# COVID19 Scholarly Article Analysis

## Innovation

The scientific community has responded to the COVID-19 pandemic with massive research efforts, including research into vaccines, medication, mitigation of the virus' spread, improved testing, and so on. This has resulted in a surge of publication of scientific articles on the topic of the COVID-19 disease. Researchers aiming to review literature on COVID-19 can benefit from analysis that reveals hidden relations between articles or groupings of articles, as they could use such analysis to help them discover articles relating to their area of research. We performed natural language processing and corpus analysis on an open dataset of scientific articles on the topic of COVID-19. This analysis may reveal underlying relations and associations between these scientific articles, which researchers can use to arrive at novel or faster insight in their efforts to counter the pandemic.

## Source Data

The source data for this analysis is the COVID-19 Open Research Dataset compiled by leading researchers in cooperation with the U.S. Government. The dataset contains over 135,000 scholarly articles relating to the topic of COVID-19 and SARS-CoV-2. It is available here:

https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge (https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge)

## Preprocessing

Since the dataset is composed of raw text json files, the data needed to be preprocessed into a numerical form before it could be analyzed. The abstract and body text sections were extracted from each json file, split into word lists, converted to all lower-case, and scrubbed of punctuation, non-alphabetic words, and stop words. Stem words were taken from the resulting word lists and stored in csv format with a corresponding file ID. At the same time, vocabulary lists were collected for each text. The lists belonging to the same sources (such as arxiv, noncomm_use_subset, custom_license, biorxiv_medrxiiv) were combined in order to create a document-term count matrix.

Below is the code for data preprocessing and vocabulary collection:

In [ ]:

```
# extract the abstract and body-text text from all json files in a folder
# clean up before analysis
# arxiv file as an example

import os
import json
import numpy
import string
import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer

# some functions setting up
table = str.maketrans('', '' ,string.punctuation)
porter = PorterStemmer()
d = numpy.empty(1)

# read every json file in the folder
path_to_json = 'CORD-19-research-challenge/arxiv/arxiv/pdf_json/'
json_files = [jf for jf in os.listdir(path_to_json) if jf.endswith('.json')]
i = 0
for index, js in enumerate(json_files):
    with open(os.path.join(path_to_json, js)) as f:
        data = json.load(f)

    file_name = data["paper_id"]
    cleaned_file = numpy.empty(1)
    i = i+1
    print(i)
    print(file_name)

    # get the text from abstract part
    for p in data["abstract"]:
        # split the sentences into words
        words = word_tokenize(p["text"])
        # lower case
        words = [w.lower() for w in words]
        # remove punctuation from each word
        words = [w.translate(table) for w in words]
        # remove the words not in alphabetic
        words = [w for w in words if w.encode('UTF-8').isalpha()]
        # filter out stop words
        stop_words = set(stopwords.words('english'))
        words = [w for w in words if not w in stop_words]
        # transfer to stem words
        words = [porter.stem(w) for w in words]
        # print("body_text: ", words)
        cleaned_file = numpy.append(cleaned_file, words)
        d = numpy.append(d, words)
        d = numpy.unique(d)

    # get the text from body_text part
    for p in data["body_text"]:
        # split the sentences into words
        words = word_tokenize(p["text"])
        # lower case
        words = [w.lower() for w in words]
        # remove punctuation from each word
        words = [w.translate(table) for w in words]
        # remove the words not in alphabetic
        words = [w for w in words if w.encode('UTF-8').isalpha()]
        # filter out stop words
        stop_words = set(stopwords.words('english'))
        words = [w for w in words if not w in stop_words]
        # transfer to stem words
        words = [porter.stem(w) for w in words]
        # print("body_text: ", words)
        cleaned_file = numpy.append(cleaned_file, words)
        d = numpy.append(d, words)
        d = numpy.unique(d)
    print("\n")
    cleaned_file = cleaned_file[1:]
    file_name = 'cleaned_file/arxiv/arxiv_pdf'+file_name +'.csv'
    numpy.savetxt(file_name, cleaned_file, fmt = "%s", delimiter = ';')

d = d[1:]
numpy.savetxt("arxiv_vocabulary.csv", d, fmt = "%s", delimiter = ';')
```

In [ ]:

```
# the amount of files is large, we have splited into several folders, run on dif
# and now we need to combine everything together
# combine all cvs files to obtain a vocabulary table
# noncomm_use_subset_pmc as an example

import os
import numpy
import string
import csv

# read every json file in the folder

dir_path = 'cleaned_file/noncomm_use_subset/noncomm_use_subset_pmc/noncomm_use_su
data = []
for i in range(0,3):
    subdir_path = dir_path + str(i)
    print(subdir_path)
    file = [f for f in os.listdir(subdir_path) if f.endswith('.csv')]
    for r in file:
        print(r)
        path = subdir_path+'/'
        path = path + r
        with open(path, 'r') as file:
            fileReader = csv.reader(file, delimiter = ';')
            data1 = numpy.array(list(fileReader)).astype(str)

        data = numpy.append(data, data1)
        data = numpy.unique(data)

numpy.savetxt("noncomm_use_subset_pmc_vocabulary.csv", data, fmt = "%s", delimit
```

## Kmeans Clustering

Using the preprocessed data, a document-term count matrix (DT matrix) was constructed, in which the rows of the DT matrix represent the vocabulary list, and the columns represent the text file list, for the files from each source. In order to speed up the process, the DT matrix was translated into a sparse matrix format, and then the sparse matrix was transformed into a term frequency-inverse document frequency matrix (TF-IDF matrix). For the clustering process, K-means was applied to the TF-IDF matrix. To find an optimal K value, the K-means algorithm was run several times, with K ranging from 1 to 19. The models were saved for further analysis.

In [ ]:

```
import re
import os
import pandas
import json
import numpy
import string
import nltk
import csv
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from nltk.stem.snowball import SnowballStemmer
from scipy.sparse import csr_matrix
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.cluster import KMeans
from sklearn.externals import joblib

# construct a dictionary vector
# path may need to be changed
path = 'covid19_schorlar_analysis/vocabulary.csv'
with open(path, 'r') as file:
    fileReader = csv.reader(file, delimiter = ';')
    dictionary = numpy.array(list(fileReader)).astype(str)

# contruct a word-count matrix
# get a list of all files in the certain folder
# path may need to be changed
path = 'covid19_schorlar_analysis/cleaned_file/biorxiv_medrxiv/biorxiv_medrxiv_p
file_list = [f for f in os.listdir(path) if f.endswith('.csv')]
# i: file index
i = 0
# f_list: file index and indexed file
f_list = numpy.zeros(2)
# matrix: word count matrix
matrix = numpy.zeros(len(dictionary))
# read every file to complete word counting
for index, f in enumerate(file_list):
    temp = numpy.array([i, f])
    print(temp)
    f_list = numpy.r_[f_list, temp.T]
    with open(os.path.join(path, f),'r') as file:
        fileReader = csv.reader(file, delimiter = ';')
        data = numpy.array(list(fileReader)).astype(str)

    count = numpy.zeros(len(dictionary))
    for word in data:
        j = numpy.where(dictionary == word)
        count[j[0]] += 1

    matrix = numpy.c_[matrix, count]
    i += 1

print(matrix)

matrix = matrix[:, 1:]
matrix = matrix.T
```

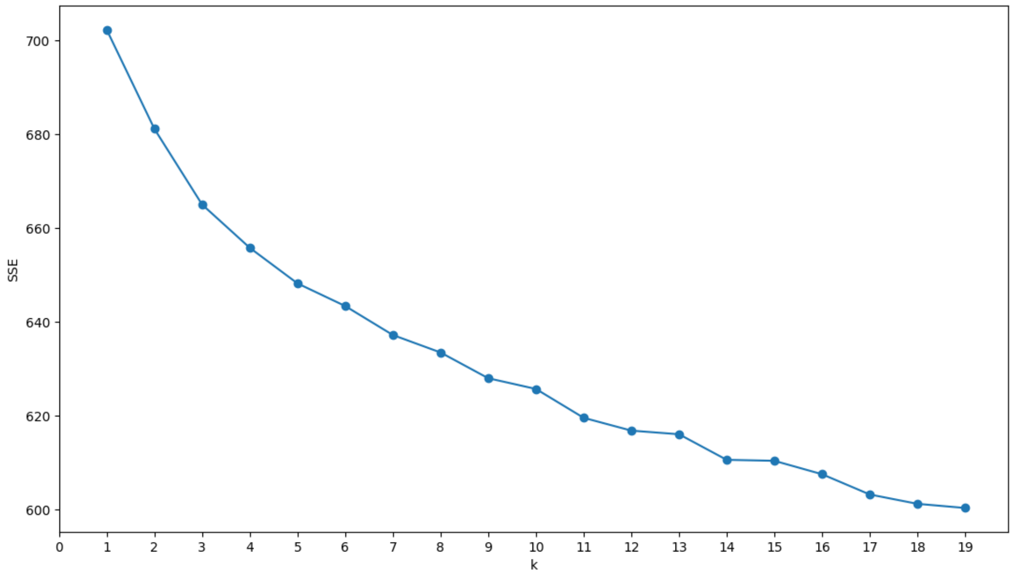
```
f_list = f_list[2:]
l = len(f_list)/2
l = int(l)
f_list = f_list.reshape([l, 2])

# index need to be changed
numpy.savetxt('biorxiv_medrxiv_index.csv', f_list, fmt = '%s', delimiter = ';')

# transfer to sparse matrix
matrix_csr = csr_matrix(matrix)
# transfer to Tfidf matrix
transformer = TfidfTransformer()
matrix_tfidf = transformer.fit_transform(matrix_csr)
# k-means cluster: k in range(1,20)
for k in range(1, 20):
    # set up
    km = KMeans(n_clusters = k)
    # fit model
    km.fit(matrix_tfidf)
    # save model
    model_name = 'k' + str(k) + '.m'
    joblib.dump(km, model_name)
```

# Analysis

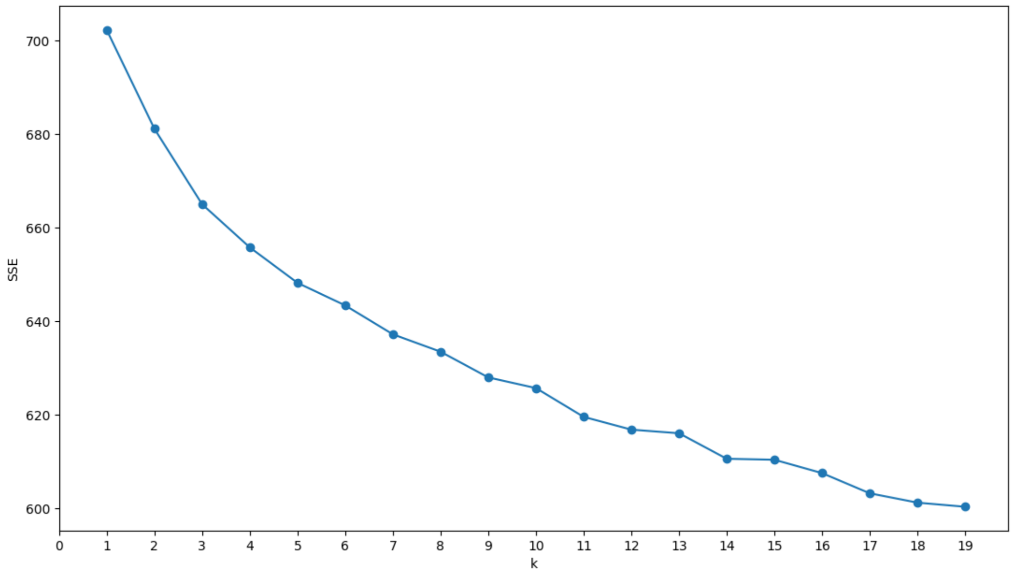
## SSE analysis

To analyze the clustering results of different K values, the sum of squares of error (SSE) was obtained from each K model and plotted into a linear graph.

