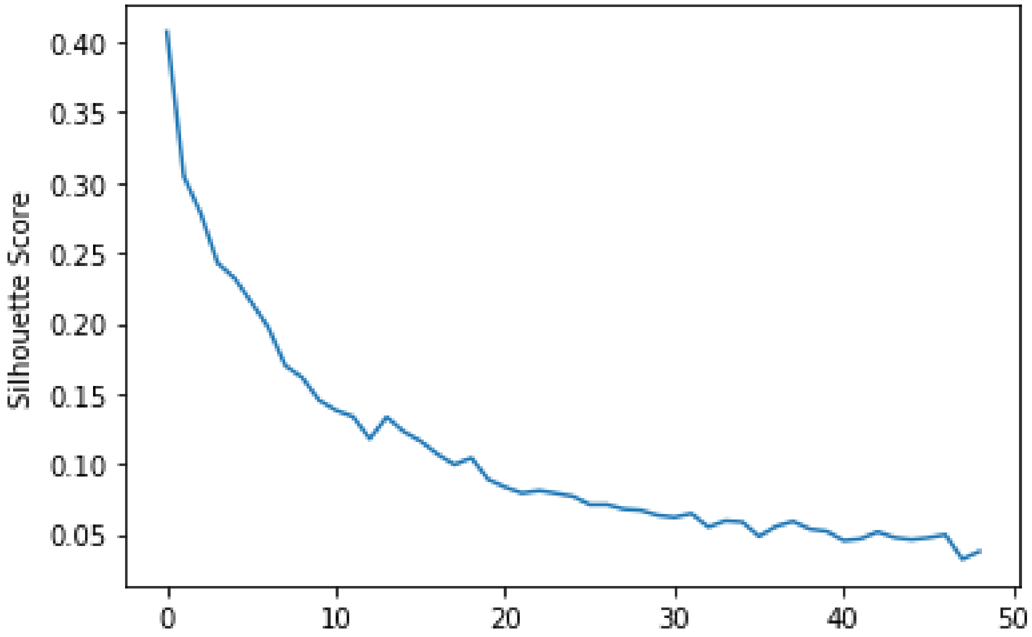
Statistical data clustering – Dataset 2

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.

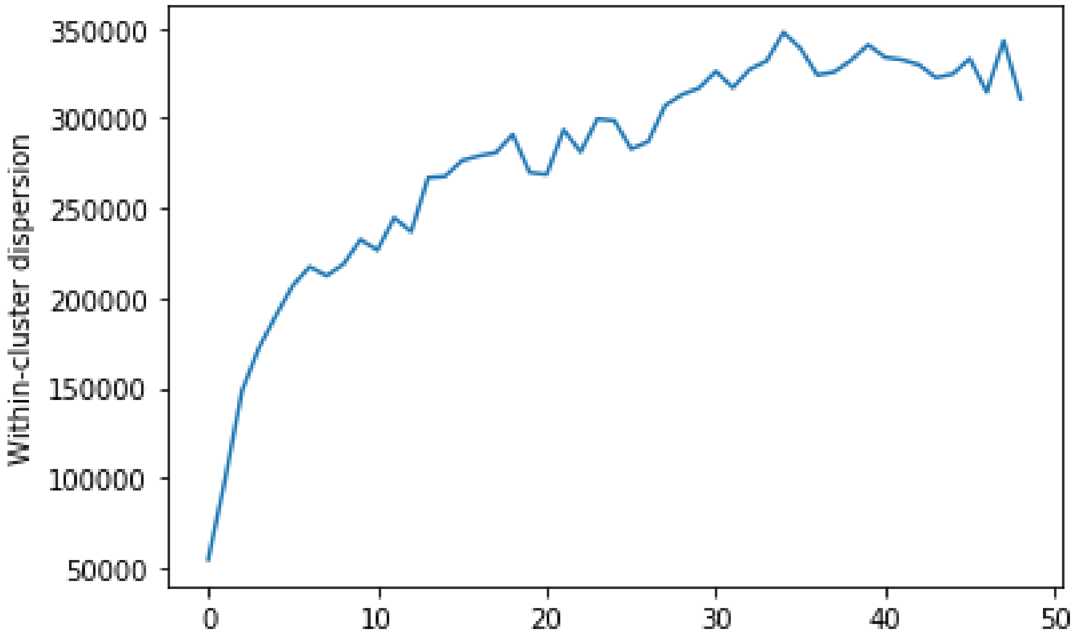
Silhouette score

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.

