626019
min depth	The minimum number of splits used to grow any internal tree	3
max depth	The maximum number of splits used to grow any internal tree	40
mean depth	The average number of splits used to grow any internal tree	17.5
min leaves	The minimum number of observations in any terminal node of an internal tree	6
max leaves	The maximum number of observations in any terminal node of an internal tree	97
mean leaves	The average number of observations in any terminal node of an internal tree	30.9

Feature Selection by Importance





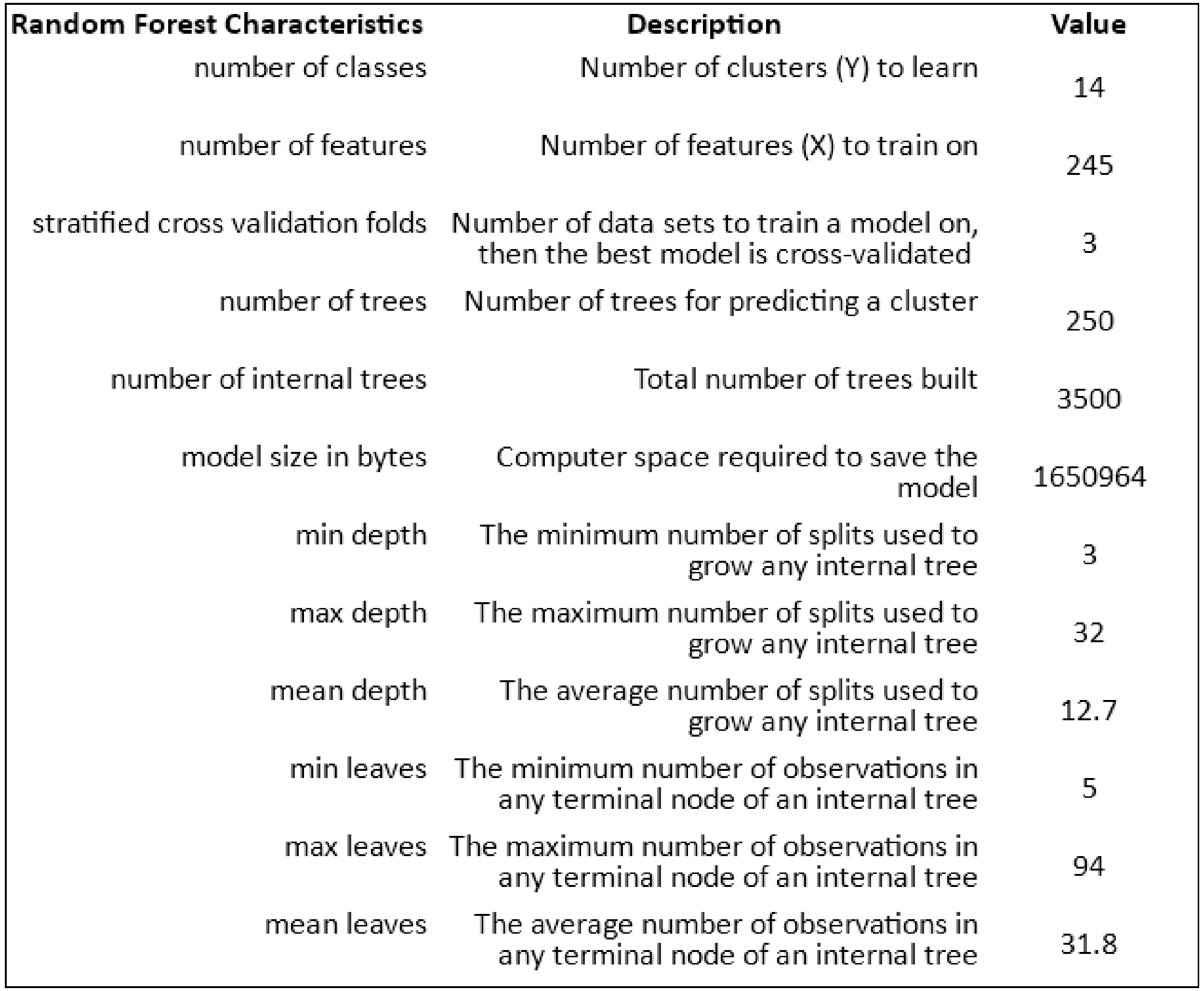
Top 10 Words for Classifying Clusters



Feature	Relative Importance	Scaled Importance	Percentage
research	5954.3	1.00	1.11%
diseas	5288.6	0.89	0.98%
pandem	4538.0	0.76	0.84%
risk	4501.3	0.76	0.84%
health	4203.1	0.71	0.78%
firm	4081.1	0.69	0.76%
vaccin	4026.1	0.68	0.75%
program	3704.3	0.62	0.69%
technolog	3674.2	0.62	0.68%
decis	3539.5	0.59	0.66%