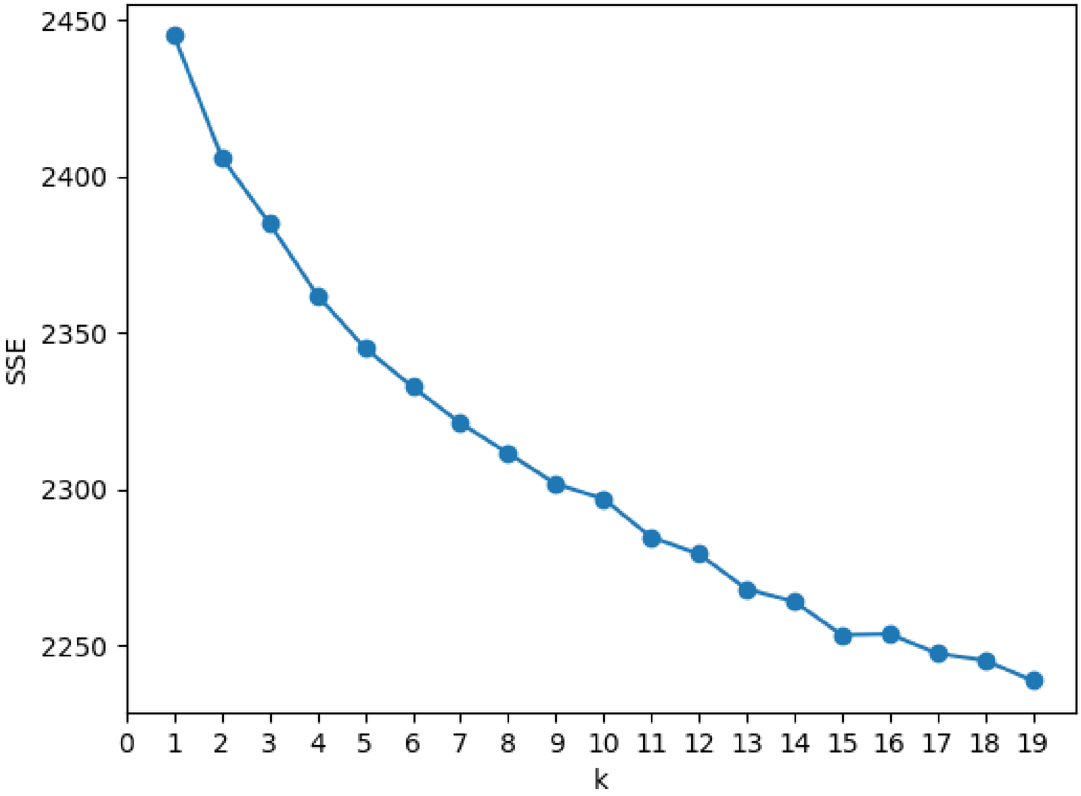
### Arxiv



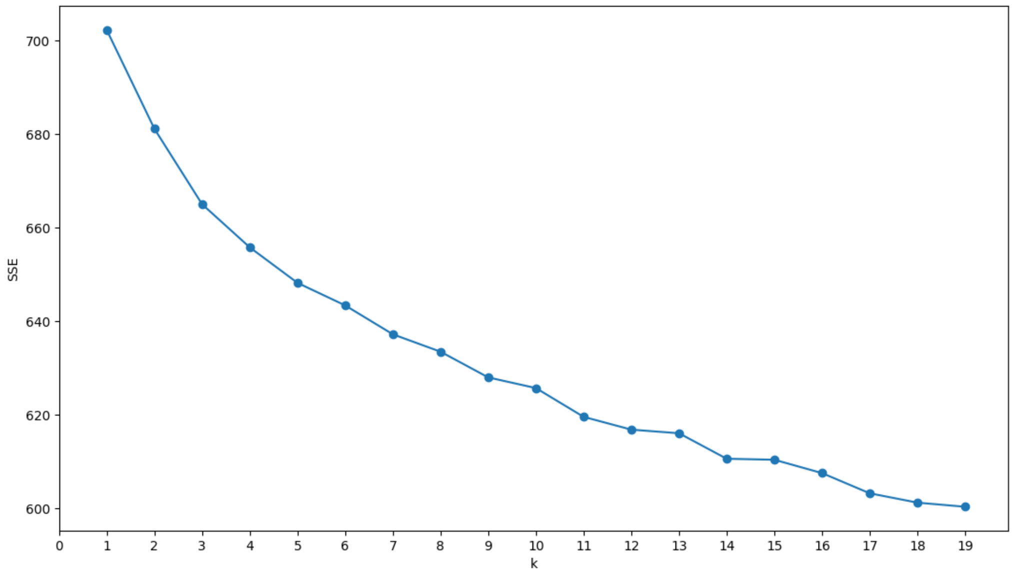
### biorxiv_medrxiv



### noncomm_use_subeset



### comm_use_subset

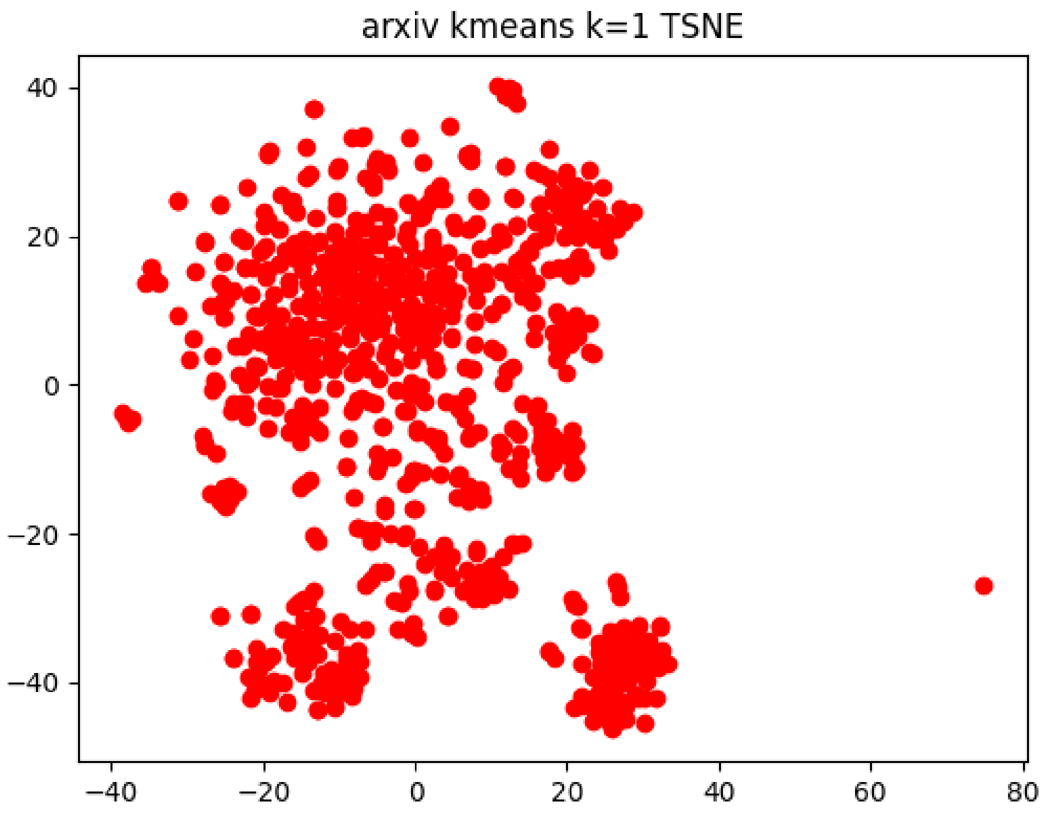


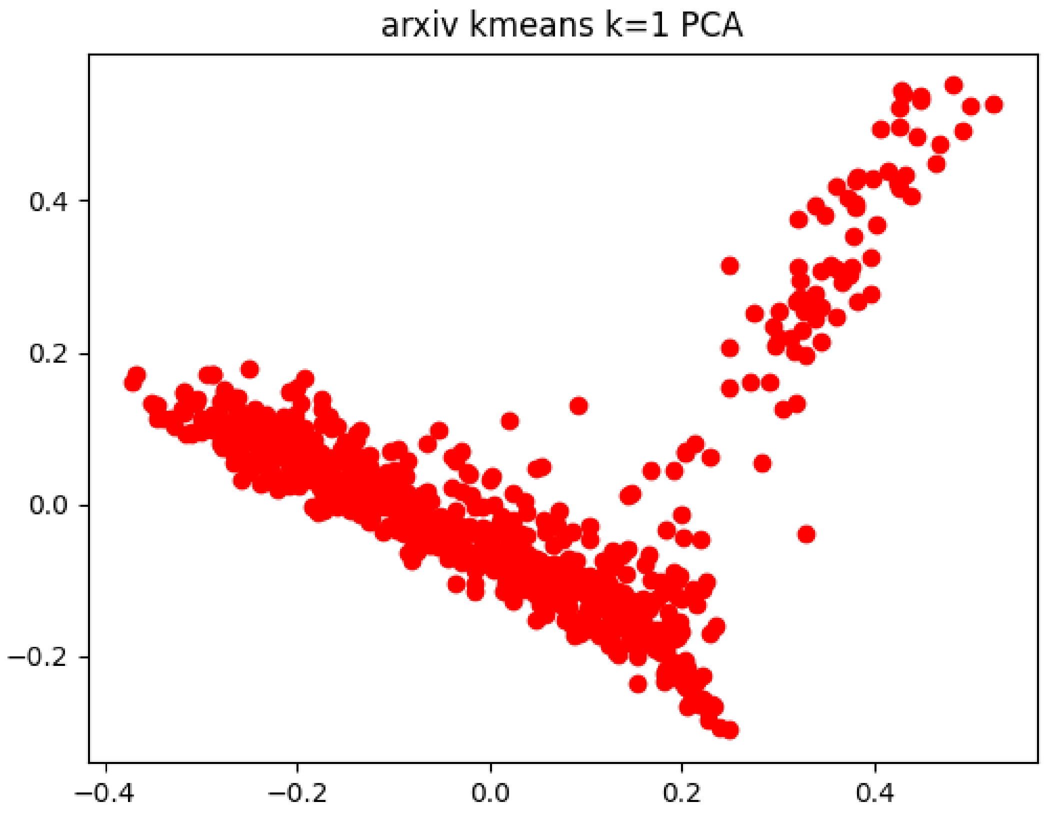
In each graph, the SSE values appear to decrease while the K values increase. The SSE values decrease rapidly at the beginning of the graph, but their decrease slows down with increasing K values. Therefore, SSE will remain stable at some K value, and the K value at this inflection point is optimal. More analysis is necessary to find the inflection point, but it is certain that its K value is smaller than or equal to 8.

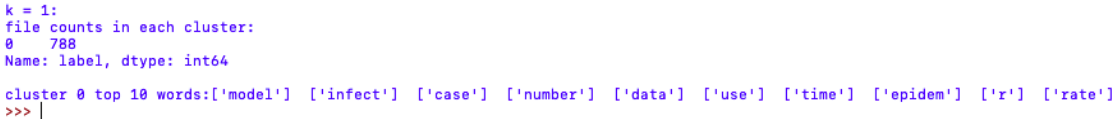
## Visualization (PCA and t-SNE)

Principle Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) were applied to reduce the dimension of the TF-IDF matrix to project the visualization results onto two-dimensional images where K ranges from 1 to 8. The 10 words which have highest frequency in each cluster at various K values are listed.

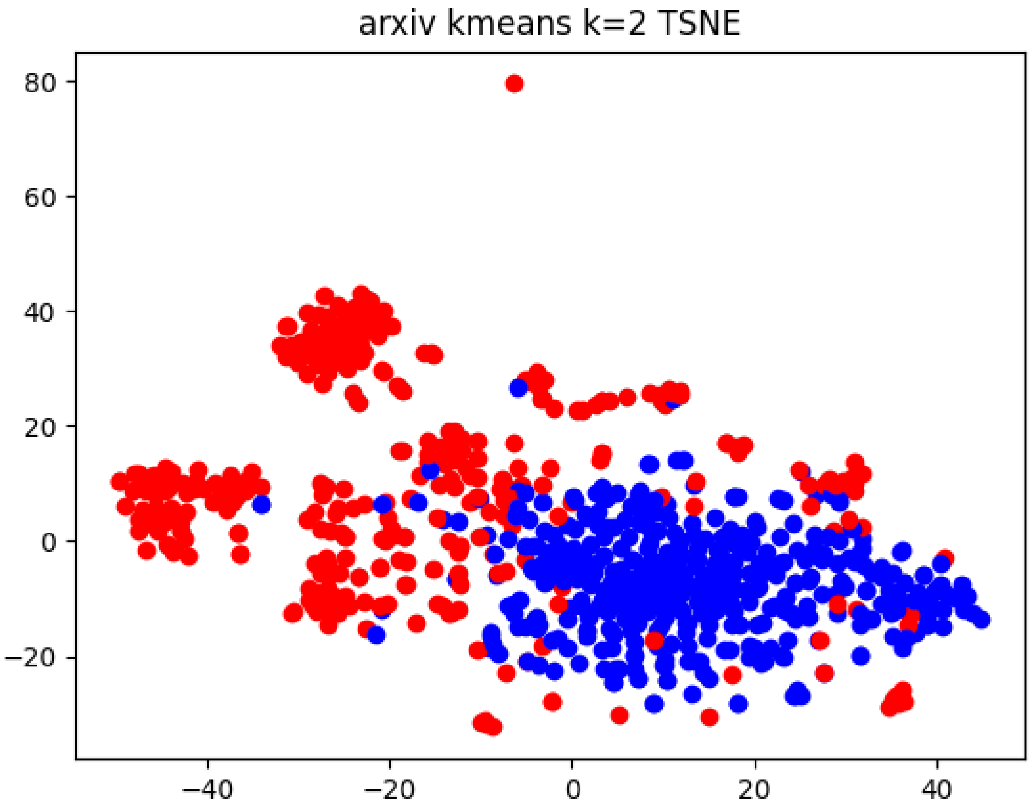
## Arxiv

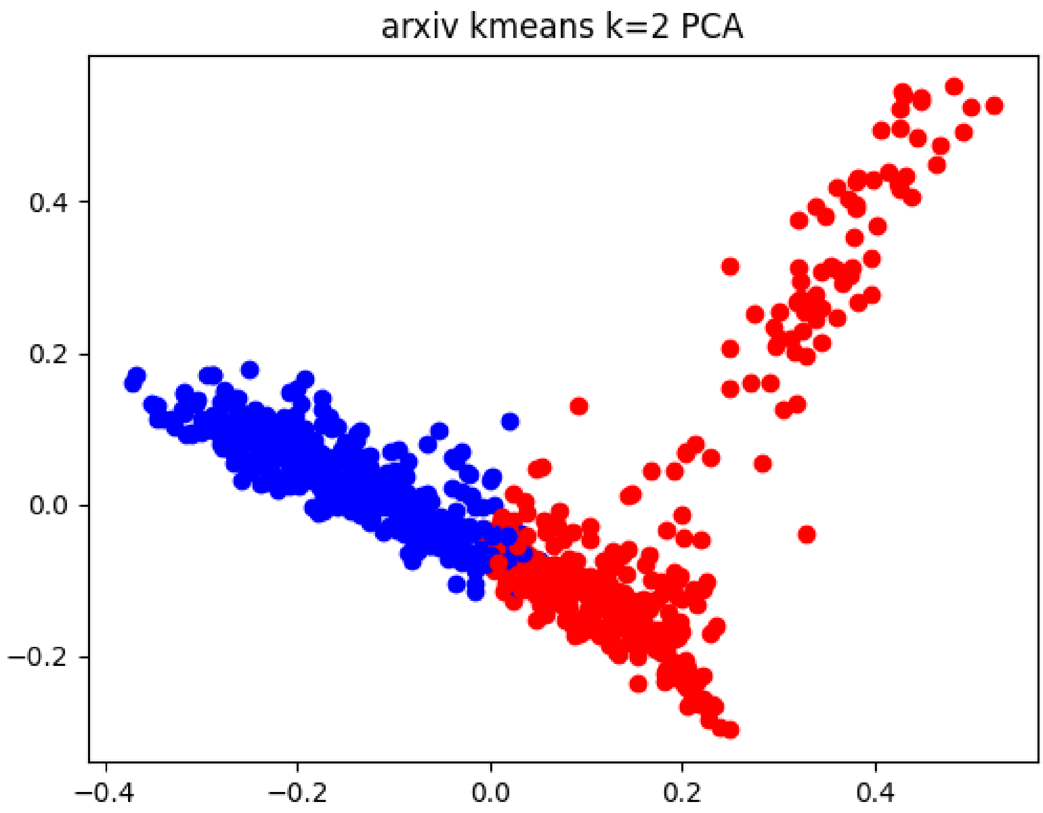

arxiv kmeans k=1 TSNE
40
20
0
−20
−40
−40
−20
0
20
40
60
80

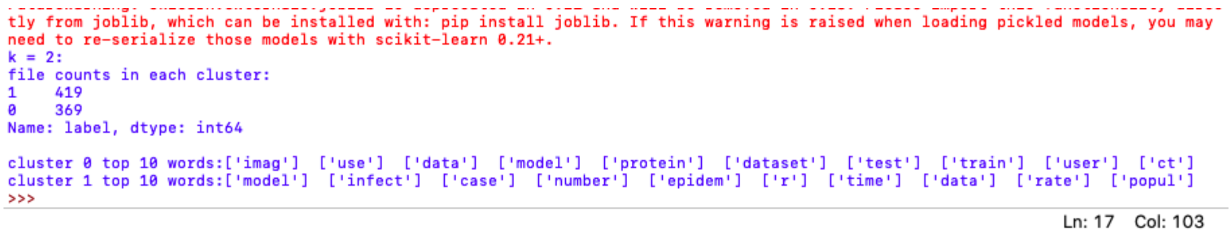

arxiv kmeans k=1 PCA
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
0.4



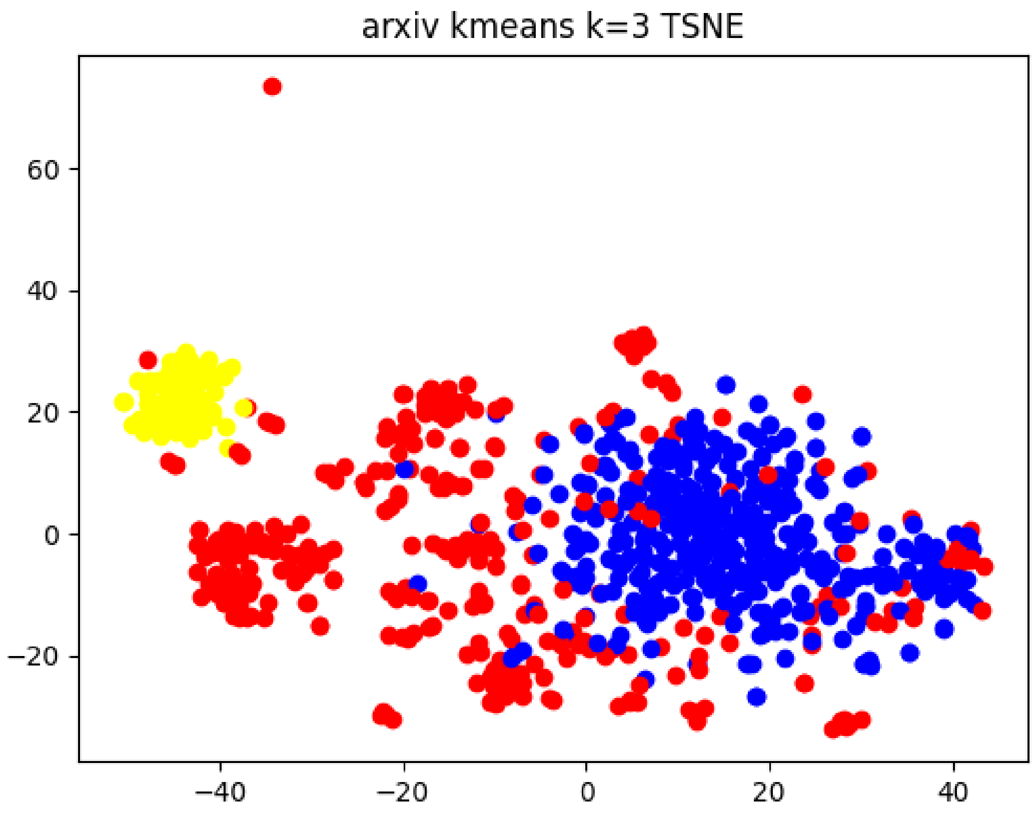
```
k = 1:
file counts in each cluster:
0    788
Name: label, dtype: int64

cluster 0 top 10 words:['model']  ['infect']  ['case']  ['number']  ['data']  ['use']  ['time']  ['epidem']  ['r']  ['rate']
>>> |
```

