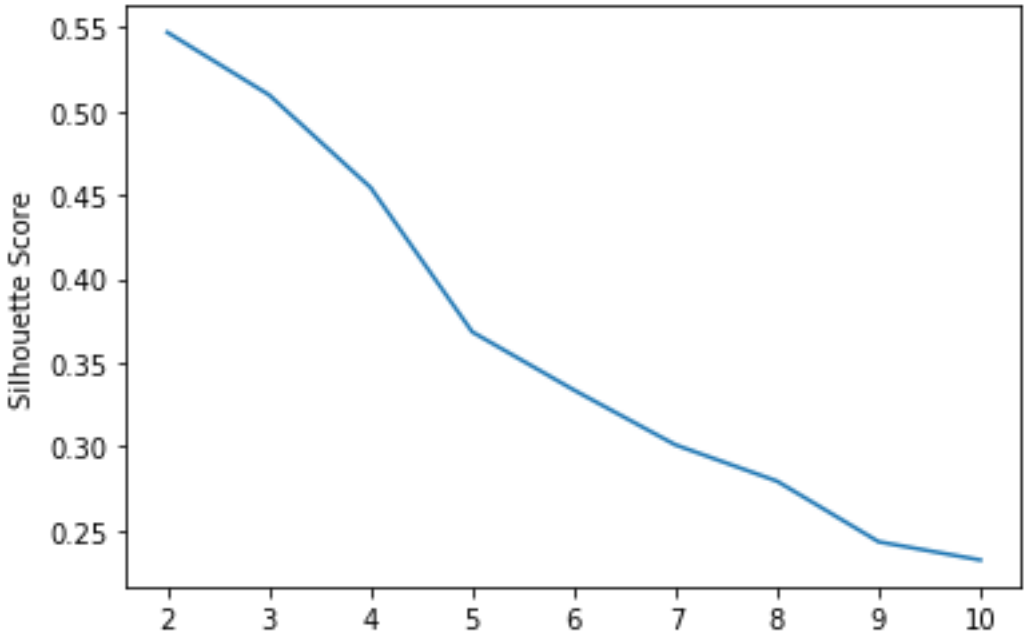
Statistical data clustering

The Nonnegative Matrix Factorization is used as the topic modelling/clustering technique. We use the Silhouette score, within-cluster, between-cluster dispersions to identify the optimal number of topics/clusters.