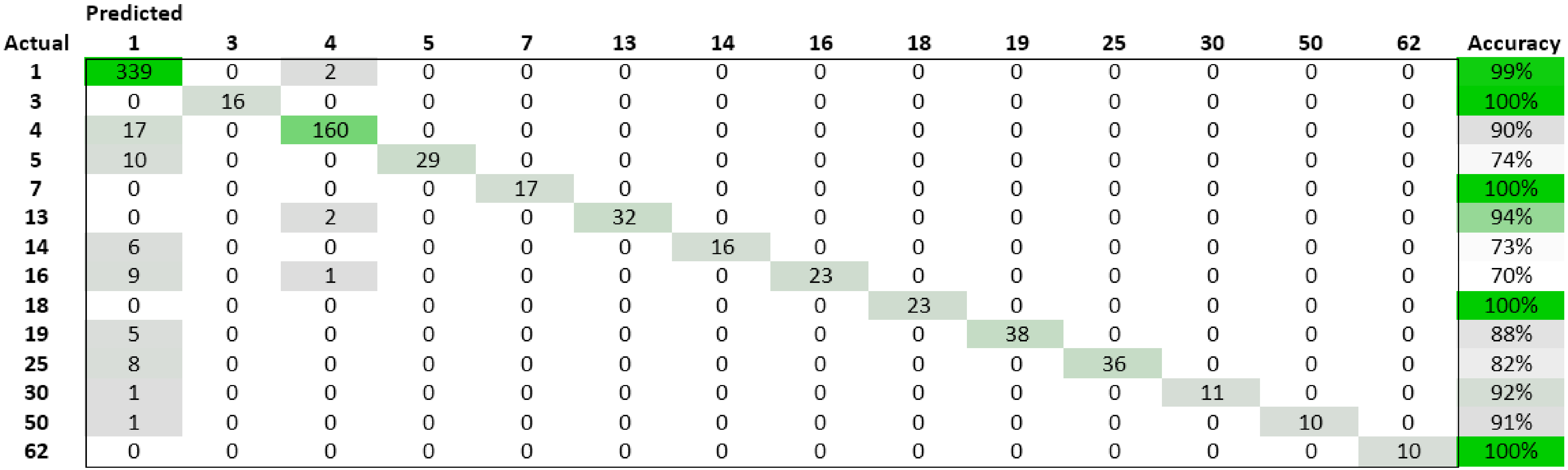
Learning Clusters with Random Forests

features: 245 words from the tfidf *matrix*



Random Forest Characteristics	Description	Value
number of classes	Number of clusters (Y) to learn	14
number of features	Number of features (X) to train on	245
stratified cross validation folds	Number of data sets to train a model on, then the best model is cross-validated	3
number of trees	Number of trees for predicting a cluster	250
number of internal trees	Total number of trees built	3500
model size in bytes	Computer space required to save the model	1650964
min depth	The minimum number of splits used to grow any internal tree	3
max depth	The maximum number of splits used to grow any internal tree	32
mean depth	The average number of splits used to grow any internal tree	12.7
min leaves	The minimum number of observations in any terminal node of an internal tree	5
max leaves	The maximum number of observations in any terminal node of an internal tree	94
mean leaves	The average number of observations in any terminal node of an internal tree	31.8