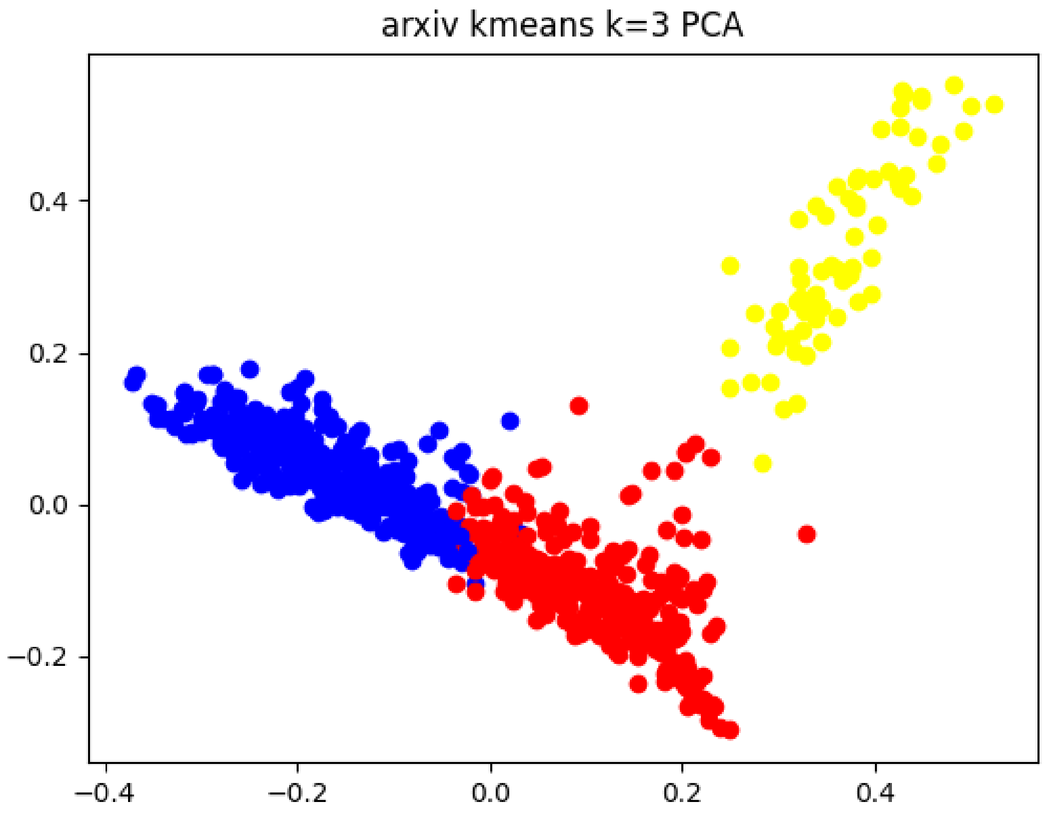


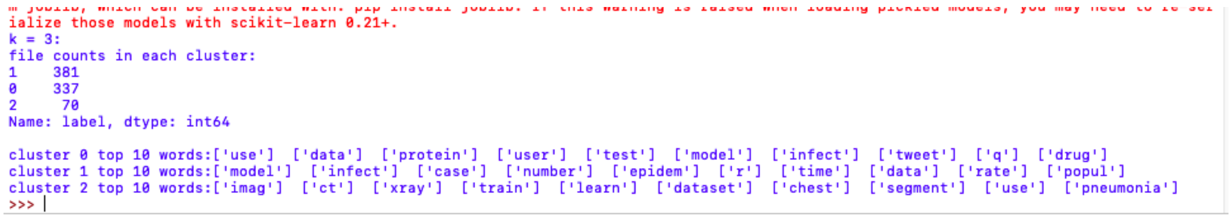




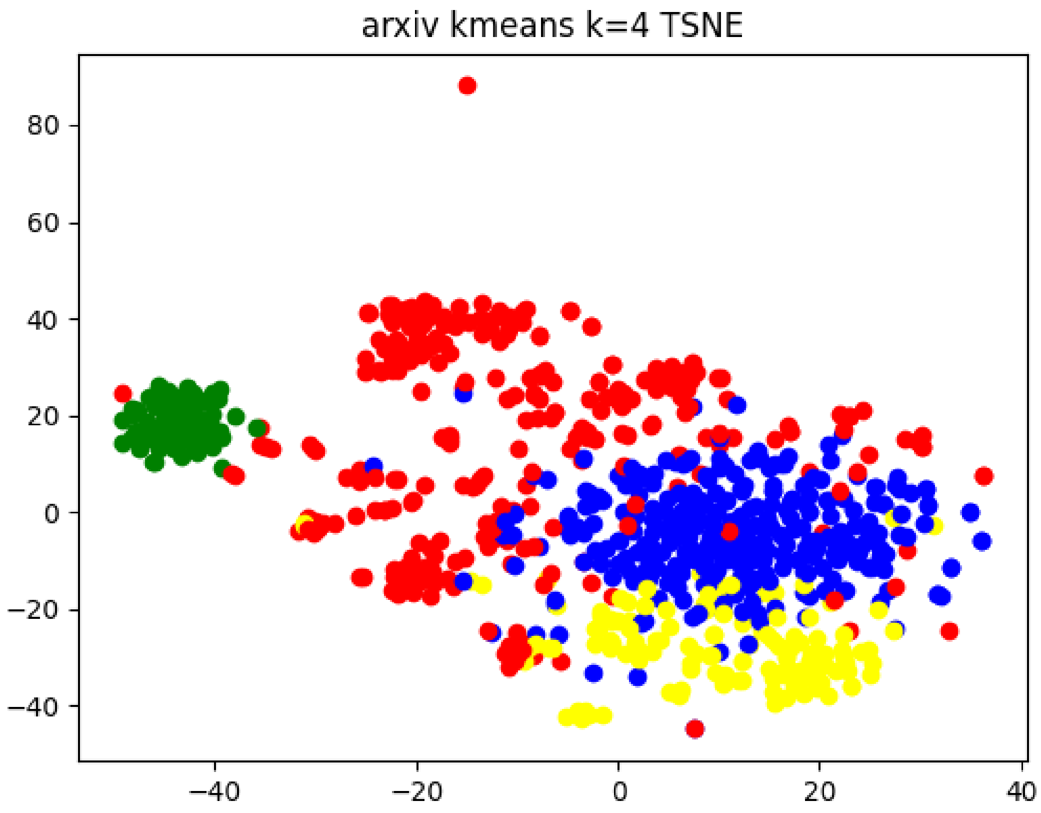
```
tly from joblib, which can be installed with: pip install joblib. If this warning is raised when loading pickled models, you may
need to re-serialize those models with scikit-learn 0.21+.
k = 2:
file counts in each cluster:
1    419
0    369
Name: label, dtype: int64

cluster 0 top 10 words:['imag']  ['use']  ['data']  ['model']  ['protein']  ['dataset']  ['test']  ['train']  ['user']  ['ct']
cluster 1 top 10 words:['model']  ['infect']  ['case']  ['number']  ['epidem']  ['r']  ['time']  ['data']  ['rate']  ['popul']
>>>
```

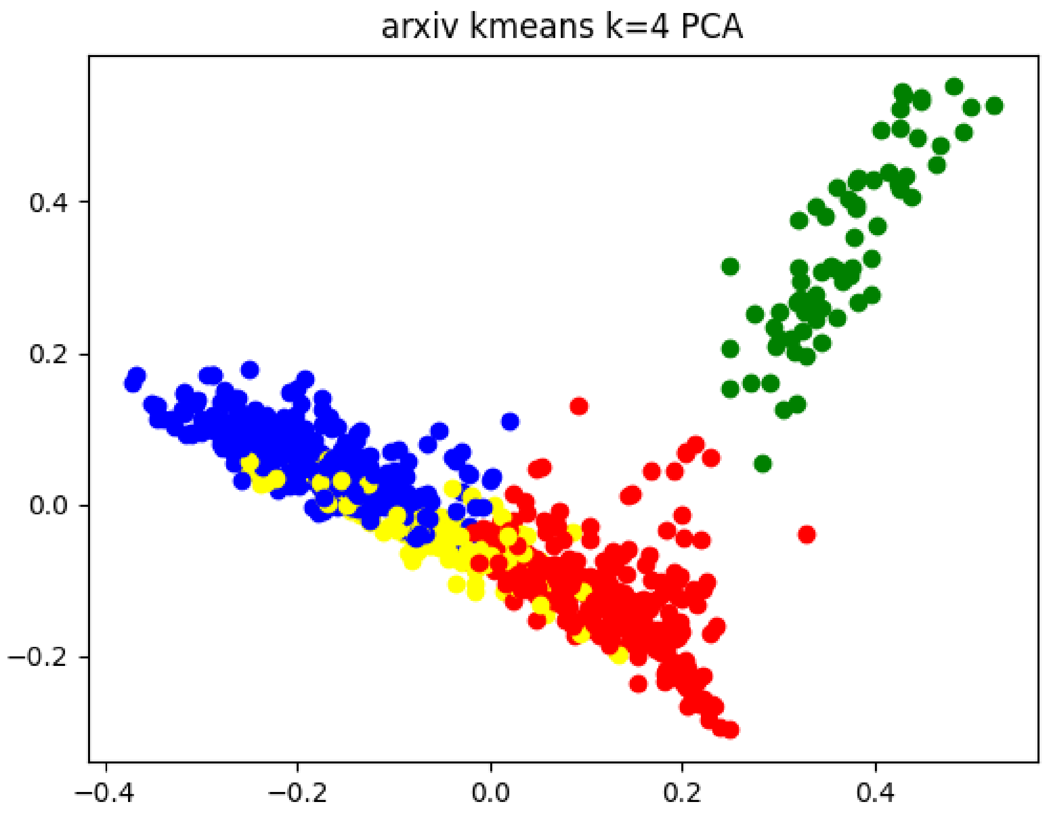

arxiv kmeans k=3 TSNE
60
40
20
0
−20
−40
−20
0
20
40

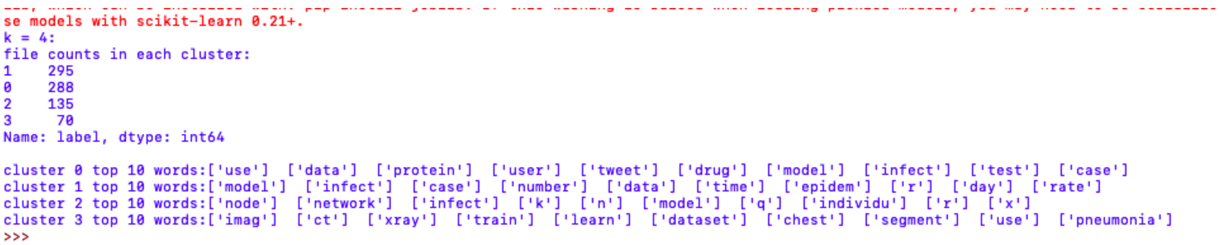

arxiv kmeans k=3 PCA
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
0.4



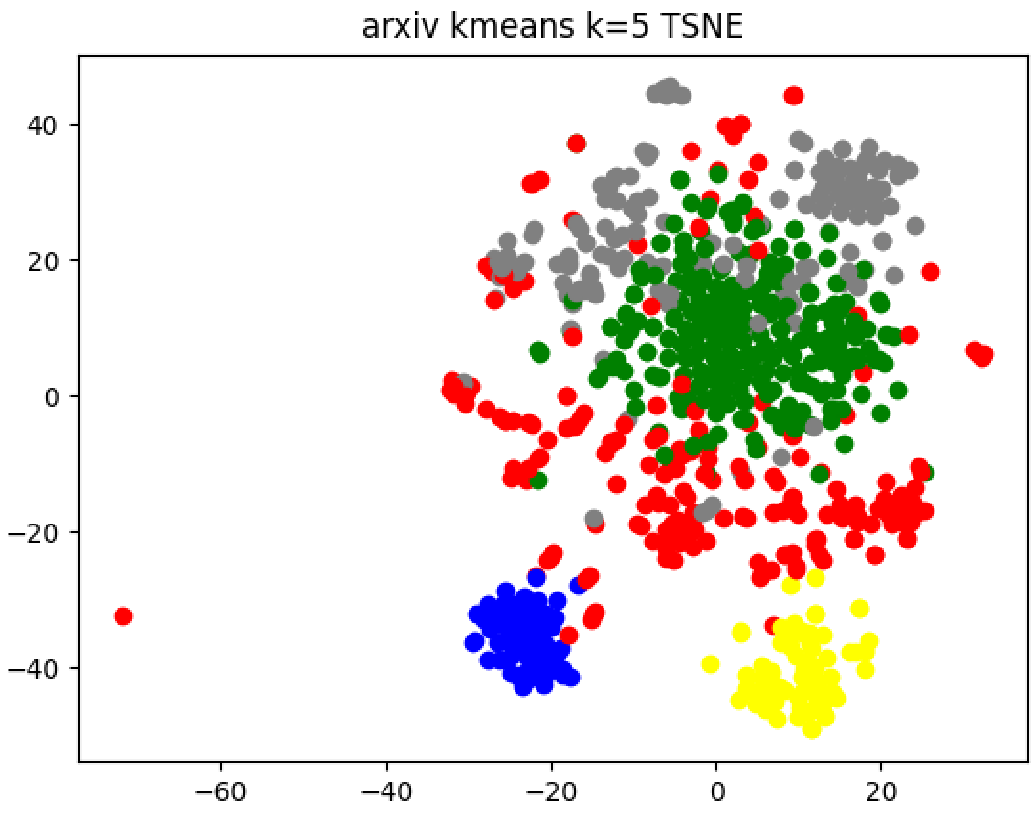
```
m joblib, which can be installed with: pip install joblib. If this warning is raised when loading pickled models, you may need to re-ser
ialize those models with scikit-learn 0.21+.
k = 3:
file counts in each cluster:
1    381
0    337
2     70
Name: label, dtype: int64

cluster 0 top 10 words:['use']  ['data']  ['protein']  ['user']  ['test']  ['model']  ['infect']  ['tweet']  ['q']  ['drug']
cluster 1 top 10 words:['model']  ['infect']  ['case']  ['number']  ['epidem']  ['r']  ['time']  ['data']  ['rate']  ['popul']
cluster 2 top 10 words:['imag']  ['ct']  ['xray']  ['train']  ['learn']  ['dataset']  ['chest']  ['segment']  ['use']  ['pneumonia']
>>> |
```