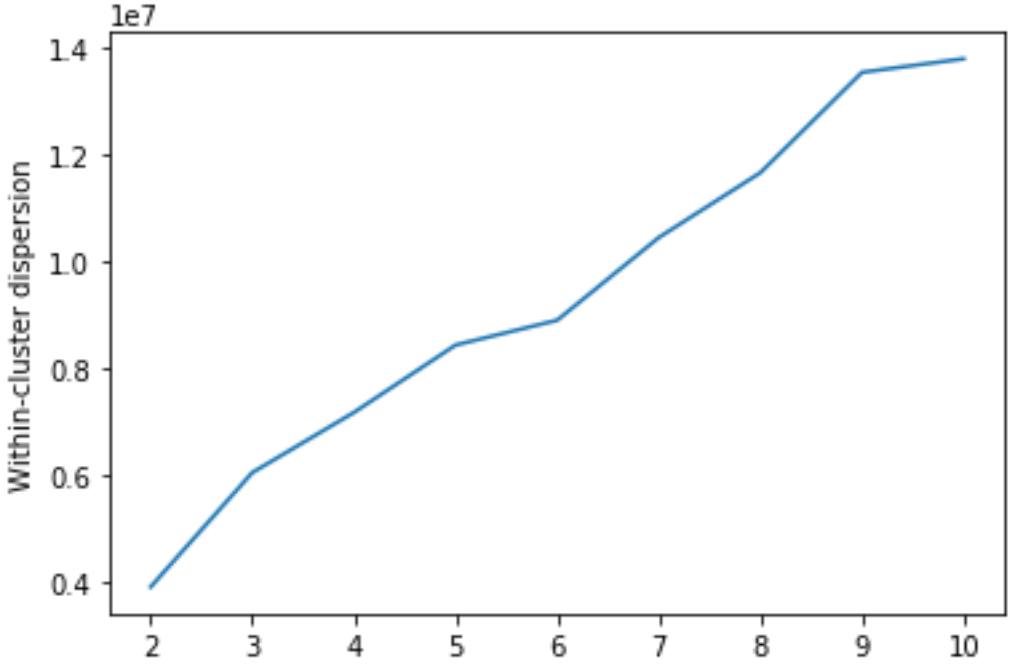
Silhouette score

X-axis is the number of clusters and y-axis is the Silhouette score. Higher the value better is the clustering quality. From the figure shown below, we can see that increasing the number of clusters decreases the cluster quality. This indicates, we have one big cluster and multiple small clusters.

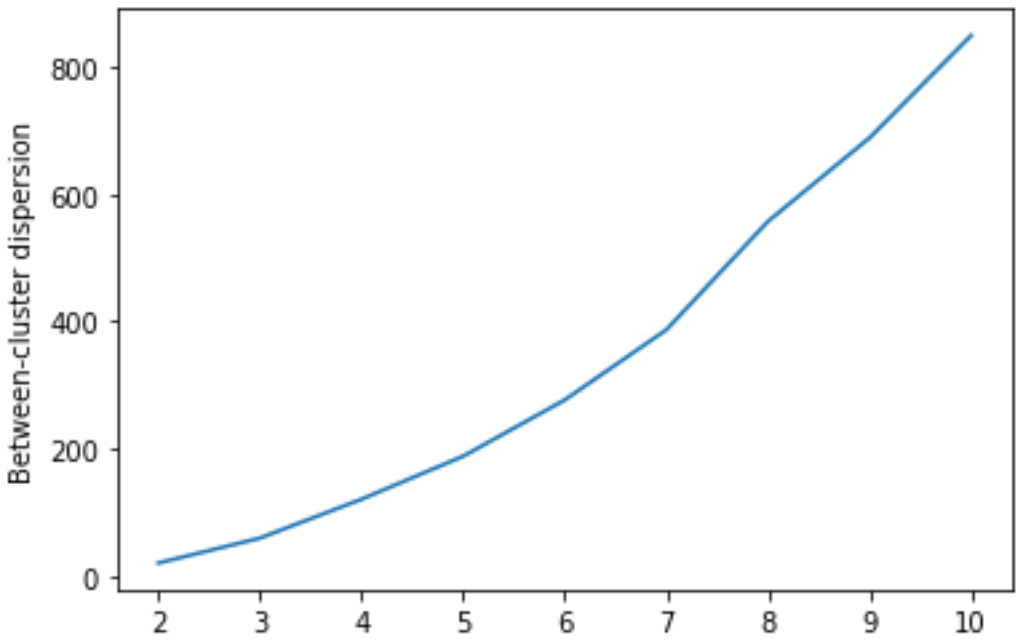


Within-cluster and between cluster dispersion

Within cluster dispersion measures the sum of distances squared between the data points within the cluster. It should be minimum.



Between-cluster dispersion measures the sum of distances squared between clusters. It should be higher.

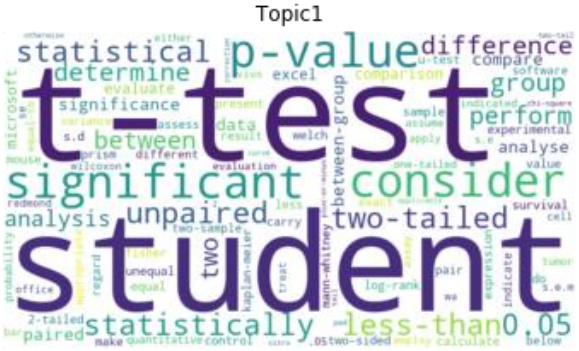


In our results, the clustering maximizes the between-cluster dispersion but does not minimizes the within-cluster dispersion.

Next, we will have the results by setting the number of clusters to 10, 2, and 5.