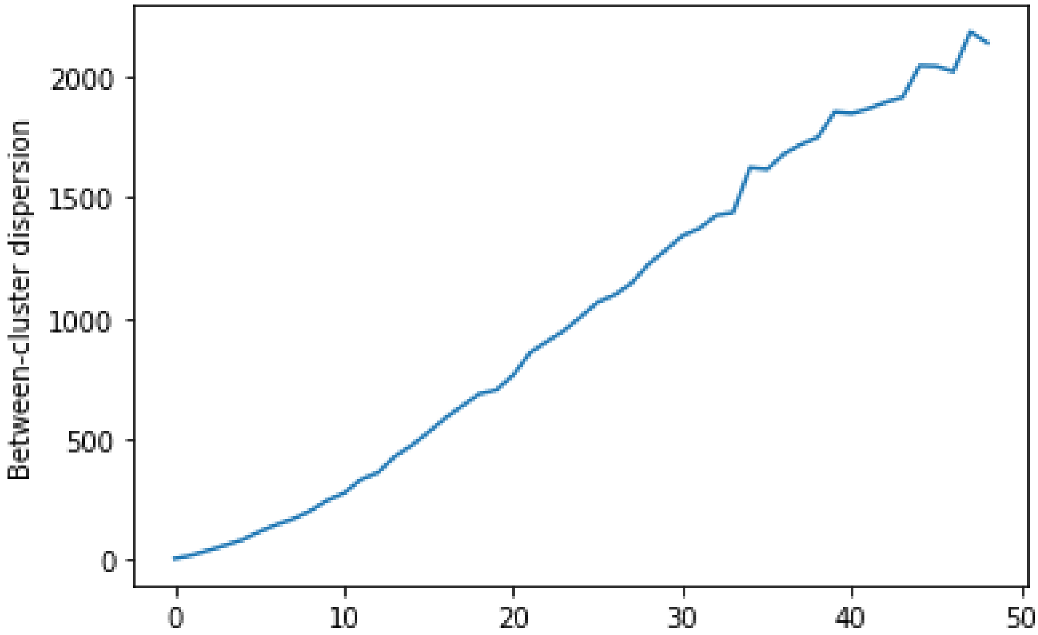


Within-cluster and between cluster dispersion

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



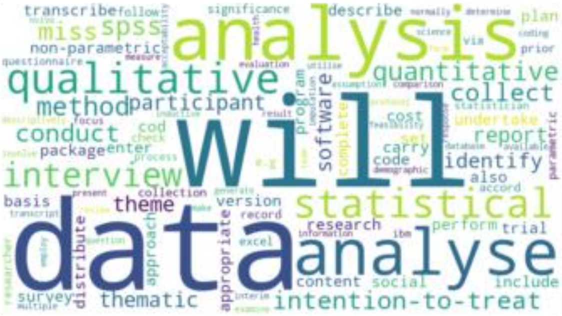
Between-cluster dispersion measures the sum of distances squared between clusters. It should be higher.



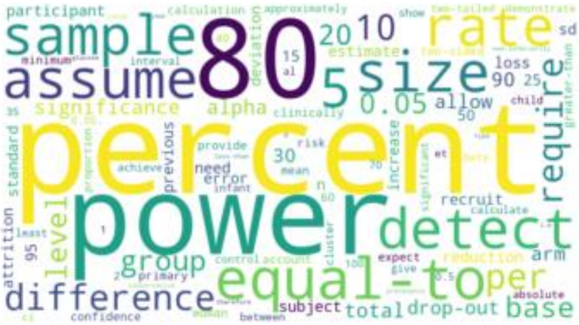
In our results, the clustering maximizes the between-cluster dispersion but does not minimizes the within-cluster dispersion.