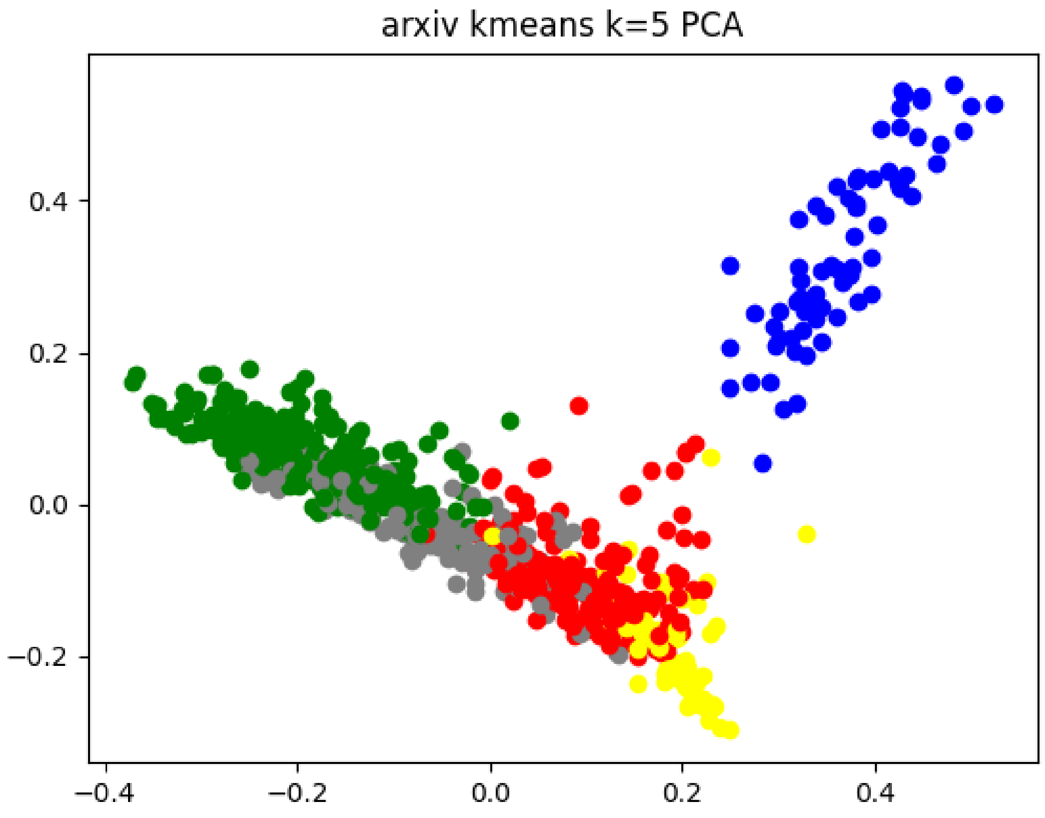




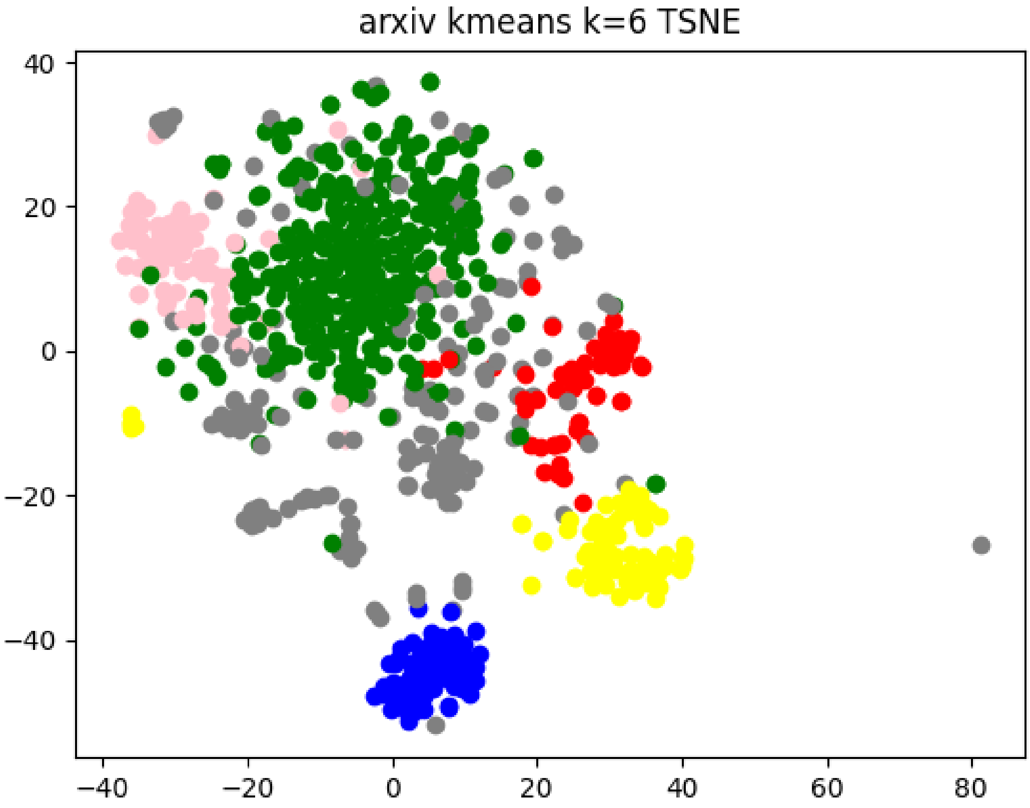


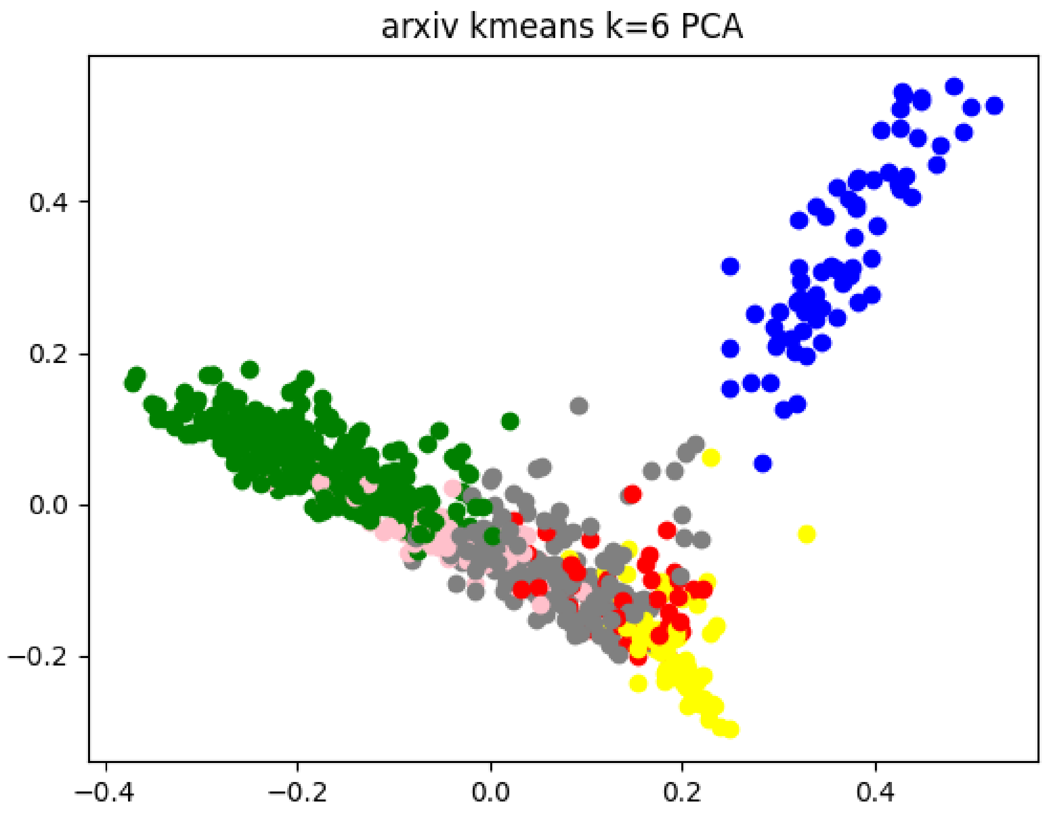
```
se models with scikit-learn 0.21+.
k = 4:
file counts in each cluster:
1    295
0    288
2    135
3     70
Name: label, dtype: int64

cluster 0 top 10 words:['use']  ['data']  ['protein']  ['user']  ['tweet']  ['drug']  ['model']  ['infect']  ['test']  ['case']
cluster 1 top 10 words:['model']  ['infect']  ['case']  ['number']  ['data']  ['time']  ['epidem']  ['r']  ['day']  ['rate']
cluster 2 top 10 words:['node']  ['network']  ['infect']  ['k']  ['n']  ['model']  ['q']  ['individu']  ['r']  ['x']
cluster 3 top 10 words:['imag']  ['ct']  ['xray']  ['train']  ['learn']  ['dataset']  ['chest']  ['segment']  ['use']  ['pneumonia']
>>>
```

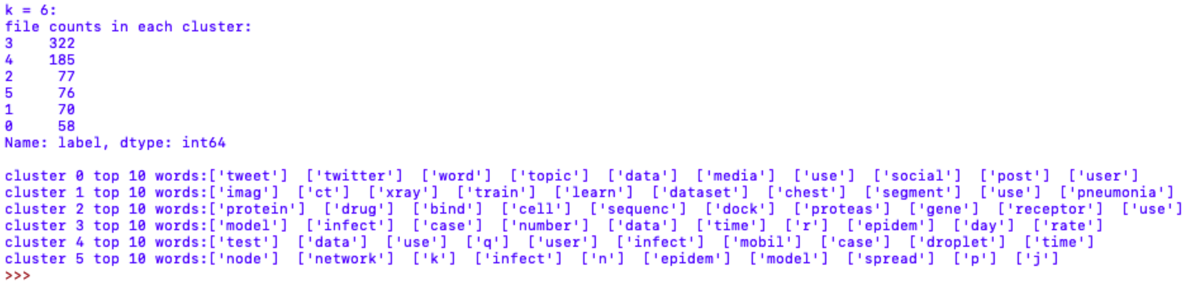

arxiv kmeans k=5 TSNE
40
20
0
−20
−40
−60
−40
−20
0
20


arxiv kmeans k=5 PCA
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
0.4

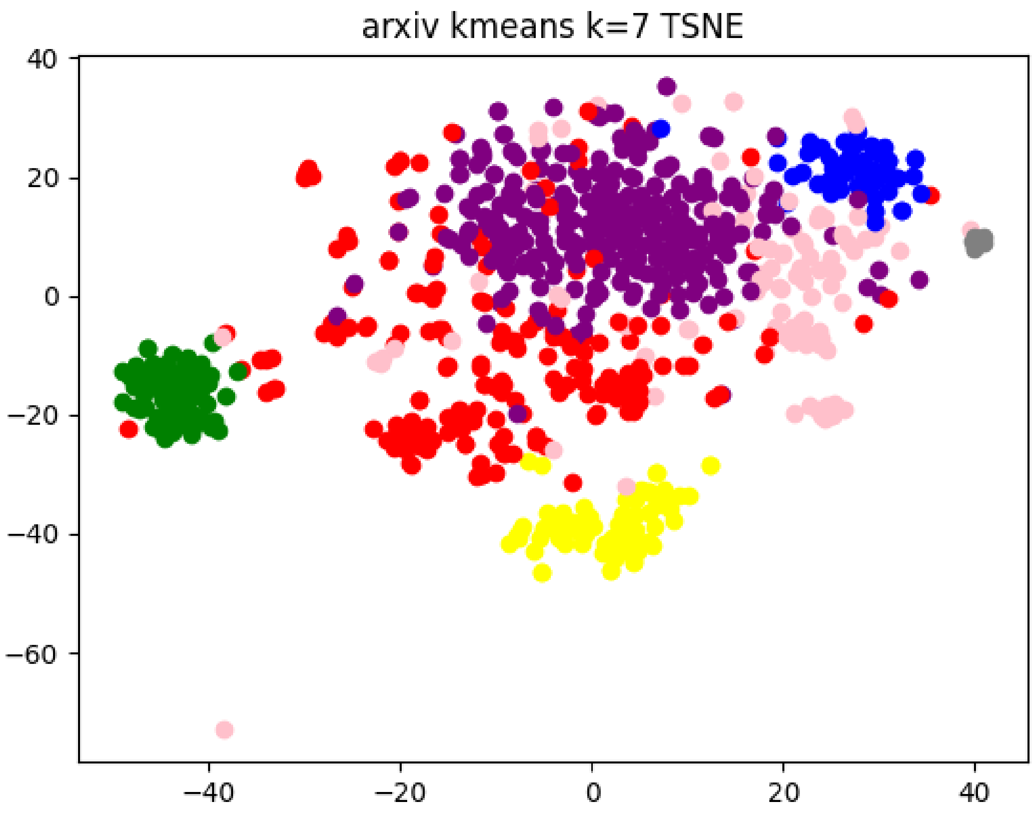
```
k = 5:
file counts in each cluster:
3    269
0    209
4    165
2     75
1     70
Name: label, dtype: int64

cluster 0 top 10 words:['data']  ['user']  ['use']  ['tweet']  ['case']  ['time']  ['inform']  ['mobil']  ['test']  ['model']
cluster 1 top 10 words:['imag']  ['ct']  ['xray']  ['train']  ['learn']  ['dataset']  ['chest']  ['segment']  ['use']  ['pneumonia']
cluster 2 top 10 words:['protein']  ['drug']  ['bind']  ['cell']  ['sequenc']  ['dock']  ['gene']  ['proteas']  ['receptor']  ['use']
cluster 3 top 10 words:['model']  ['infect']  ['case']  ['number']  ['data']  ['time']  ['epidem']  ['day']  ['r']  ['rate']
cluster 4 top 10 words:['infect']  ['node']  ['network']  ['n']  ['model']  ['k']  ['r']  ['q']  ['individu']  ['epidem']
>>>
```

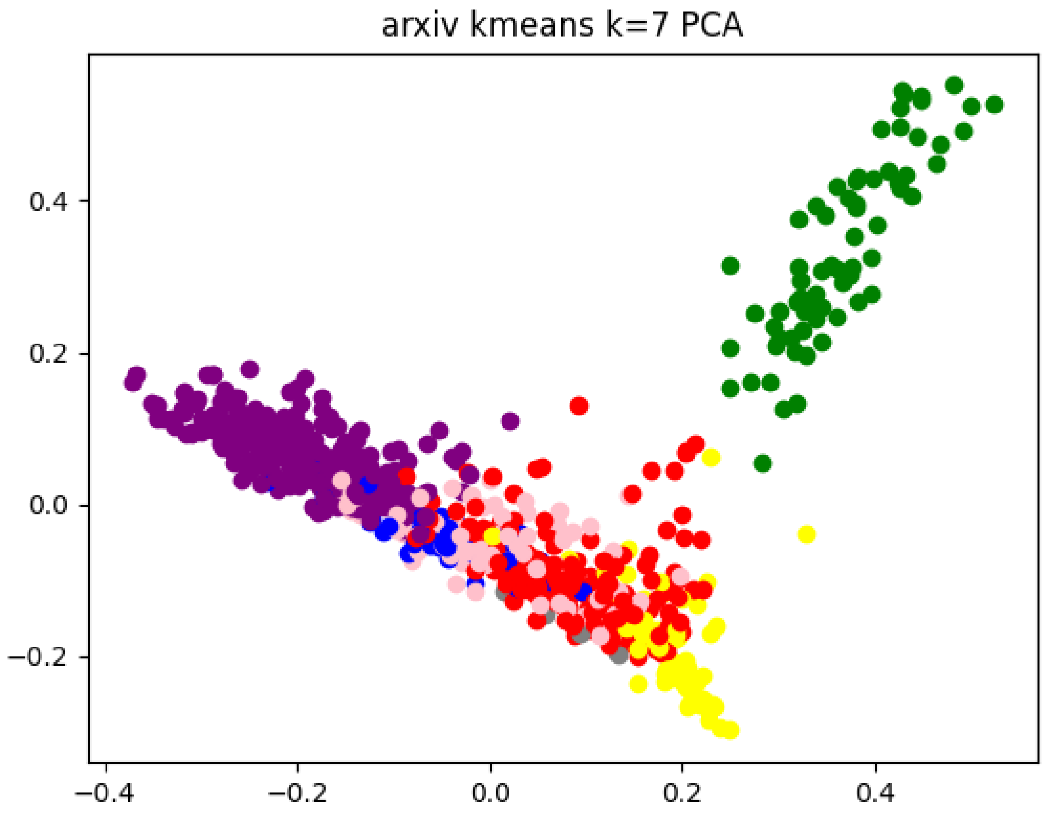


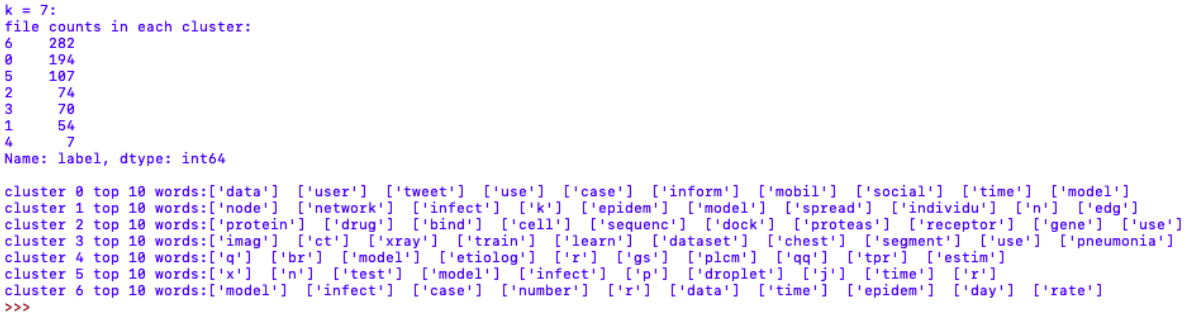




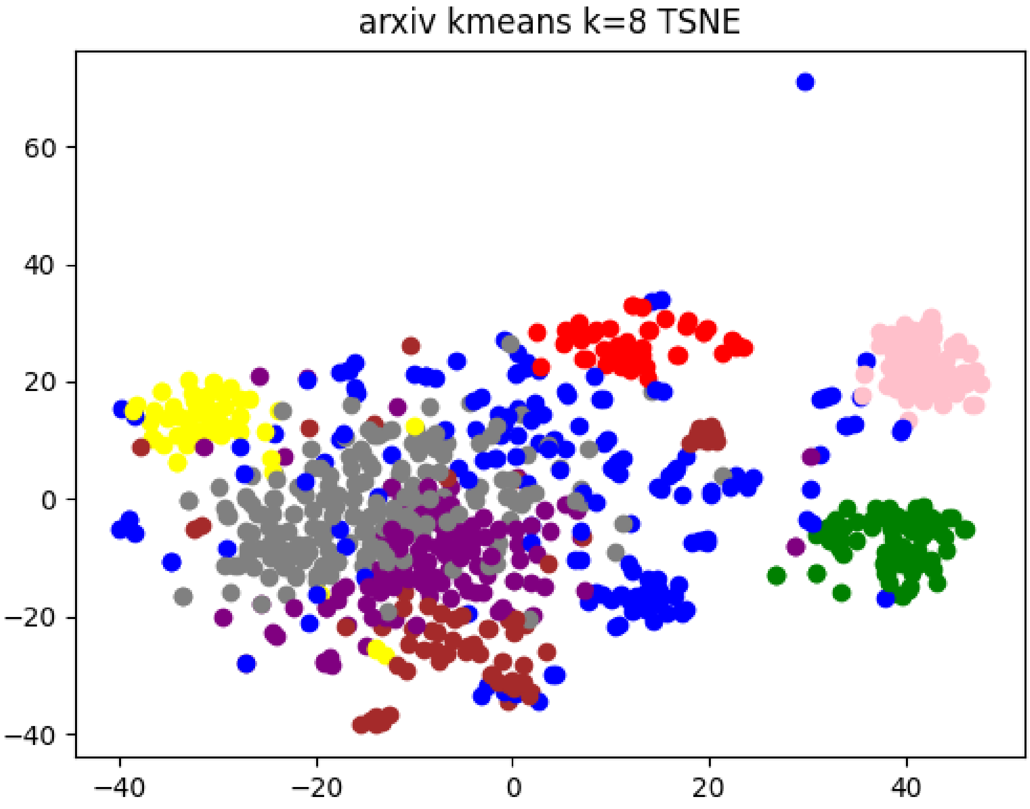
```
k = 6:
file counts in each cluster:
3    322
4    185
2     77
5     76
1     70
0     58
Name: label, dtype: int64

cluster 0 top 10 words:['tweet']  ['twitter']  ['word']  ['topic']  ['data']  ['media']  ['use']  ['social']  ['post']  ['user']
cluster 1 top 10 words:['imag']  ['ct']  ['xray']  ['train']  ['learn']  ['dataset']  ['chest']  ['segment']  ['use']  ['pneumonia']
cluster 2 top 10 words:['protein']  ['drug']  ['bind']  ['cell']  ['sequenc']  ['dock']  ['proteas']  ['gene']  ['receptor']  ['use']
cluster 3 top 10 words:['model']  ['infect']  ['case']  ['number']  ['data']  ['time']  ['r']  ['epidem']  ['day']  ['rate']
cluster 4 top 10 words:['test']  ['data']  ['use']  ['q']  ['user']  ['infect']  ['mobil']  ['case']  ['droplet']  ['time']
cluster 5 top 10 words:['node']  ['network']  ['k']  ['infect']  ['n']  ['epidem']  ['model']  ['spread']  ['p']  ['j']
>>>
```