Topics (10)

Topic1



Topic2

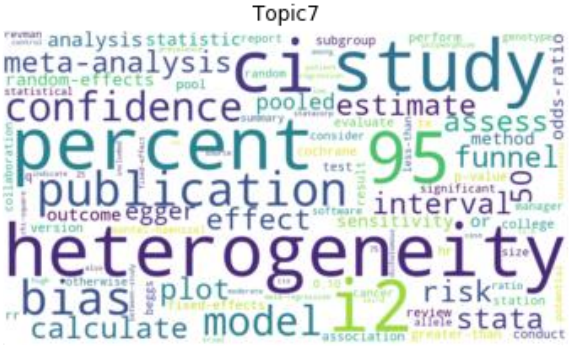
Topic3

Topic4

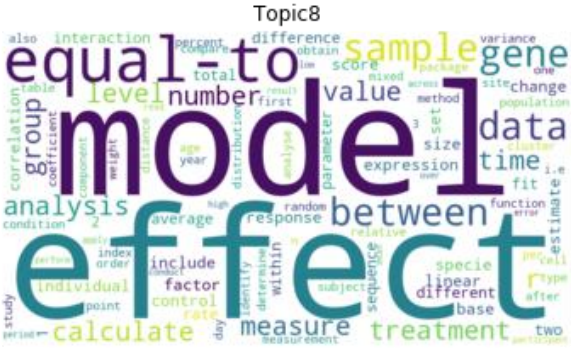
Topic5

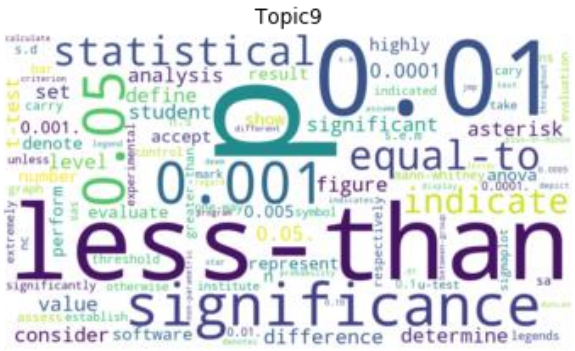
Topic6

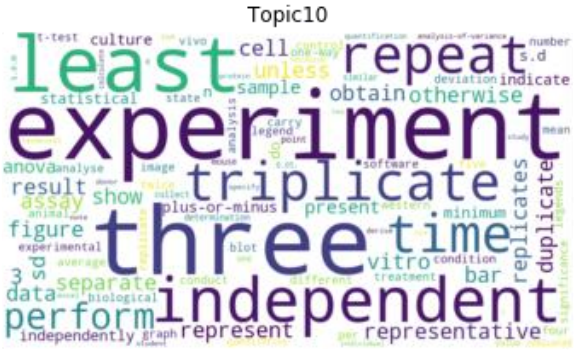
Topic7



Topic8





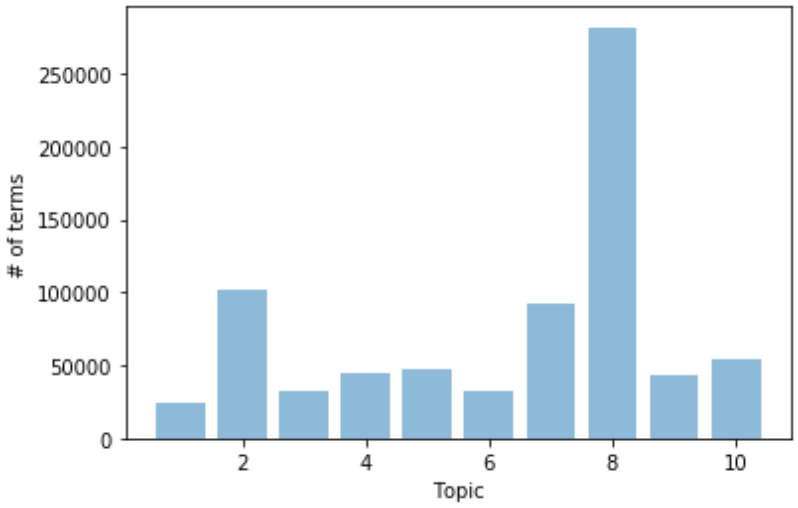


Number of documents in each cluster

8	38183
2	19478
4	10163
3	9879
5	9574
9	6104
7	5165
6	4746
10	4664
1	3775

Size of the topics

#terms	topic
24686	1
101372	2
32808	3
45178	4
47946	5
32226	6
91952	7
281656	8
43533	9
53673	10