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

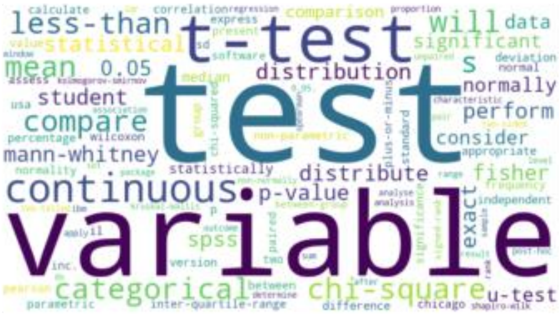
Topics (10)



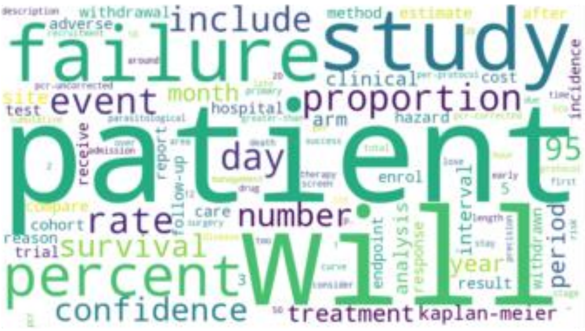
Topic1



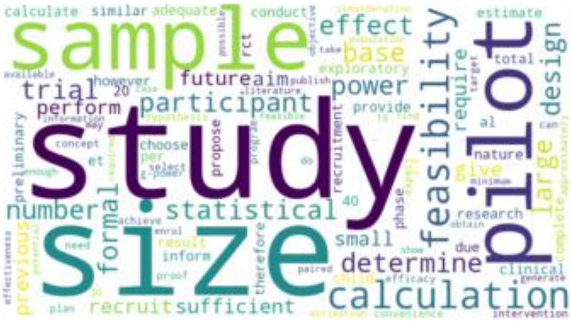
Topic2



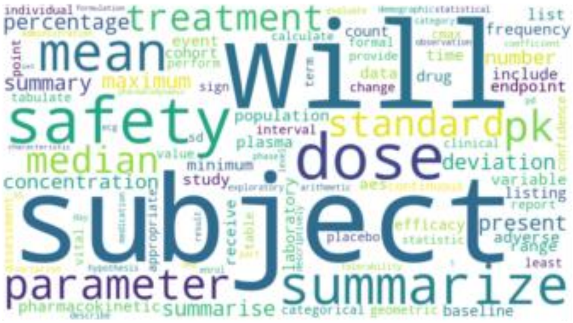
Topic3



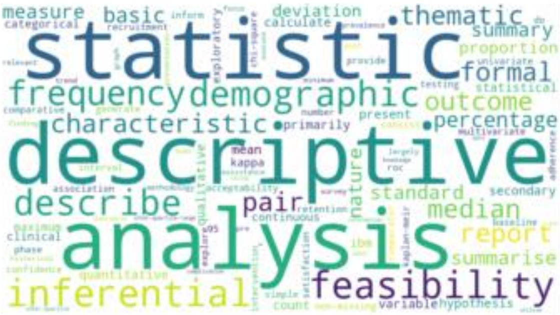
Topic4



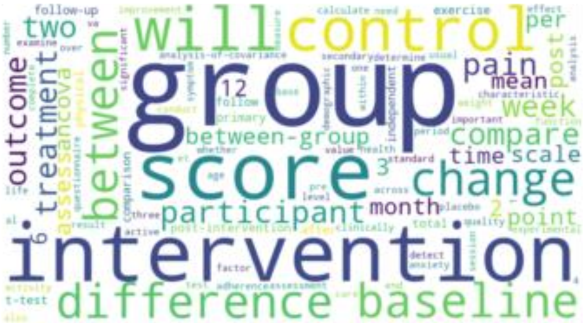
Topic5



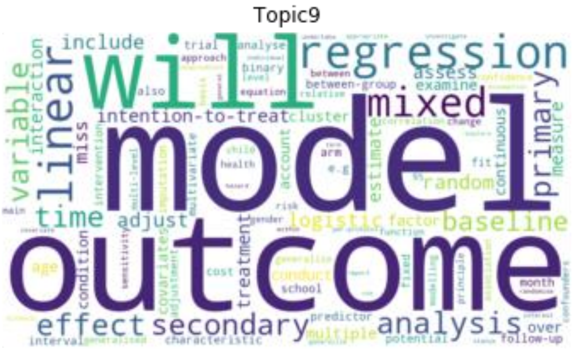
Topic6

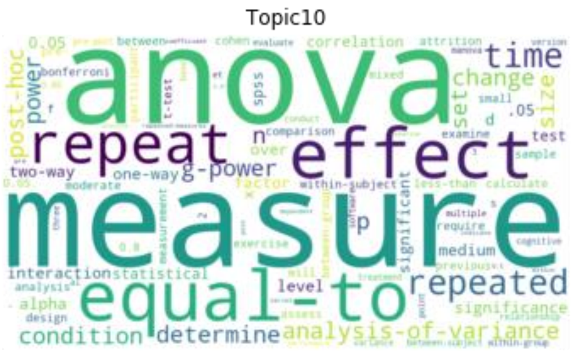


Topic7



Topic8


Topic9

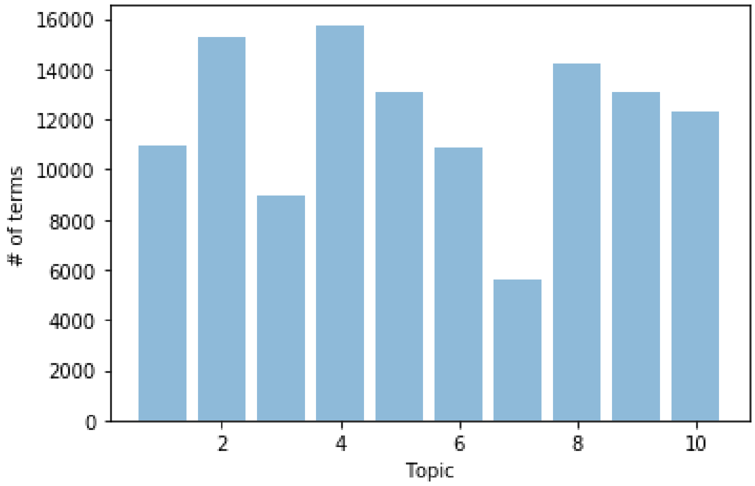

Topic10

Number of documents in each cluster

```
2       1834
3       1277
9       1195
8       1020
4        909
1        854
10       852
5        834
6        524
7        333
```

Size of the topics

	#terms	topic
0	10973	1
1	15259	2
2	8967	3
3	15765	4
4	13077	5
5	10845	6
6	5656	7
7	14253	8
8	13110	9
9	12346	10



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

```
Total number of words: 35881
Median number of words by topic: 12711.5
IQR of words by topic: 2772.5
```

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

```
Topic 1
                    DOI  clusterID     value
1   ACTRN12617001072303          1  0.127538
2   ACTRN12618001741279          1  0.126456
3   ACTRN12620000095965          1  0.113941
4   ACTRN12615000502538          1  0.112210
5   ACTRN12616001734459          1  0.111140
6   ACTRN12620001054909          1  0.109982
7   ACTRN12617001062314          1  0.109206
8   ACTRN12620000897965          1  0.107033
9   ACTRN12616000811404          1  0.104623
10  ACTRN12620000326998          1  0.103678
```