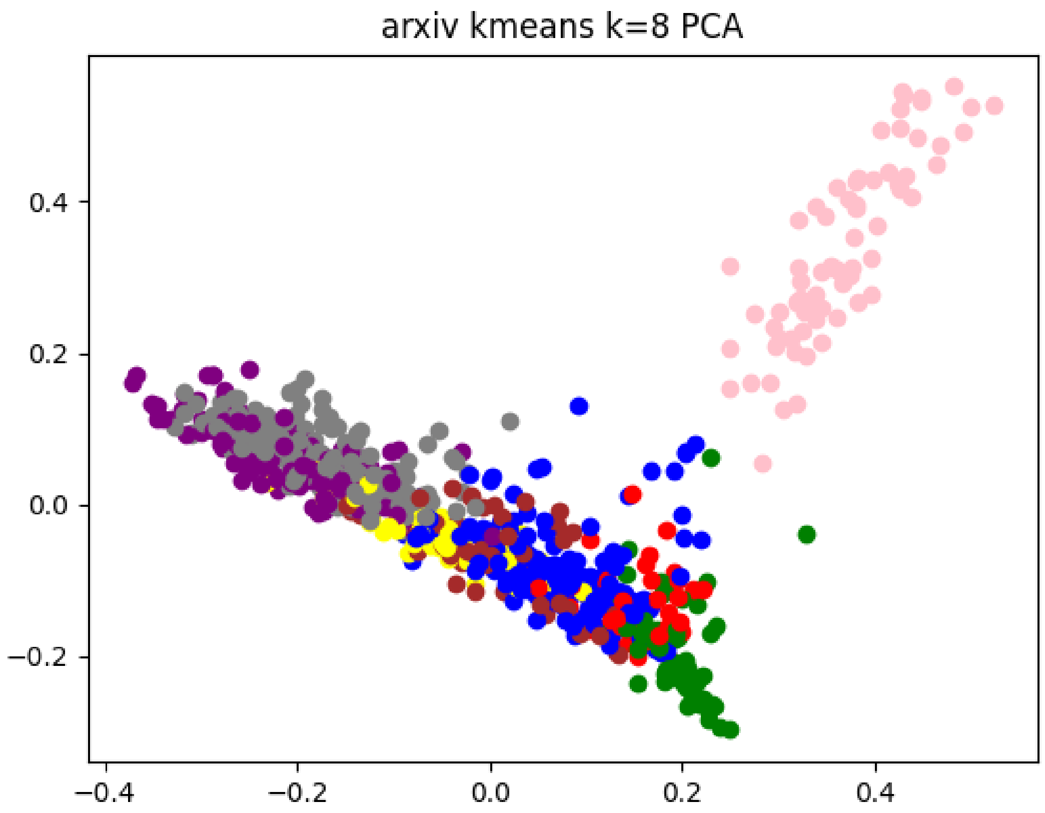

arxiv kmeans k=7 TSNE
40
20
0
−20
−40
−60
−40
−20
0
20
40

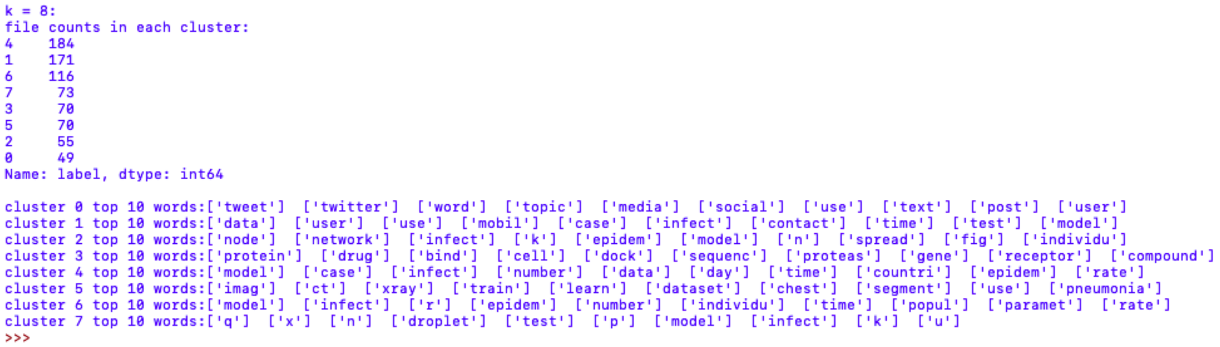

arxiv kmeans k=7 PCA
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
0.4



```
k = 7:
file counts in each cluster:
6    282
0    194
5    107
2     74
3     70
1     54
4      7
Name: label, dtype: int64

cluster 0 top 10 words:['data']  ['user']  ['tweet']  ['use']  ['case']  ['inform']  ['mobil']  ['social']  ['time']  ['model']
cluster 1 top 10 words:['node']  ['network']  ['infect']  ['k']  ['epidem']  ['model']  ['spread']  ['individu']  ['n']  ['edg']
cluster 2 top 10 words:['protein']  ['drug']  ['bind']  ['cell']  ['sequenc']  ['dock']  ['proteas']  ['receptor']  ['gene']  ['use']
cluster 3 top 10 words:['imag']  ['ct']  ['xray']  ['train']  ['learn']  ['dataset']  ['chest']  ['segment']  ['use']  ['pneumonia']
cluster 4 top 10 words:['q']  ['br']  ['model']  ['etiolog']  ['r']  ['gs']  ['plcm']  ['qq']  ['tpr']  ['estim']
cluster 5 top 10 words:['x']  ['n']  ['test']  ['model']  ['infect']  ['p']  ['droplet']  ['j']  ['time']  ['r']
cluster 6 top 10 words:['model']  ['infect']  ['case']  ['number']  ['r']  ['data']  ['time']  ['epidem']  ['day']  ['rate']
>>>
```







```
k = 8:
file counts in each cluster:
4    184
1    171
6    116
7     73
3     70
5     70
2     55
0     49
Name: label, dtype: int64

cluster 0 top 10 words:['tweet']  ['twitter']  ['word']  ['topic']  ['media']  ['social']  ['use']  ['text']  ['post']  ['user']
cluster 1 top 10 words:['data']  ['user']  ['use']  ['mobil']  ['case']  ['infect']  ['contact']  ['time']  ['test']  ['model']
cluster 2 top 10 words:['node']  ['network']  ['infect']  ['k']  ['epidem']  ['model']  ['n']  ['spread']  ['fig']  ['individu']
cluster 3 top 10 words:['protein']  ['drug']  ['bind']  ['cell']  ['dock']  ['sequenc']  ['proteas']  ['gene']  ['receptor']  ['compound']
cluster 4 top 10 words:['model']  ['case']  ['infect']  ['number']  ['data']  ['day']  ['time']  ['countri']  ['epidem']  ['rate']
cluster 5 top 10 words:['imag']  ['ct']  ['xray']  ['train']  ['learn']  ['dataset']  ['chest']  ['segment']  ['use']  ['pneumonia']
cluster 6 top 10 words:['model']  ['infect']  ['r']  ['epidem']  ['number']  ['individu']  ['time']  ['popul']  ['paramet']  ['rate']
cluster 7 top 10 words:['q']  ['x']  ['n']  ['droplet']  ['test']  ['p']  ['model']  ['infect']  ['k']  ['u']
>>>
```

Based on the graphs, K = 3 should be the optimal value. The following is the result from recovering the stem words of each cluster when K = 3:

Cluster 0: use, data, protein, user, test, model, infect, tweet, q, drug

Cluster 1: model, infect, case, number, epidemic, r, time, data, rate, popular

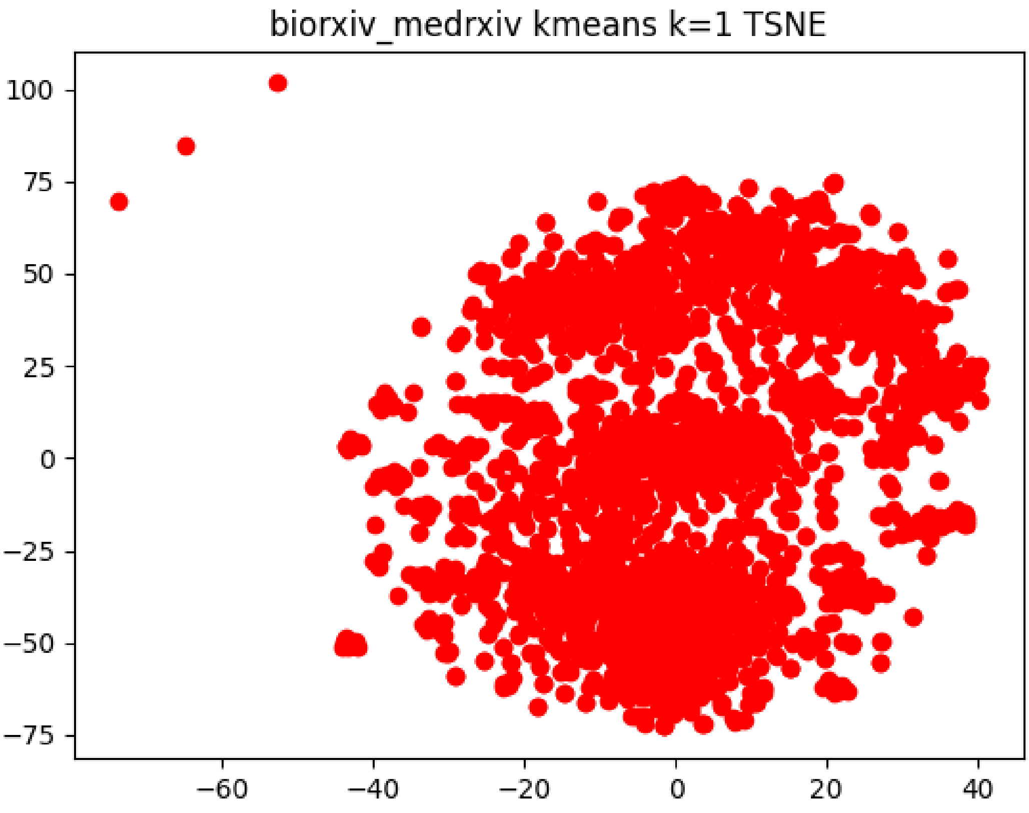
Cluster 2: image, ct, xray, train, learn, dataset, chest, segment, use, pneumonia

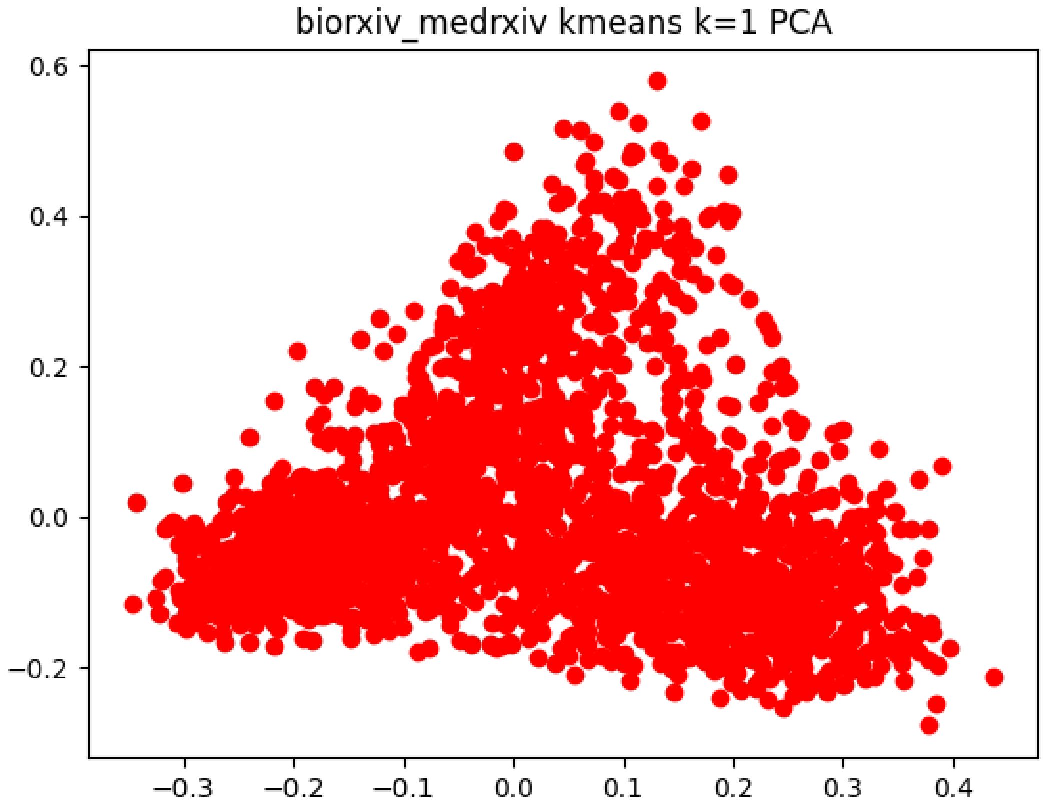
Cluster 0 outlines that researchers attempt to gather infection-related information from Twitter's user data, such as people's drug use, and build models.

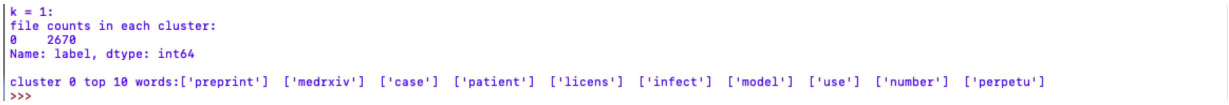
Cluster 1 outlines several key factors in epidemiological studies, the speed of transmission, the time span of transmission, and the number of people affected.

Cluster 2 mentions pneumonia, which is very similar to COVID-19, and X-ray images of the chest, possibly in an attempt to aid diagnosis and treatment with machine learning.