Median number of terms/words by topic (and inter-quartile range)

```
Total number of words: 179360
Median number of words by topic: 46562.0
IQR of words by topic: 34642.0
```

A few DOIs for papers that were a strong match in each topic

Topic 1

	DOI	clusterID	value
1	10.1371/journal.pone.0077872	1	0.102699
2	10.1371/journal.pone.0144950	1	0.102699
3	10.1371/journal.pone.0019412	1	0.102699
4	10.1371/journal.pone.0055941	1	0.102153
5	10.1371/journal.pone.0101993	1	0.100685
6	10.1371/journal.pone.0080609	1	0.100685
7	10.1371/journal.pone.0112761	1	0.100685
8	10.1371/journal.pone.0023991	1	0.100685
9	10.1371/journal.pone.0061696	1	0.100685
10	10.1371/journal.pone.0136728	1	0.100685

Topic 2

	DOI	clusterID	value
1	10.1371/journal.pone.0020433	2	0.113397
2	10.1371/journal.pone.0197700	2	0.110135
3	10.1371/journal.pone.0147387	2	0.106858
4	10.1371/journal.pone.0108339	2	0.106573
5	10.1371/journal.pone.0232768	2	0.105949
6	10.1371/journal.pone.0148400	2	0.105390
7	10.1371/journal.pone.0220691	2	0.105375
8	10.1371/journal.pone.0231092	2	0.101563
9	10.1371/journal.pone.0203437	2	0.100246
10	10.1371/journal.pone.0146836	2	0.099964

Topic 3

	DOI	clusterID	value
1	10.1371/journal.pone.0045453	3	0.152681
2	10.1371/journal.pone.0043519	3	0.152216
3	10.1371/journal.pone.0170640	3	0.147192
4	10.1371/journal.pone.0109516	3	0.137825
5	10.1371/journal.pone.0126871	3	0.137092
6	10.1371/journal.pone.0058148	3	0.136884
7	10.1371/journal.pone.0036750	3	0.136884
8	10.1371/journal.pone.0075907	3	0.136184
9	10.1371/journal.pone.0012701	3	0.136150
10	10.1371/journal.pone.0012357	3	0.134891

Topic 4

	DOI	clusterID	value
1	10.1371/journal.pone.0084327	4	0.113318
2	10.1371/journal.pone.0153818	4	0.108406
3	10.1371/journal.pone.0223954	4	0.106164
4	10.1371/journal.pone.0040801	4	0.105727
5	10.1371/journal.pone.0131760	4	0.105097
6	10.1371/journal.pone.0066413	4	0.103207
7	10.1371/journal.pone.0121161	4	0.102749
8	10.1371/journal.pone.0098207	4	0.101825
9	10.1371/journal.pone.0101260	4	0.101468
10	10.1371/journal.pone.0155490	4	0.101443

Topic 5

	DOI	clusterID	value
1	10.1371/journal.pone.0192908	5	0.147440
2	10.1371/journal.pone.0161684	5	0.143619
3	10.1371/journal.pone.0165798	5	0.140842
4	10.1371/journal.pone.0119595	5	0.140158
5	10.1371/journal.pone.0034931	5	0.139169
6	10.1371/journal.pone.0158777	5	0.134007
7	10.1371/journal.pone.0201281	5	0.133451
8	10.1371/journal.pone.0119013	5	0.132399
9	10.1371/journal.pone.0041607	5	0.131723
10	10.1371/journal.pone.0102103	5	0.130998