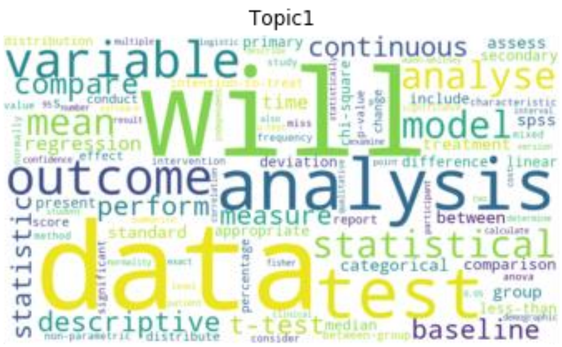
```
Topic 2
                     DOI  clusterID     value
1    ACTRN12615000345583          2  0.143446
2    ACTRN12613000536763          2  0.141373
3    ACTRN12616000789460          2  0.138741
4   ACTRN12616000890437p          2  0.137840
5    ACTRN12620001369910          2  0.134507
6    ACTRN12616000405415          2  0.134331
7    ACTRN12613000655741          2  0.133645
8    ACTRN12617000884303          2  0.131010
9   ACTRN12618001686291p          2  0.126321
10   ACTRN12619001400156          2  0.125444
```

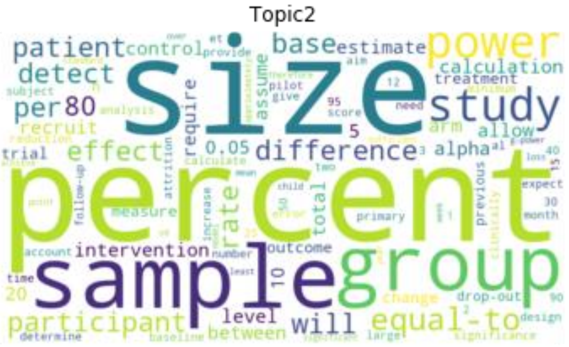
```
Topic 3
                     DOI  clusterID     value
1    ACTRN12618001067268          3  0.197486
2    ACTRN12618000271202          3  0.193882
3    ACTRN12618000158268          3  0.191970
4   ACTRN12619001763134p          3  0.177674
5    ACTRN12617000096358          3  0.177140
6    ACTRN12616001425482          3  0.174889
7    ACTRN12620000086965          3  0.173373
8    ACTRN12618001767291          3  0.169734
9    ACTRN12613000392763          3  0.169223
10   ACTRN12619001105134          3  0.168086
```

```
Topic 4
                    DOI  clusterID     value
1   ACTRN12616001451493          4  0.180063
2   ACTRN12616000526471          4  0.178407
3   ACTRN12618001223224          4  0.172474
4   ACTRN12616001423404          4  0.171042
5   ACTRN12616001005448          4  0.171037
6   ACTRN12616000553471          4  0.170360
7   ACTRN12617001055392          4  0.170176
8   ACTRN12616001478404          4  0.170097
9   ACTRN12617000456358          4  0.169716
10  ACTRN12619000859189          4  0.169584


Topic 5
                     DOI  clusterID     value
1    ACTRN12614000396628          5  0.191177
2    ACTRN12618000853246          5  0.191177
3    ACTRN12620001365954          5  0.173641
4    ACTRN12614000283673          5  0.164784
5    ACTRN12614000349640          5  0.153910
6    ACTRN12613001364763          5  0.145901
7    ACTRN12618000008224          5  0.144136
8   ACTRN12617001533381p          5  0.141107
9    ACTRN12617001470381          5  0.138162
10   ACTRN12617001482358          5  0.138162


Topic 6
                    DOI  clusterID     value
1   ACTRN12613000991718          6  0.189175
2   ACTRN12615000606583          6  0.189175
3   ACTRN12616000104459          6  0.180722
4   ACTRN12616001040459          6  0.179802
5   ACTRN12617001471370          6  0.179573
6   ACTRN12617001178336          6  0.175392
7   ACTRN12620000725965          6  0.175140
8   ACTRN12616001437459          6  0.173411
9   ACTRN12615001025527          6  0.173181
10  ACTRN12619001178134          6  0.168932


Topic 7
                     DOI  clusterID     value
1    ACTRN12619001572156          7  0.334002
2    ACTRN12614001207606          7  0.334002
3   ACTRN12620000745943p          7  0.334002
4    ACTRN12613000405718          7  0.334002
5    ACTRN12619001407189          7  0.334002
6    ACTRN12616001615471          7  0.334002
7    ACTRN12618001720202          7  0.334002
8    ACTRN12615000304538          7  0.334002
9   ACTRN12618000284268p          7  0.334002
10   ACTRN12619001524189          7  0.334002
```

```
Topic 8
                     DOI  clusterID     value
1   ACTRN12616000119493          8  0.163654
2   ACTRN12613000275763          8  0.117682
3   ACTRN12621000002886          8  0.111242
4   ACTRN12613000296730          8  0.110135
5   ACTRN12613000844741          8  0.108817
6   ACTRN12617000033347          8  0.103795
7   ACTRN12616000607471          8  0.101170
8   ACTRN12620000953932          8  0.100537
9   ACTRN12617000735358          8  0.099751
10  ACTRN12617000246381          8  0.099050


Topic 9
                     DOI  clusterID     value
1   ACTRN12619001074189          9  0.159672
2   ACTRN12616000246482          9  0.152831
3   ACTRN12619000373178          9  0.150919
4   ACTRN12615001032549          9  0.146555
5   ACTRN12618000944235          9  0.137932
6   ACTRN12620001281987          9  0.133287
7   ACTRN12615000803594          9  0.131356
8   ACTRN12617000364370          9  0.131080
9   ACTRN12620000669998          9  0.128340
10  ACTRN12613000691741          9  0.128106


Topic 10
                     DOI  clusterID     value
1   ACTRN12613000286741         10  0.204550
2   ACTRN12618000710224         10  0.204550
3   ACTRN12617000007336         10  0.204550
4   ACTRN12619001007123         10  0.204550
5   ACTRN12617001623381         10  0.204550
6   ACTRN12617000108314         10  0.176315
7   ACTRN12616001486415         10  0.165148
8   ACTRN12617001432303         10  0.155135
9   ACTRN12616001528448         10  0.154091
10  ACTRN12613000499785         10  0.152290
```