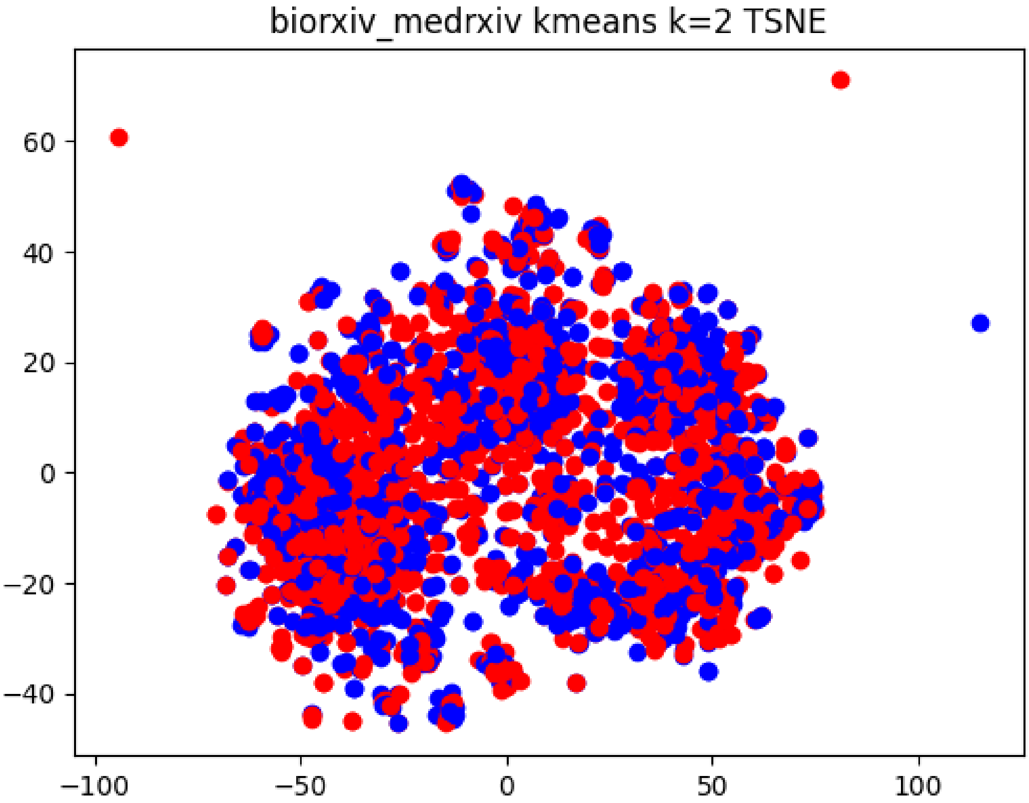
## biorxiv_medrxiv

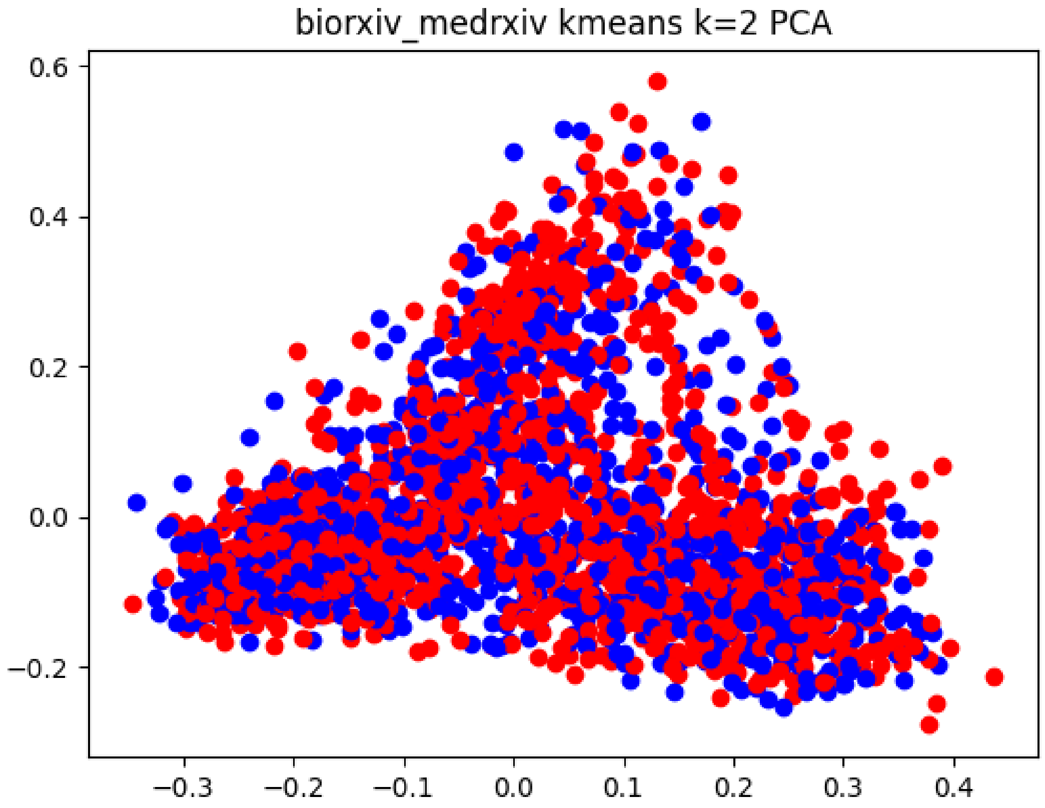


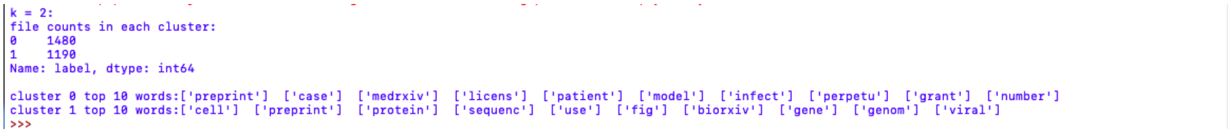




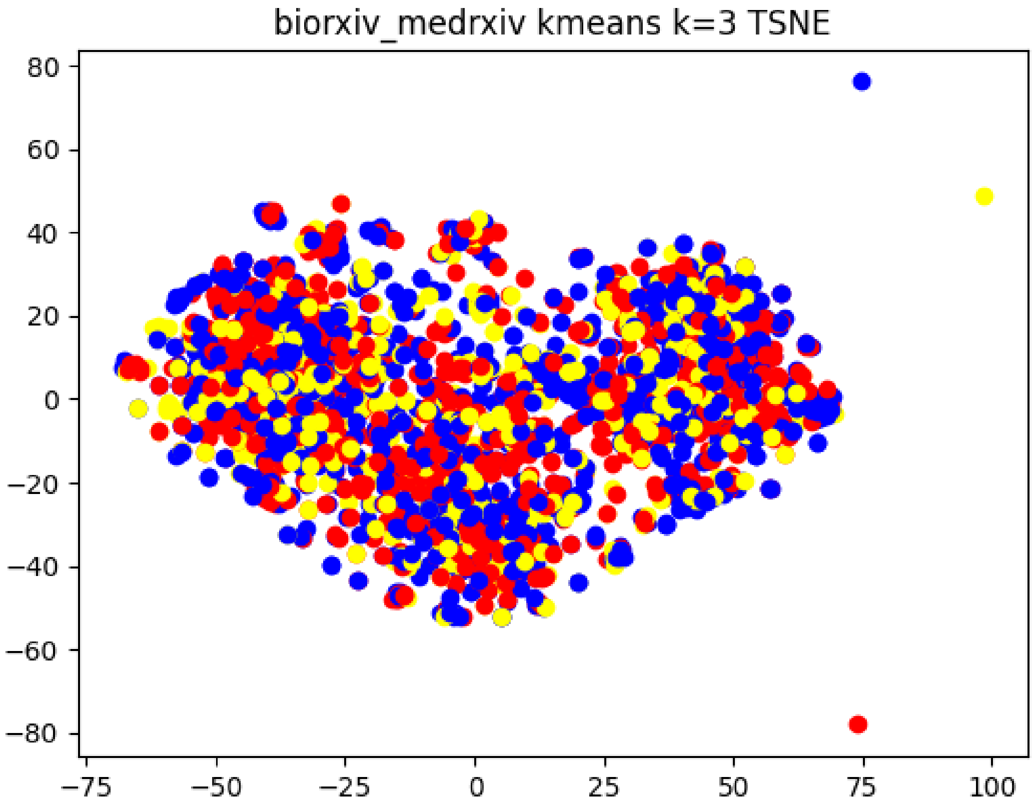
```
k = 1:
file counts in each cluster:
0    2670
Name: label, dtype: int64

cluster 0 top 10 words:['preprint']  ['medrxiv']  ['case']  ['patient']  ['licens']  ['infect']  ['model']  ['use']  ['number']  ['perpetu']
>>>
```

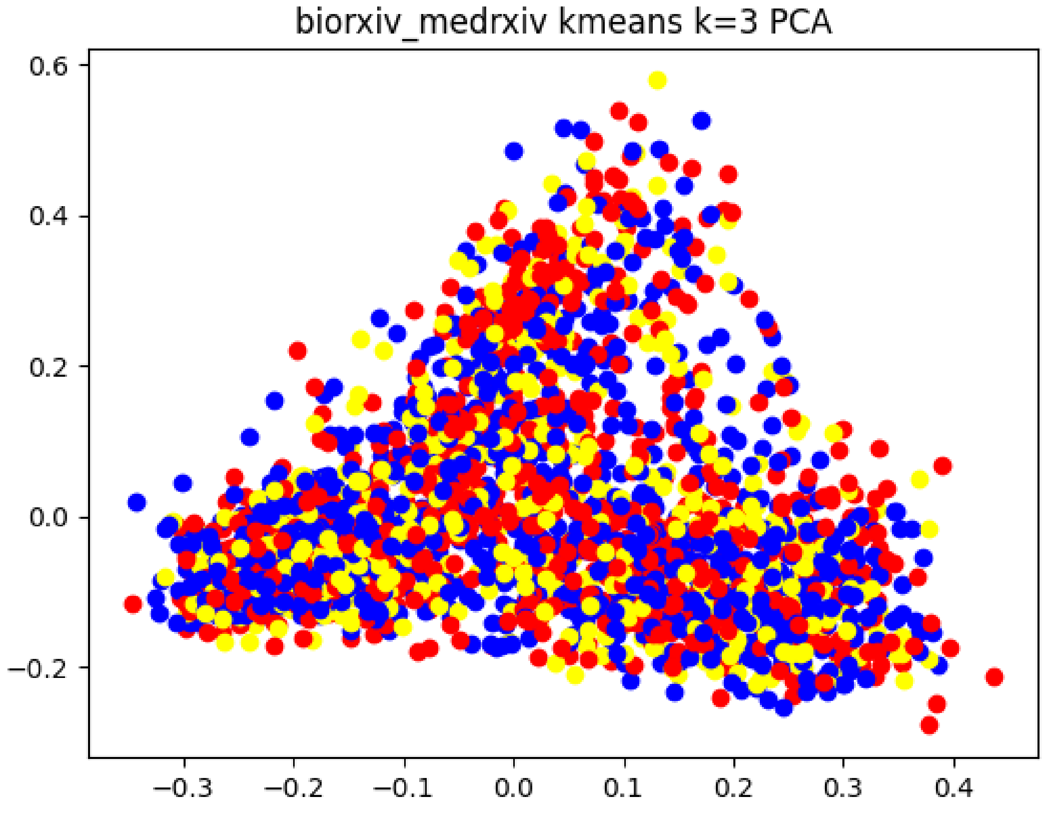


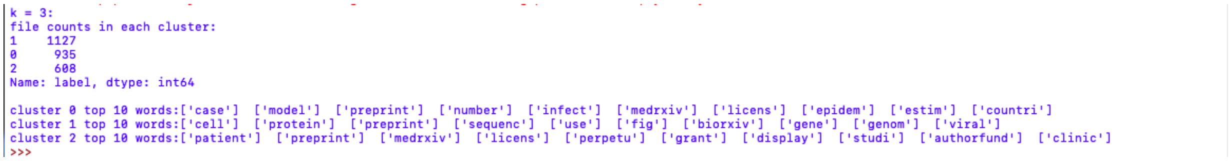




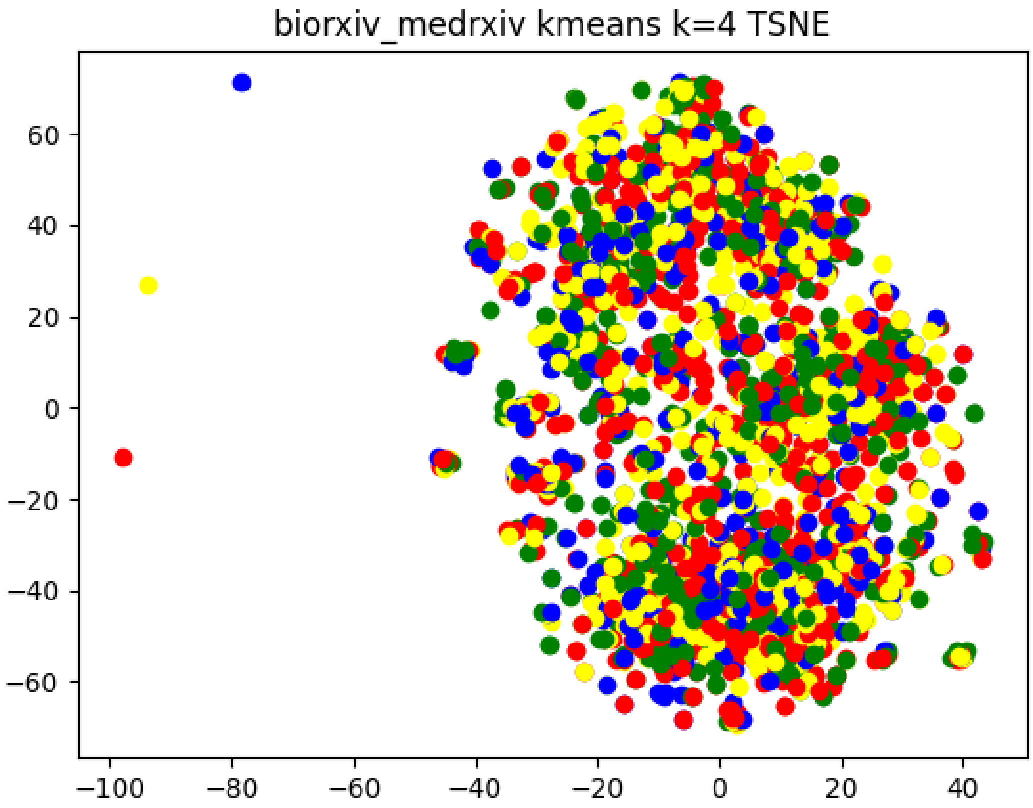
```
k = 2:
file counts in each cluster:
0    1480
1    1190
Name: label, dtype: int64

cluster 0 top 10 words:['preprint']  ['case']  ['medrxiv']  ['licens']  ['patient']  ['model']  ['infect']  ['perpetu']  ['grant']  ['number']
cluster 1 top 10 words:['cell']  ['preprint']  ['protein']  ['sequenc']  ['use']  ['fig']  ['biorxiv']  ['gene']  ['genom']  ['viral']
>>>
```

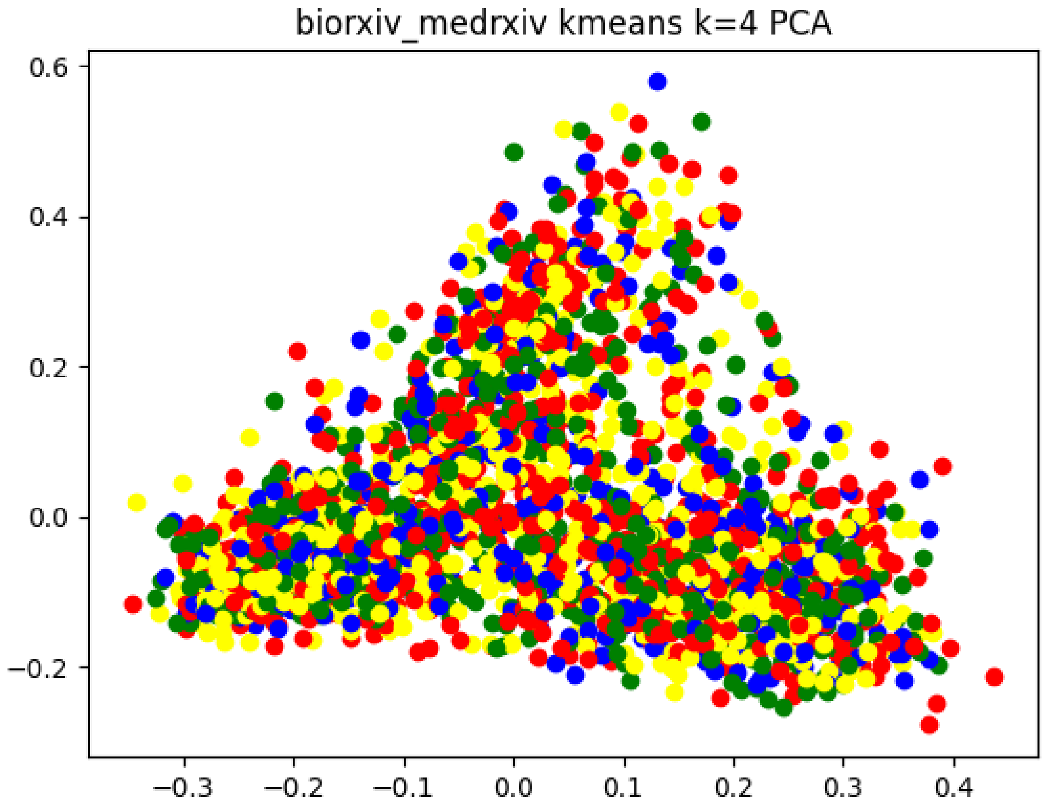

biorxiv_medrxiv kmeans k=3 TSNE
80
60
40
20
0
−20
−40
−60
−80
−75
−50
−25
0
25
50
75
100

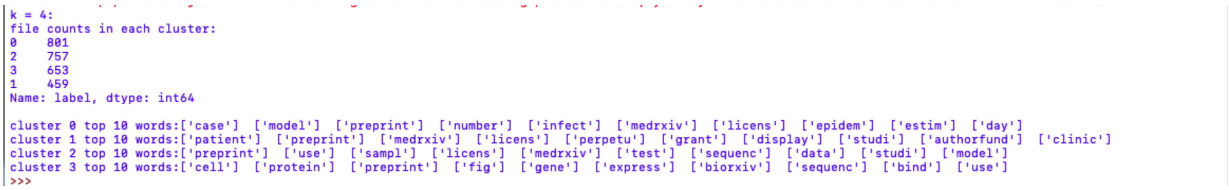

biorxiv_medrxiv kmeans k=3 PCA
0.6
0.4
0.2
0.0
−0.2
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4



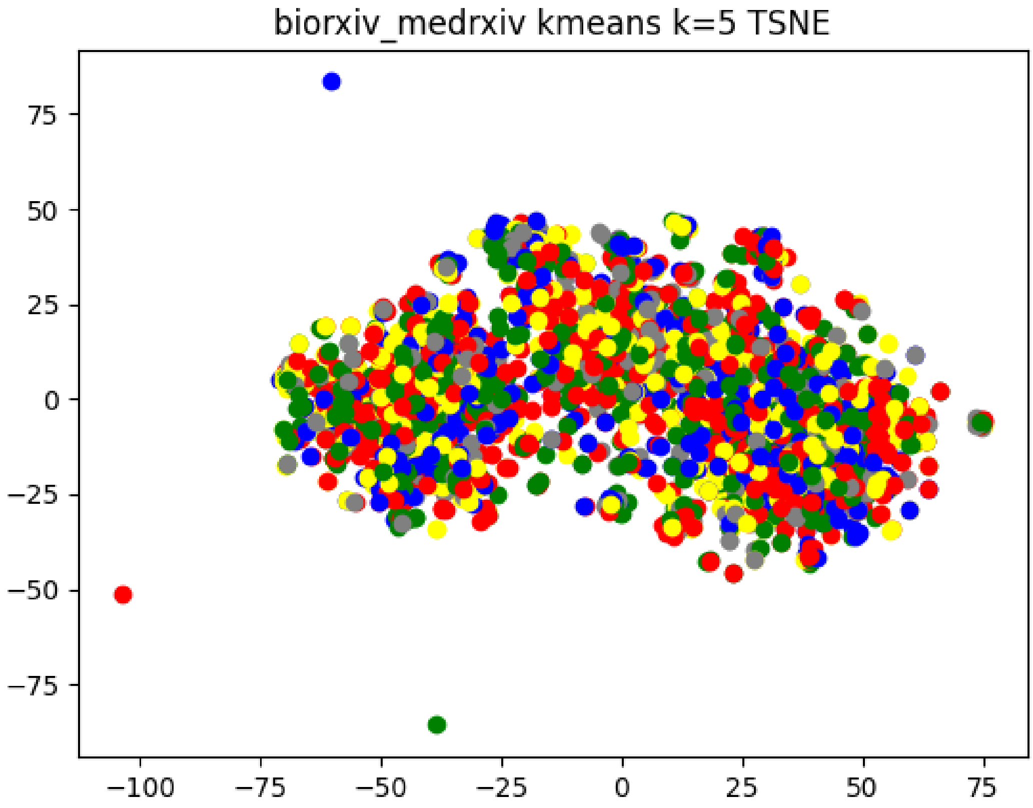
```
k = 3:
file counts in each cluster:
1    1127
0     935
2     608
Name: label, dtype: int64

cluster 0 top 10 words:['case']  ['model']  ['preprint']  ['number']  ['infect']  ['medrxiv']  ['licens']  ['epidem']  ['estim']  ['countri']
cluster 1 top 10 words:['cell']  ['protein']  ['preprint']  ['sequenc']  ['use']  ['fig']  ['biorxiv']  ['gene']  ['genom']  ['viral']
cluster 2 top 10 words:['patient']  ['preprint']  ['medrxiv']  ['licens']  ['perpetu']  ['grant']  ['display']  ['studi']  ['authorfund']  ['clinic']
>>>
```