Topic 6

	DOI	clusterID	value
1	10.1371/journal.pone.0083363	6	0.105452
2	10.1371/journal.pone.0066159	6	0.101841
3	10.1371/journal.pone.0218531	6	0.095548
4	10.1371/journal.pone.0094134	6	0.093996
5	10.1371/journal.pone.0186021	6	0.091638
6	10.1371/journal.pone.0058540	6	0.091026
7	10.1371/journal.pone.0118674	6	0.090642
8	10.1371/journal.pone.0078961	6	0.090313
9	10.1371/journal.pone.0087342	6	0.089961
10	10.1371/journal.pone.0023858	6	0.089546

Topic 7

	DOI	clusterID	value
1	10.1371/journal.pone.0094005	7	0.143821
2	10.1371/journal.pone.0102323	7	0.138627
3	10.1371/journal.pone.0109744	7	0.137107
4	10.1371/journal.pone.0174519	7	0.136900
5	10.1371/journal.pone.0130636	7	0.136674
6	10.1371/journal.pone.0090396	7	0.135906
7	10.1371/journal.pone.0161564	7	0.135696
8	10.1371/journal.pone.0050857	7	0.135665
9	10.1371/journal.pone.0144406	7	0.135501
10	10.1371/journal.pone.0095966	7	0.135077

Topic 8

	DOI	clusterID	value
1	10.1371/journal.pone.0004760	8	0.067780
2	10.1371/journal.pone.0228157	8	0.067095
3	10.1371/journal.pone.0211363	8	0.066322
4	10.1371/journal.pone.0090081	8	0.065350
5	10.1371/journal.pone.0167882	8	0.064975
6	10.1371/journal.pone.0032206	8	0.063299
7	10.1371/journal.pone.0143241	8	0.063259
8	10.1371/journal.pone.0089060	8	0.063049
9	10.1371/journal.pone.0054469	8	0.062640
10	10.1371/journal.pone.0057832	8	0.062576

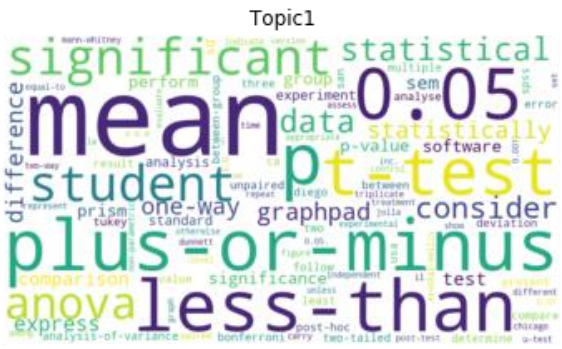
Topic 9

	DOI	clusterID	value
1	10.1371/journal.pone.0159980	9	0.174789
2	10.1371/journal.pone.0145239	9	0.165920
3	10.1371/journal.pone.0108698	9	0.163333
4	10.1371/journal.pone.0034031	9	0.159656
5	10.1371/journal.pone.0133917	9	0.158715
6	10.1371/journal.pone.0037349	9	0.158207
7	10.1371/journal.pone.0047977	9	0.157657
8	10.1371/journal.pone.0096418	9	0.157361
9	10.1371/journal.pone.0134113	9	0.157032
10	10.1371/journal.pone.0032437	9	0.156186

Topic 10

	DOI	clusterID	value
1	10.1371/journal.pone.0060528	10	0.120455
2	10.1371/journal.pone.0109170	10	0.119927
3	10.1371/journal.pone.0186506	10	0.117436
4	10.1371/journal.pone.0118864	10	0.113238
5	10.1371/journal.pone.0159927	10	0.111837
6	10.1371/journal.pone.0051895	10	0.111386
7	10.1371/journal.pone.0130937	10	0.110669
8	10.1371/journal.pone.0093364	10	0.110409
9	10.1371/journal.pone.0084771	10	0.109735
10	10.1371/journal.pone.0029037	10	0.109449

Topics (2)