Topics (2)

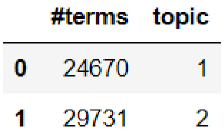




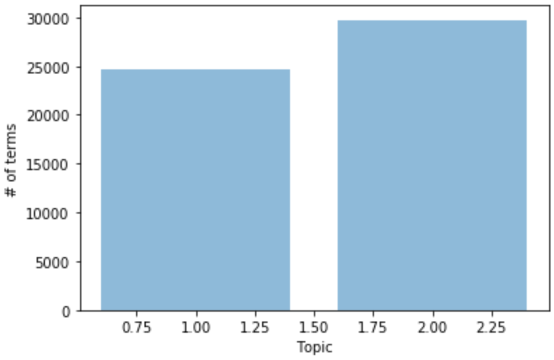
Number of documents in each cluster

```
2     4979
1     4653
```

Size of the topics



	#terms	topic
0	24670	1
1	29731	2



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

```
Total number of words: 35881
Median number of words by topic: 27200.5
IQR of words by topic: 0.0
```

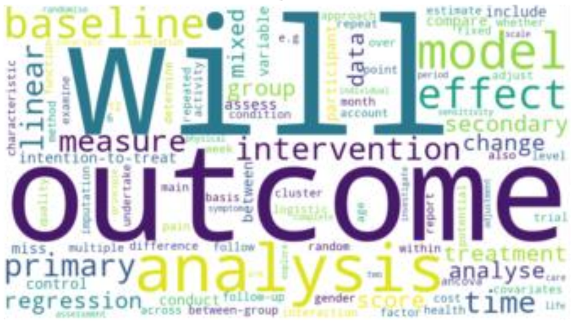
A few DOIs for papers that were a strong match in each topic

```
Topic 1
                         DOI  clusterID     value
1     ACTRN12615000502538          1  0.122607
2     ACTRN12614000517673          1  0.118921
3    ACTRN12620001278921p          1  0.116262
4     ACTRN12618001419257          1  0.115750
5     ACTRN12620000086965          1  0.113821
6     ACTRN12618001363279          1  0.113589
7     ACTRN12619000113156          1  0.112714
8     ACTRN12616000721404          1  0.112215
9     ACTRN12620001040954          1  0.111489
10    ACTRN12619000514101          1  0.111136
```