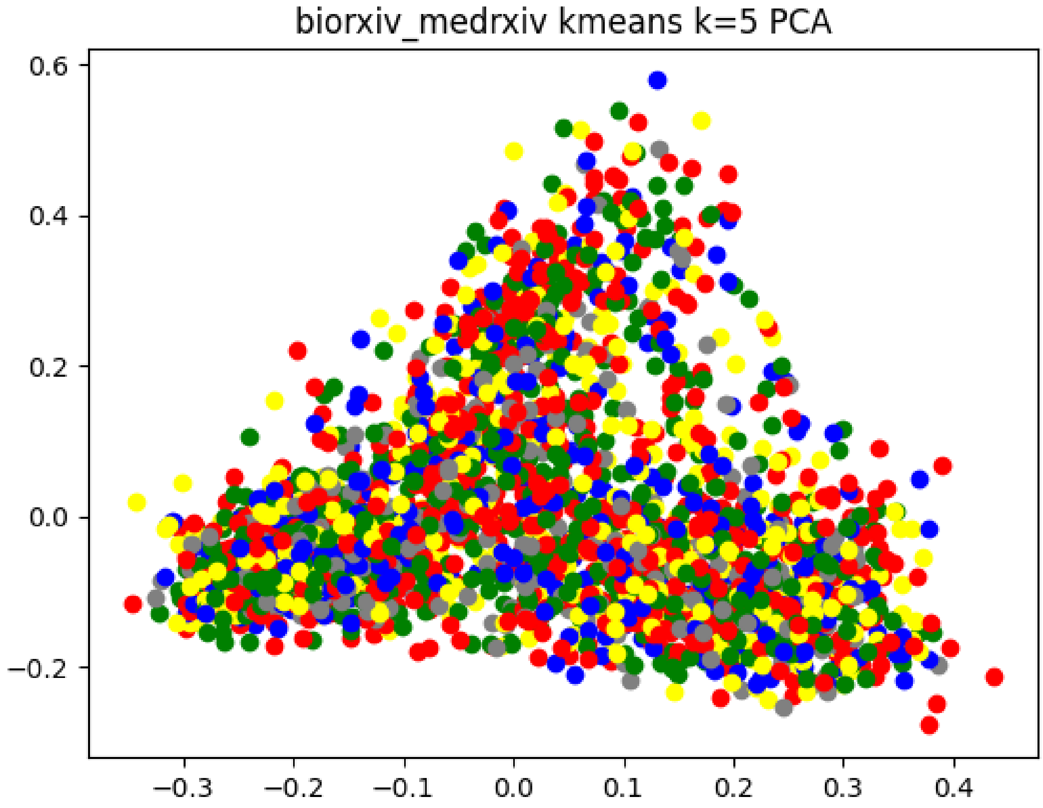


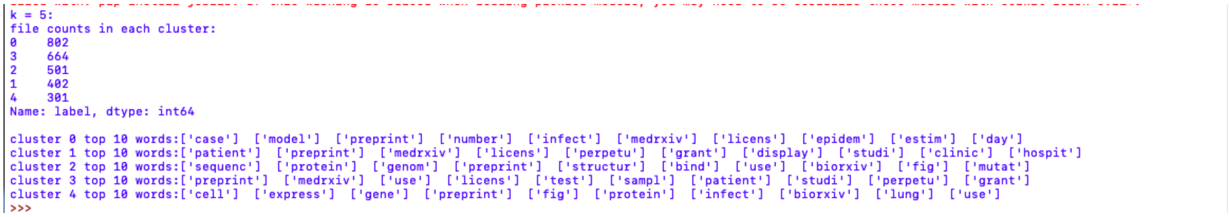




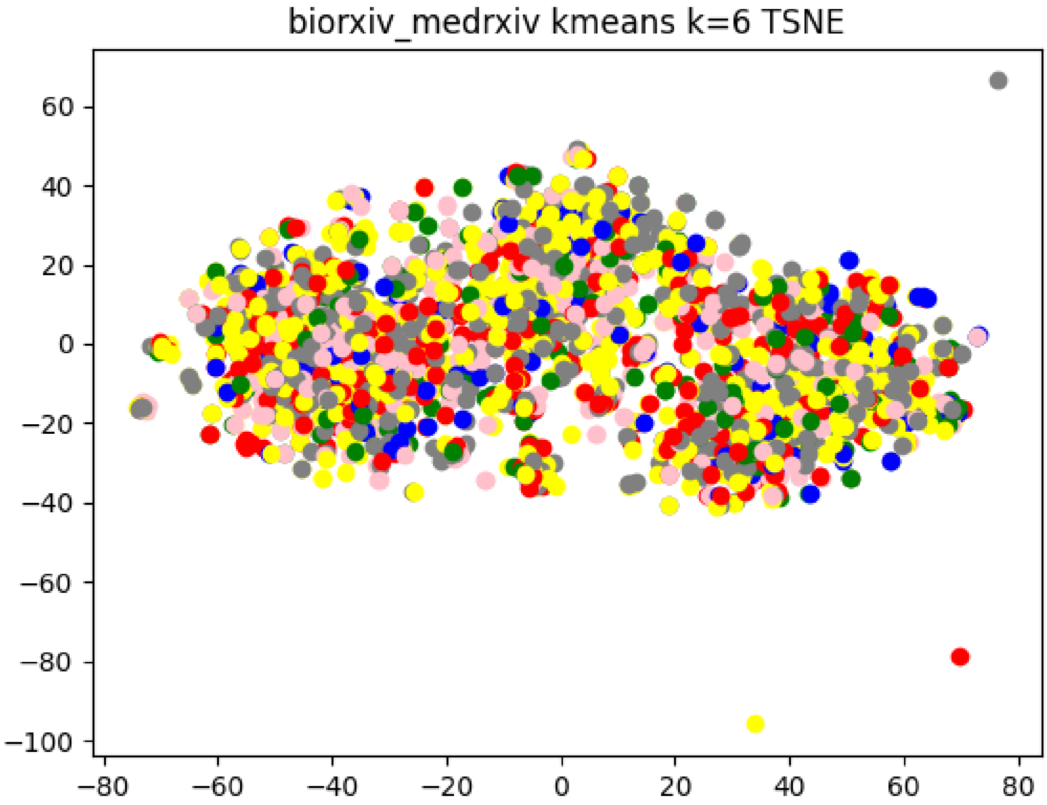
```
k = 4:
file counts in each cluster:
0    801
2    757
3    653
1    459
Name: label, dtype: int64

cluster 0 top 10 words:['case']  ['model']  ['preprint']  ['number']  ['infect']  ['medrxiv']  ['licens']  ['epidem']  ['estim']  ['day']
cluster 1 top 10 words:['patient']  ['preprint']  ['medrxiv']  ['licens']  ['perpetu']  ['grant']  ['display']  ['studi']  ['authorfund']  ['clinic']
cluster 2 top 10 words:['preprint']  ['use']  ['sampl']  ['licens']  ['medrxiv']  ['test']  ['sequenc']  ['data']  ['studi']  ['model']
cluster 3 top 10 words:['cell']  ['protein']  ['preprint']  ['fig']  ['gene']  ['express']  ['biorxiv']  ['sequenc']  ['bind']  ['use']
>>>
```

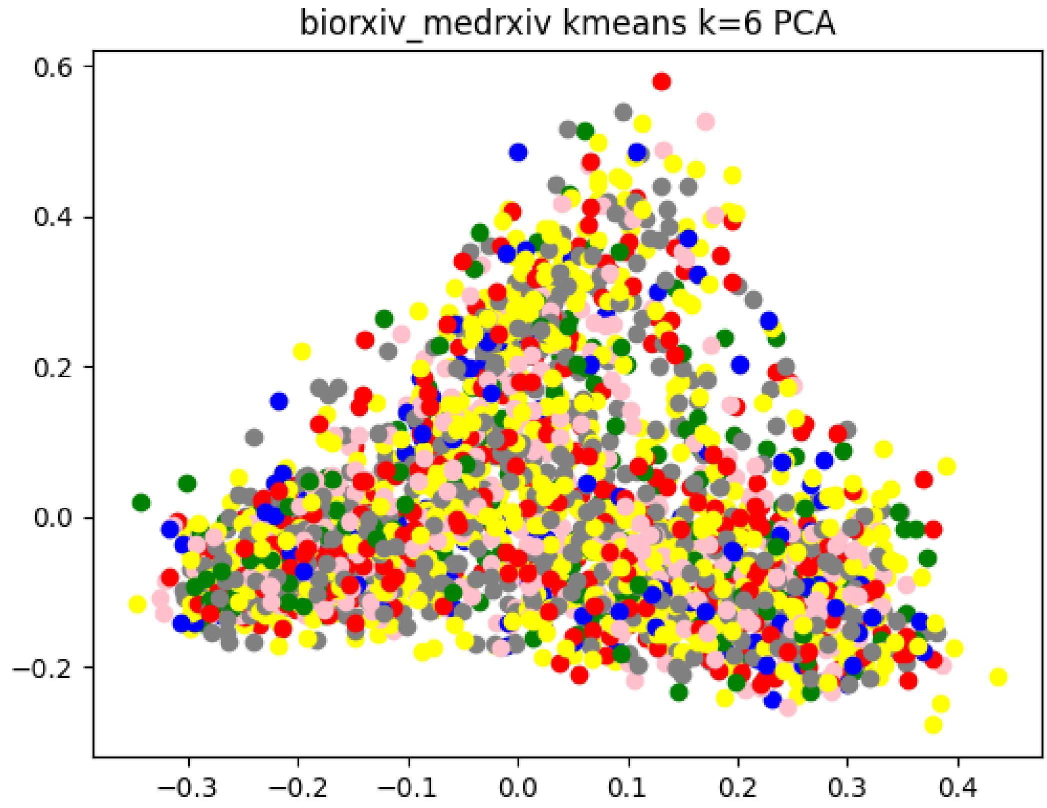

biorxiv_medrxiv kmeans k=5 TSNE
75
50
25
0
−25
−50
−75
−100
−75
−50
−25
0
25
50
75

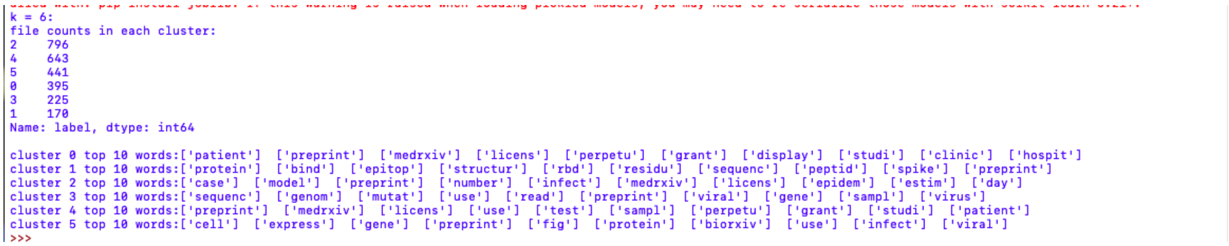

biorxiv_medrxiv kmeans k=5 PCA
0.6
0.4
0.2
0.0
−0.2
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4



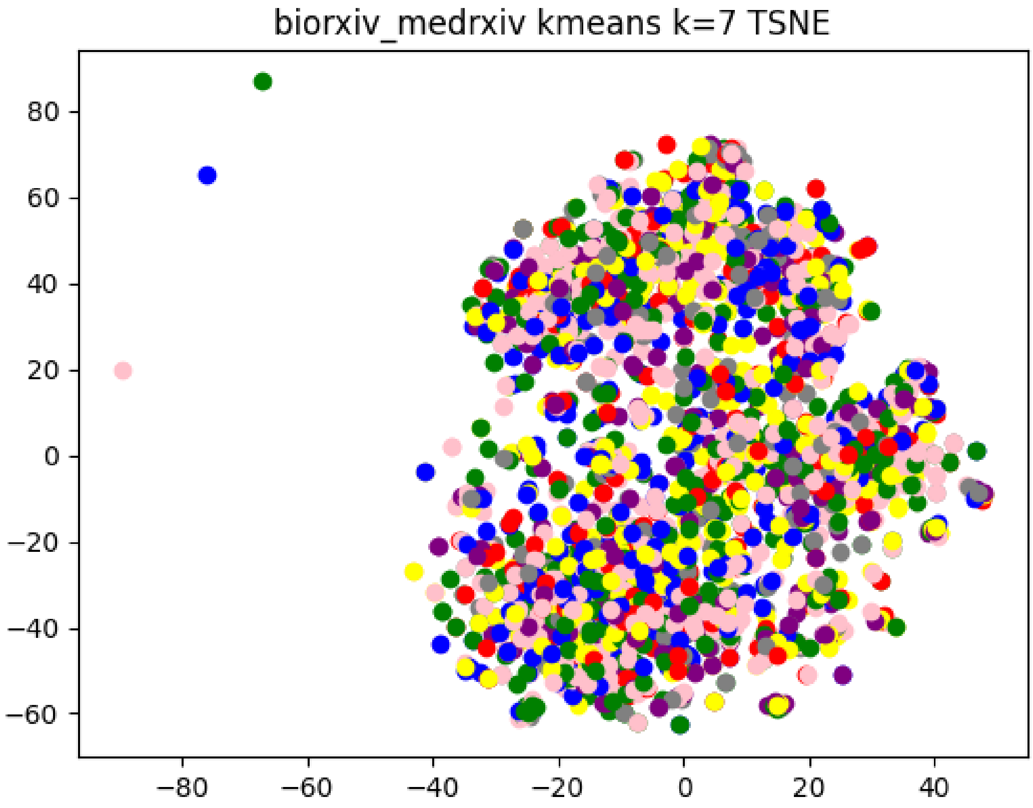
```
k = 5:
file counts in each cluster:
0    802
3    664
2    501
1    402
4    301
Name: label, dtype: int64

cluster 0 top 10 words:['case']  ['model']  ['preprint']  ['number']  ['infect']  ['medrxiv']  ['licens']  ['epidem']  ['estim']  ['day']
cluster 1 top 10 words:['patient']  ['preprint']  ['medrxiv']  ['licens']  ['perpetu']  ['grant']  ['display']  ['studi']  ['clinic']  ['hospit']
cluster 2 top 10 words:['sequenc']  ['protein']  ['genom']  ['preprint']  ['structur']  ['bind']  ['use']  ['biorxiv']  ['fig']  ['mutat']
cluster 3 top 10 words:['preprint']  ['medrxiv']  ['use']  ['licens']  ['test']  ['sampl']  ['patient']  ['studi']  ['perpetu']  ['grant']
cluster 4 top 10 words:['cell']  ['express']  ['gene']  ['preprint']  ['fig']  ['protein']  ['infect']  ['biorxiv']  ['lung']  ['use']
>>>
```