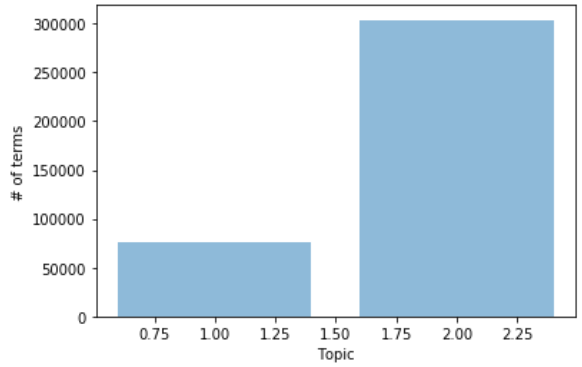


Number of documents in each cluster

```
2    71244
1    40487
```

Size of the topics

#terms	topic
75648	1
303753	2



Median number of terms/words by topic (and inter-quartile range)

```
Total number of words: 179360
Median number of words by topic: 189700.5
IQR of words by topic: 0.0
```

A few DOIs for papers that were a strong match in each topic

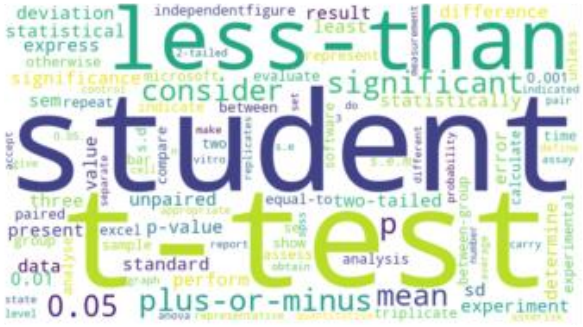
```
Topic 1
                                DOI  clusterID     value
1    10.1371/journal.pone.0066895          1  0.083666
2    10.1371/journal.pone.0195657          1  0.082681
3    10.1371/journal.pone.0161396          1  0.082638
4    10.1371/journal.pone.0120629          1  0.081026
5    10.1371/journal.pone.0097906          1  0.080988
6    10.1371/journal.pone.0071342          1  0.080834
7    10.1371/journal.pone.0193184          1  0.079767
8    10.1371/journal.pone.0115648          1  0.079618
9    10.1371/journal.pone.0038787          1  0.079127
10   10.1371/journal.pone.0184363          1  0.079086
```

Topic 2

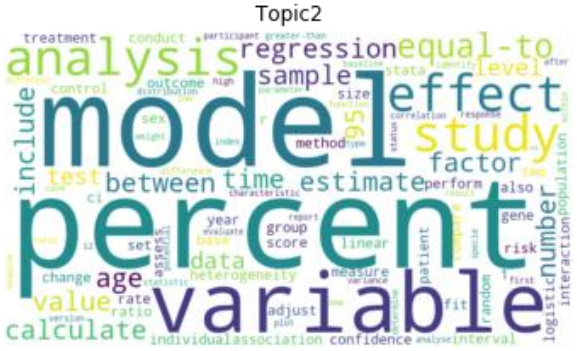
	DOI	clusterID	value
1	10.1371/journal.pone.0020433	2	0.076935
2	10.1371/journal.pone.0181158	2	0.074067
3	10.1371/journal.pone.0205006	2	0.069439
4	10.1371/journal.pone.0217525	2	0.069069
5	10.1371/journal.pone.0182352	2	0.068373
6	10.1371/journal.pone.0184168	2	0.068340
7	10.1371/journal.pone.0220691	2	0.067760
8	10.1371/journal.pone.0215639	2	0.067553
9	10.1371/journal.pone.0230538	2	0.066750
10	10.1371/journal.pone.0100039	2	0.066672

Topics (5)