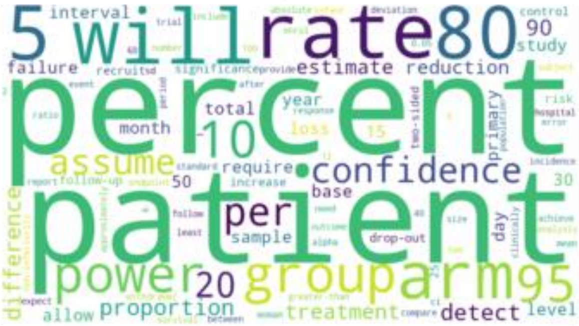
```
Topic 2
                     DOI  clusterID     value
1    ACTRN12617000561381          2  0.127578
2    ACTRN12616001122448          2  0.124067
3    ACTRN12616001367437          2  0.121665
4    ACTRN12616000789460          2  0.120881
5    ACTRN12617000884303          2  0.119740
6   ACTRN12619000289112p          2  0.119547
7    ACTRN12617000706370          2  0.118432
8    ACTRN12615000230550          2  0.116450
9    ACTRN12615000012572          2  0.116427
10   ACTRN12617001546347          2  0.114661
```

Topics (5)

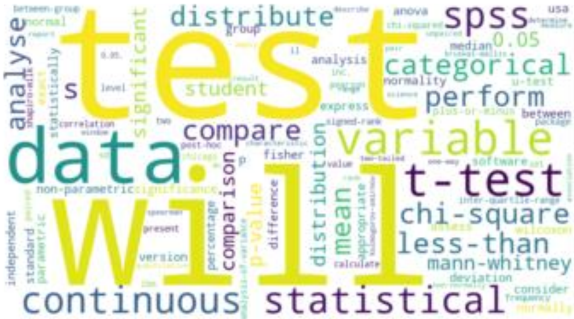
Topic1



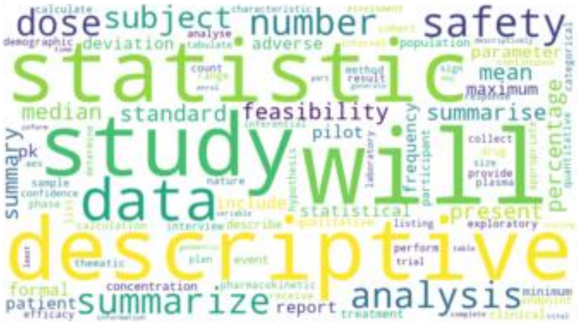
Topic2



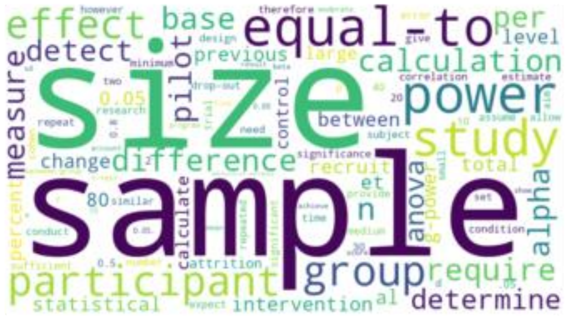
Topic3



Topic4



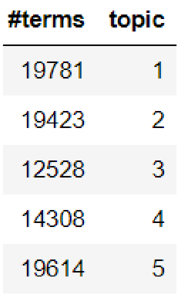
Topic5



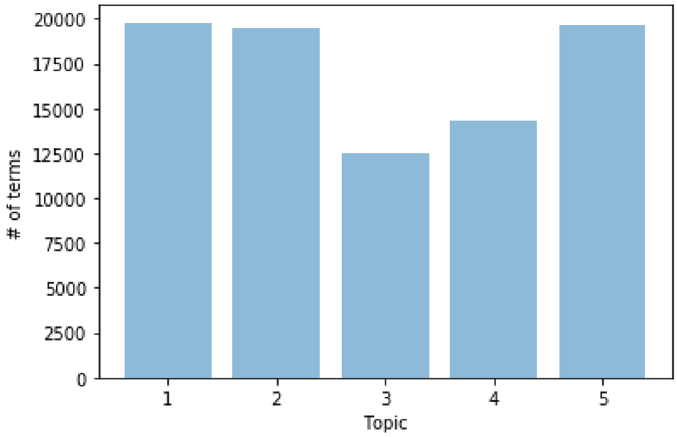
Number of documents in each cluster

2	2147
1	2146
5	1944
3	1830
4	1565

Size of the topics



#terms	topic
19781	1
19423	2
12528	3
14308	4
19614	5



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

```
Total number of words: 35881
Median number of words by topic: 19423.0
IQR of words by topic: 5306.0
```

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

```
Topic 1
                         DOI  clusterID     value
1     ACTRN12619001716156          1  0.130403
2    ACTRN12618001628235p          1  0.122605
3     ACTRN12618001751268          1  0.122543
4     ACTRN12619000373178          1  0.121323
5     ACTRN12620000012976          1  0.120270
6     ACTRN12619001112156          1  0.119396
7     ACTRN12619001238167          1  0.118076
8     ACTRN12617000364370          1  0.117785
9     ACTRN12617000480381          1  0.116247
10    ACTRN12615000490572          1  0.114767
```

```
Topic 2
                         DOI  clusterID     value
1     ACTRN12616000789460          2  0.160160
2     ACTRN12619000496112          2  0.141134