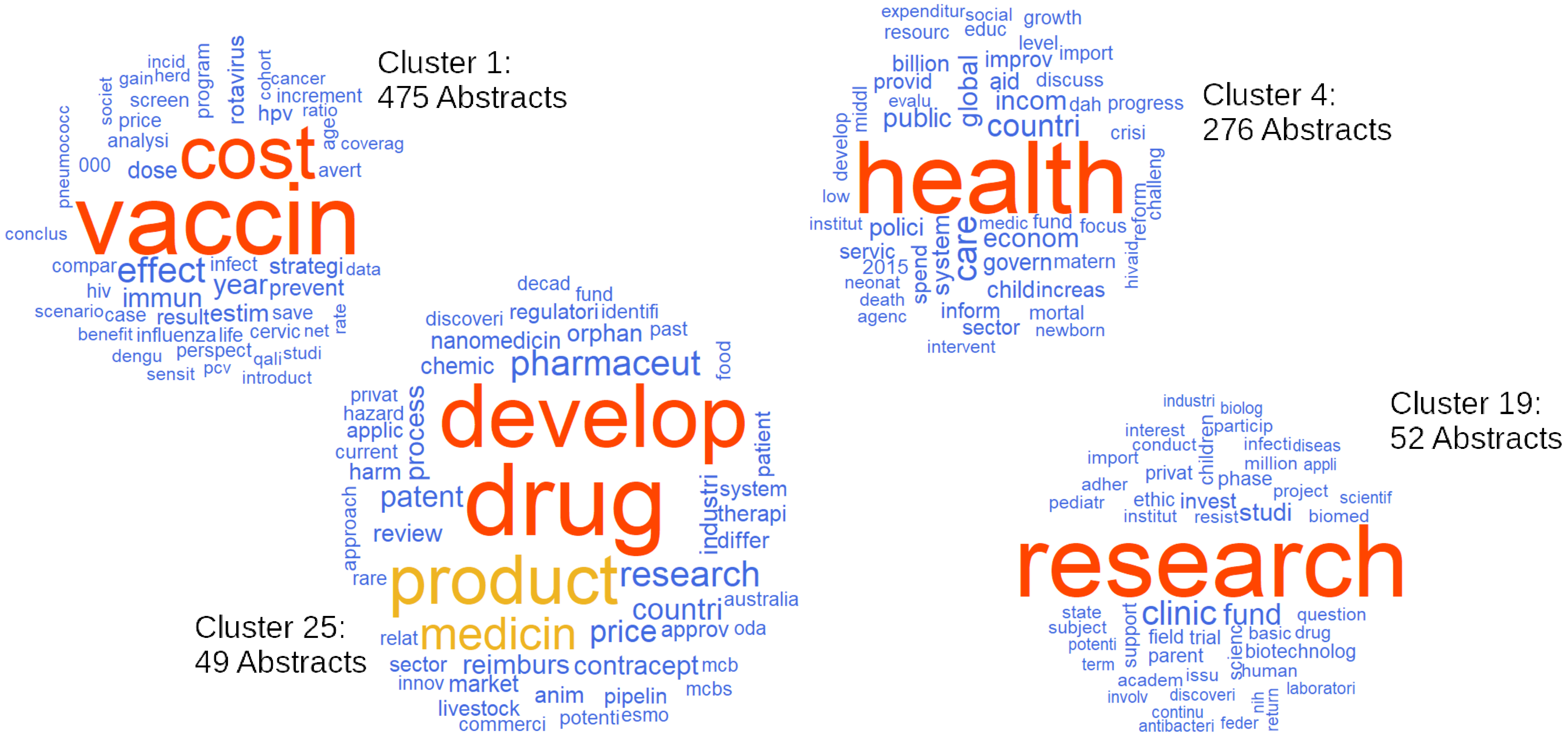
Random Forest Confusion Matrix



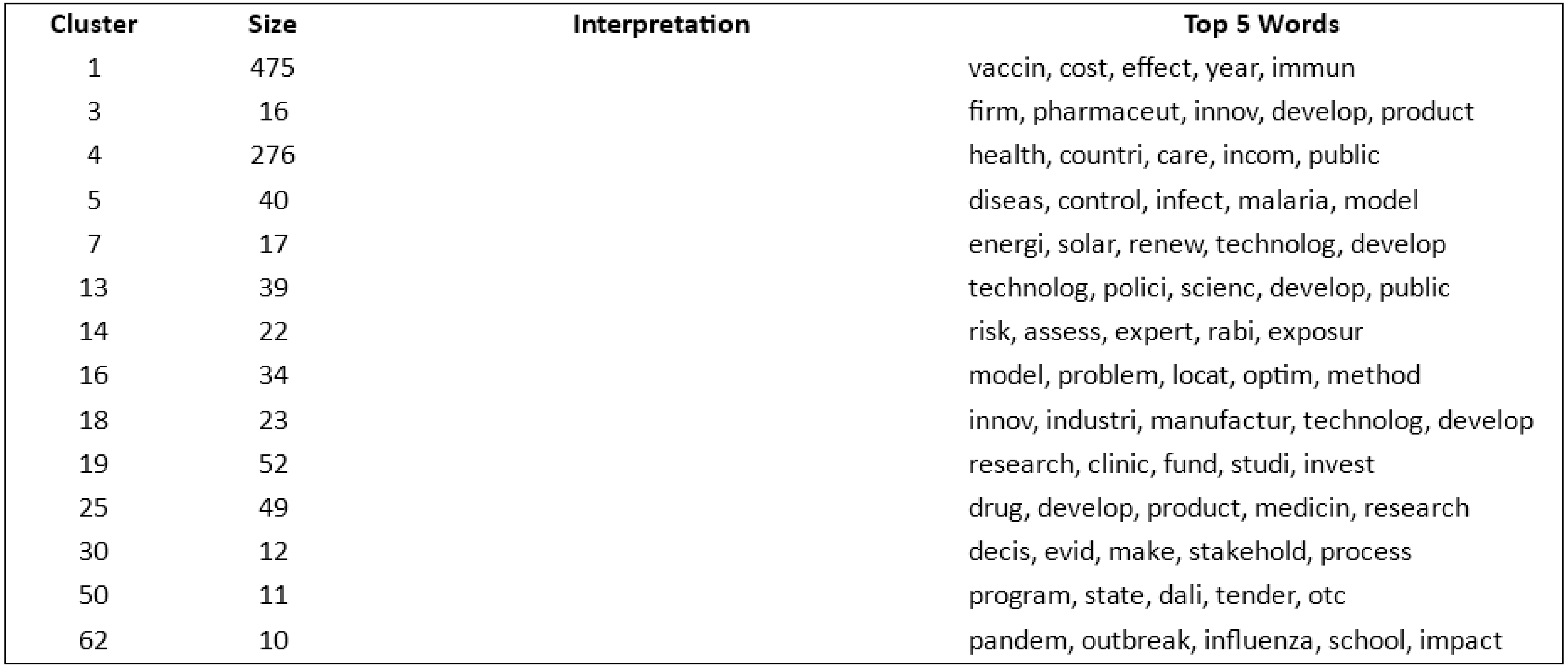
	Predicted														
Actual	**1**	**3**	**4**	**5**	**7**	**13**	**14**	**16**	**18**	**19**	**25**	**30**	**50**	**62**	**Accuracy**
1	339	0	2	0	0	0	0	0	0	0	0	0	0	0	99%
3	0	16	0	0	0	0	0	0	0	0	0	0	0	0	100%
4	17	0	160	0	0	0	0	0	0	0	0	0	0	0	90%
5	10	0	0	29	0	0	0	0	0	0	0	0	0	0	74%
7	0	0	0	0	17	0	0	0	0	0	0	0	0	0	100%
13	0	0	2	0	0	32	0	0	0	0	0	0	0	0	94%
14	6	0	0	0	0	0	16	0	0	0	0	0	0	0	73%
16	9	0	1	0	0	0	0	23	0	0	0	0	0	0	70%
18	0	0	0	0	0	0	0	0	23	0	0	0	0	0	100%
19	5	0	0	0	0	0	0	0	0	38	0	0	0	0	88%
25	8	0	0	0	0	0	0	0	0	0	36	0	0	0	82%
30	1	0	0	0	0	0	0	0	0	0	0	11	0	0	92%
50	1	0	0	0	0	0	0	0	0	0	0	0	10	0	91%
62	0	0	0	0	0	0	0	0	0	0	0	0	0	10	100%

Words in Clusters

Random Forest re-classified the ignored clusters



Reading Abstracts to Interpret Clusters



Cluster	Size	Interpretation	Top 5 Words
1	475		vaccin, cost, effect, year, immun
3	16		firm, pharmaceut, innov, develop, product
4	276		health, countri, care, incom, public
5	40		diseas, control, infect, malaria, model
7	17		energi, solar, renew, technolog, develop
13	39		technolog, polici, scienc, develop, public
14	22		risk, assess, expert, rabi, exposur
16	34		model, problem, locat, optim, method
18	23		innov, industri, manufactur, technolog, develop
19	52		research, clinic, fund, studi, invest
25	49		drug, develop, product, medicin, research
30	12		decis, evid, make, stakehold, process
50	11		program, state, dali, tender, otc
62	10		pandem, outbreak, influenza, school, impact