Topic1



Topic2



Topic3

Topic4

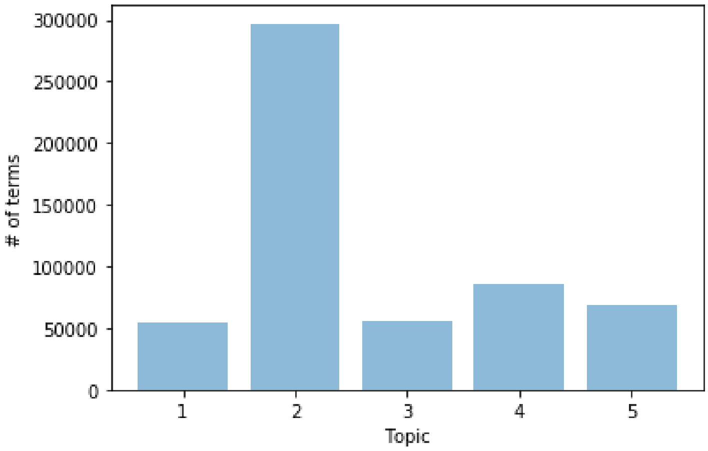
Topic5

Number of documents in each cluster

2	53491
5	18029
4	16740
1	12531
3	10940

Size of the topics

#terms	topic
54780	1
296922	2
56009	3
85558	4
67962	5



Median number of terms/words by topic (and inter-quartile range)

```
Total number of words: 179360
Median number of words by topic: 67962.0
IQR of words by topic: 29549.0
```

A few DOIs for papers that were a strong match in each topic

```
Topic 1
                              DOI  clusterID     value
1   10.1371/journal.pone.0080784          1  0.092128
2   10.1371/journal.pone.0035722          1  0.092128
3   10.1371/journal.pone.0110483          1  0.091057
4   10.1371/journal.pone.0159967          1  0.090409
5   10.1371/journal.pone.0144295          1  0.090253
6   10.1371/journal.pone.0068229          1  0.089819
7   10.1371/journal.pone.0169099          1  0.089812
8   10.1371/journal.pone.0035030          1  0.089671
9   10.1371/journal.pone.0135259          1  0.089572
10  10.1371/journal.pone.0151927          1  0.088694
```

```
Topic 2
                              DOI  clusterID     value
1   10.1371/journal.pone.0228157          2  0.066358
2   10.1371/journal.pone.0161929          2  0.065607
3   10.1371/journal.pone.0165040          2  0.064895
4   10.1371/journal.pone.0195294          2  0.064777
5   10.1371/journal.pone.0039173          2  0.064744
6   10.1371/journal.pone.0040936          2  0.064533
7   10.1371/journal.pone.0030800          2  0.064401
8   10.1371/journal.pone.0087352          2  0.064121
9   10.1371/journal.pone.0057246          2  0.064018
10  10.1371/journal.pone.0080158          2  0.063715
```

Topic 3

	DOI	clusterID	value
1	10.1371/journal.pone.0043519	3	0.145880
2	10.1371/journal.pone.0045453	3	0.145829
3	10.1371/journal.pone.0170640	3	0.141296
4	10.1371/journal.pone.0036750	3	0.133662
5	10.1371/journal.pone.0058148	3	0.133662
6	10.1371/journal.pone.0126871	3	0.131677
7	10.1371/journal.pone.0109516	3	0.131624
8	10.1371/journal.pone.0012701	3	0.131557
9	10.1371/journal.pone.0075907	3	0.131448
10	10.1371/journal.pone.0036902	3	0.131015

Topic 4

	DOI	clusterID	value
1	10.1371/journal.pone.0098207	4	0.109730
2	10.1371/journal.pone.0051180	4	0.108776
3	10.1371/journal.pone.0207844	4	0.108718
4	10.1371/journal.pone.0046301	4	0.108362
5	10.1371/journal.pone.0145237	4	0.105562
6	10.1371/journal.pone.0020198	4	0.105168
7	10.1371/journal.pone.0106533	4	0.105102
8	10.1371/journal.pone.0069592	4	0.104635
9	10.1371/journal.pone.0171819	4	0.104527
10	10.1371/journal.pone.0066413	4	0.104498

Topic 5

	DOI	clusterID	value
1	10.1371/journal.pone.0120046	5	0.111358
2	10.1371/journal.pone.0204950	5	0.110686
3	10.1371/journal.pone.0062685	5	0.110222
4	10.1371/journal.pone.0133783	5	0.110072
5	10.1371/journal.pone.0098797	5	0.109021
6	10.1371/journal.pone.0110978	5	0.107235
7	10.1371/journal.pone.0229517	5	0.105044
8	10.1371/journal.pone.0100707	5	0.104626
9	10.1371/journal.pone.0026906	5	0.104609
10	10.1371/journal.pone.0045760	5	0.104609