3    ACTRN12618001686291p          2  0.139961
4    ACTRN12616000890437p          2  0.139930
5     ACTRN12615001379505          2  0.138362
6     ACTRN12616000338460          2  0.135227
7     ACTRN12618000790246          2  0.133682
8     ACTRN12616000405415          2  0.132684
9     ACTRN12615000345583          2  0.132154
10    ACTRN12615001333505          2  0.131747
```

```
Topic 3
                    DOI  clusterID     value
1   ACTRN12618000271202          3  0.207470
2   ACTRN12618000158268          3  0.206897
3   ACTRN12613000392763          3  0.189801
4   ACTRN12618001067268          3  0.186927
5   ACTRN12618001767291          3  0.185273
6   ACTRN12614001149651          3  0.183062
7   ACTRN12617000771358          3  0.178789
8   ACTRN12616001425482          3  0.176994
9   ACTRN12616000730404          3  0.176673
10  ACTRN12619001105134          3  0.175424
```

```
Topic 4
                    DOI  clusterID     value
1   ACTRN12615000606583          4  0.193738
2   ACTRN12613000991718          4  0.193738
3   ACTRN12614001299695          4  0.191839
4   ACTRN12614000725662          4  0.191839
5   ACTRN12616000676415          4  0.187882
6   ACTRN12617001178336          4  0.185571
7   ACTRN12614001207606          4  0.181331
8   ACTRN12615000304538          4  0.181331
9   ACTRN12618001720202          4  0.181331
10  ACTRN12617001621303          4  0.181331
```

```
Topic 5
                    DOI  clusterID     value
1   ACTRN12614001010684          5  0.131373
2   ACTRN12617000876392          5  0.127356
3   ACTRN12615000878572          5  0.127257
4   ACTRN12614000830695          5  0.124506
5   ACTRN12615000009516          5  0.116323
6   ACTRN12613000978763          5  0.115061
7   ACTRN12613001364763          5  0.113473
8   ACTRN12616000803493          5  0.113168
9   ACTRN12618001280291          5  0.111153
10  ACTRN12617001603303          5  0.110520
```