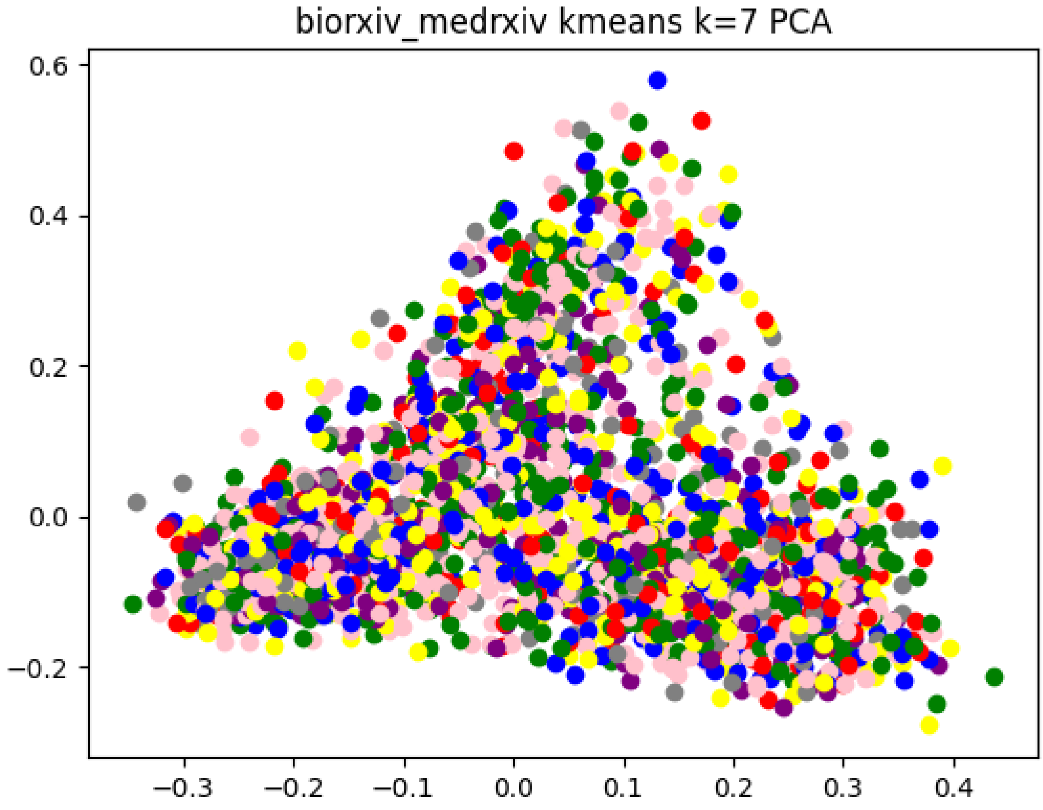


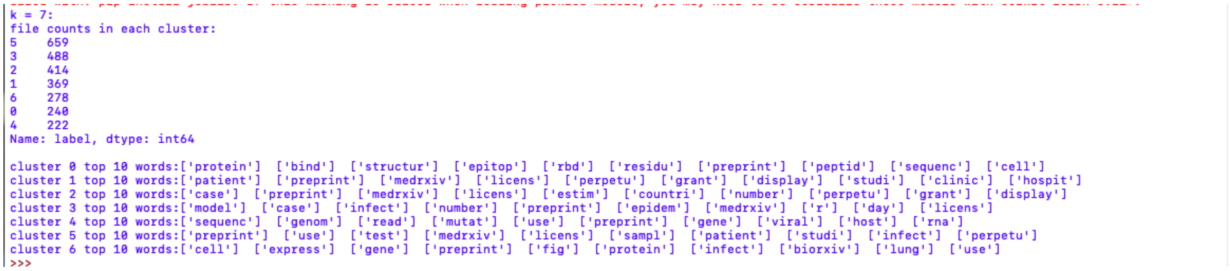




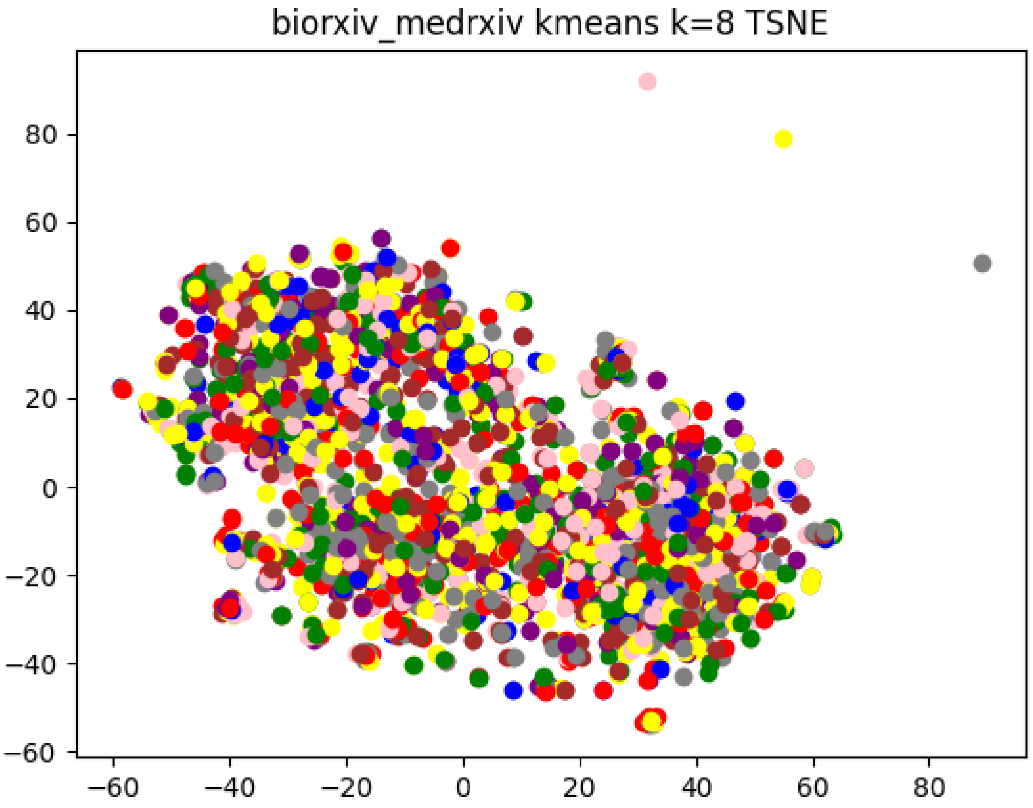
```
alled with: pip install joblib. If this warning is raised when loading pickled models, you may need to re-serialize those models with scikit-learn 0.21+.
k = 6:
file counts in each cluster:
2    796
4    643
5    441
0    395
3    225
1    170
Name: label, dtype: int64

cluster 0 top 10 words:['patient']  ['preprint']  ['medrxiv']  ['licens']  ['perpetu']  ['grant']  ['display']  ['studi']  ['clinic']  ['hospit']
cluster 1 top 10 words:['protein']  ['bind']  ['epitop']  ['structur']  ['rbd']  ['residu']  ['sequenc']  ['peptid']  ['spike']  ['preprint']
cluster 2 top 10 words:['case']  ['model']  ['preprint']  ['number']  ['infect']  ['medrxiv']  ['licens']  ['epidem']  ['estim']  ['day']
cluster 3 top 10 words:['sequenc']  ['genom']  ['mutat']  ['use']  ['read']  ['preprint']  ['viral']  ['gene']  ['sampl']  ['virus']
cluster 4 top 10 words:['preprint']  ['medrxiv']  ['licens']  ['use']  ['test']  ['sampl']  ['perpetu']  ['grant']  ['studi']  ['patient']
cluster 5 top 10 words:['cell']  ['express']  ['gene']  ['preprint']  ['fig']  ['protein']  ['biorxiv']  ['use']  ['infect']  ['viral']
>>>
```

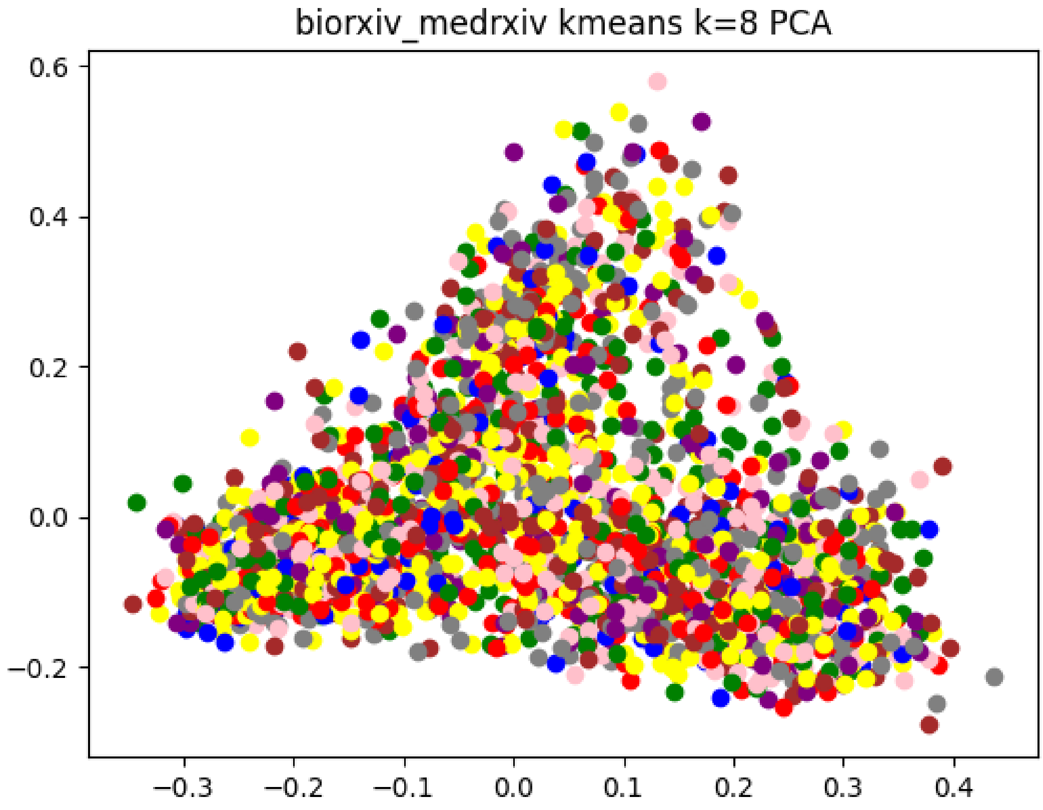

biorxiv_medrxiv kmeans k=7 TSNE
80
60
40
20
0
−20
−40
−60
−80
−60
−40
−20
0
20
40

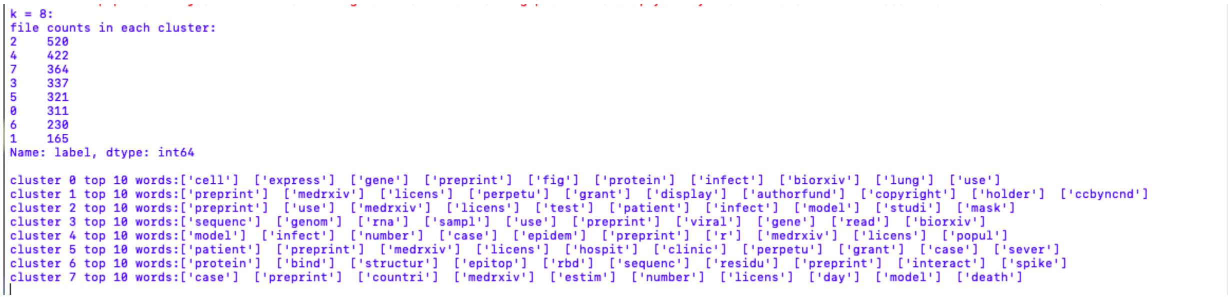

biorxiv_medrxiv kmeans k=7 PCA
0.6
0.4
0.2
0.0
−0.2
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4



```
k = 7:
file counts in each cluster:
5    659
3    488
2    414
1    369
6    278
0    240
4    222
Name: label, dtype: int64

cluster 0 top 10 words:['protein']  ['bind']  ['structur']  ['epitop']  ['rbd']  ['residu']  ['preprint']  ['peptid']  ['sequenc']  ['cell']
cluster 1 top 10 words:['patient']  ['preprint']  ['medrxiv']  ['licens']  ['perpetu']  ['grant']  ['display']  ['studi']  ['clinic']  ['hospit']
cluster 2 top 10 words:['case']  ['preprint']  ['medrxiv']  ['licens']  ['estim']  ['countri']  ['number']  ['perpetu']  ['grant']  ['display']
cluster 3 top 10 words:['model']  ['case']  ['infect']  ['number']  ['preprint']  ['epidem']  ['medrxiv']  ['r']  ['day']  ['licens']
cluster 4 top 10 words:['sequenc']  ['genom']  ['read']  ['mutat']  ['use']  ['preprint']  ['gene']  ['viral']  ['host']  ['rna']
cluster 5 top 10 words:['preprint']  ['use']  ['test']  ['medrxiv']  ['licens']  ['sampl']  ['patient']  ['studi']  ['infect']  ['perpetu']
cluster 6 top 10 words:['cell']  ['express']  ['gene']  ['preprint']  ['fig']  ['protein']  ['infect']  ['biorxiv']  ['lung']  ['use']
>>>
```



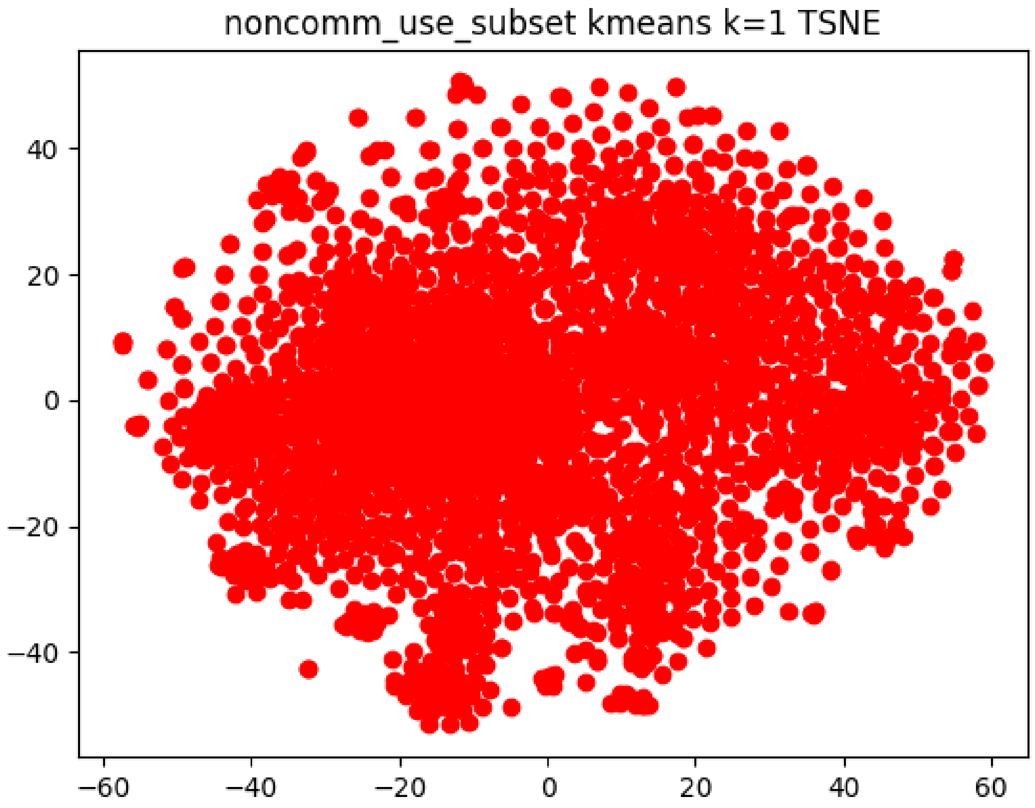


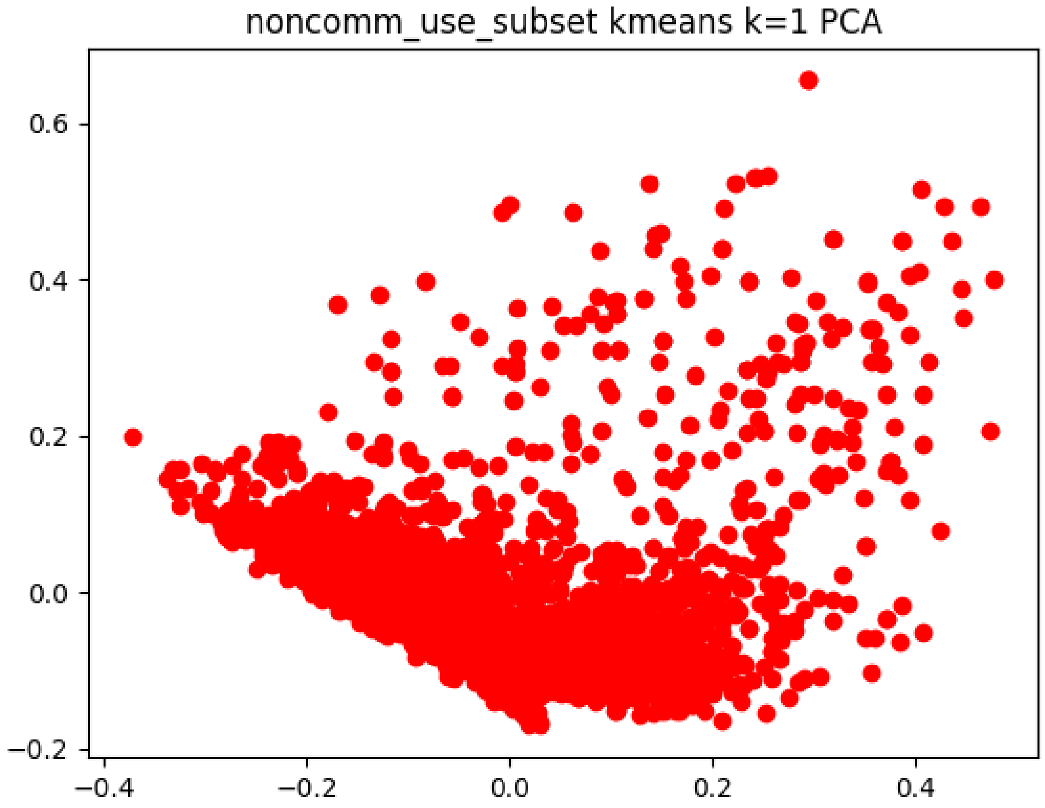


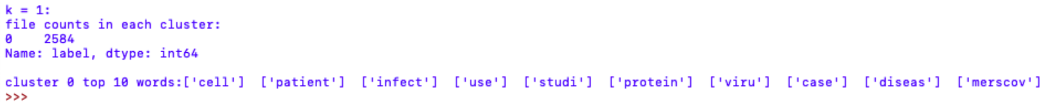
```
k = 8:
file counts in each cluster:
2    520
4    422
7    364
3    337
5    321
0    311
6    230
1    165
Name: label, dtype: int64

cluster 0 top 10 words:['cell']  ['express']  ['gene']  ['preprint']  ['fig']  ['protein']  ['infect']  ['biorxiv']  ['lung']  ['use']
cluster 1 top 10 words:['preprint']  ['medrxiv']  ['licens']  ['perpetu']  ['grant']  ['display']  ['authorfund']  ['copyright']  ['holder']  ['ccbyncnd']
cluster 2 top 10 words:['preprint']  ['use']  ['medrxiv']  ['licens']  ['test']  ['patient']  ['infect']  ['model']  ['studi']  ['mask']
cluster 3 top 10 words:['sequenc']  ['genom']  ['rna']  ['sampl']  ['use']  ['preprint']  ['viral']  ['gene']  ['read']  ['biorxiv']
cluster 4 top 10 words:['model']  ['infect']  ['number']  ['case']  ['epidem']  ['preprint']  ['r']  ['medrxiv']  ['licens']  ['popul']
cluster 5 top 10 words:['patient']  ['preprint']  ['medrxiv']  ['licens']  ['hospit']  ['clinic']  ['perpetu']  ['grant']  ['case']  ['sever']
cluster 6 top 10 words:['protein']  ['bind']  ['structur']  ['epitop']  ['rbd']  ['sequenc']  ['residu']  ['preprint']  ['interact']  ['spike']
cluster 7 top 10 words:['case']  ['preprint']  ['countri']  ['medrxiv']  ['estim']  ['number']  ['licens']  ['day']  ['model']  ['death']
```

The classified images of this source showed no obvious pattern, and it was generally seen from the keywords that these papers were also about virus research and hospital data analysis.

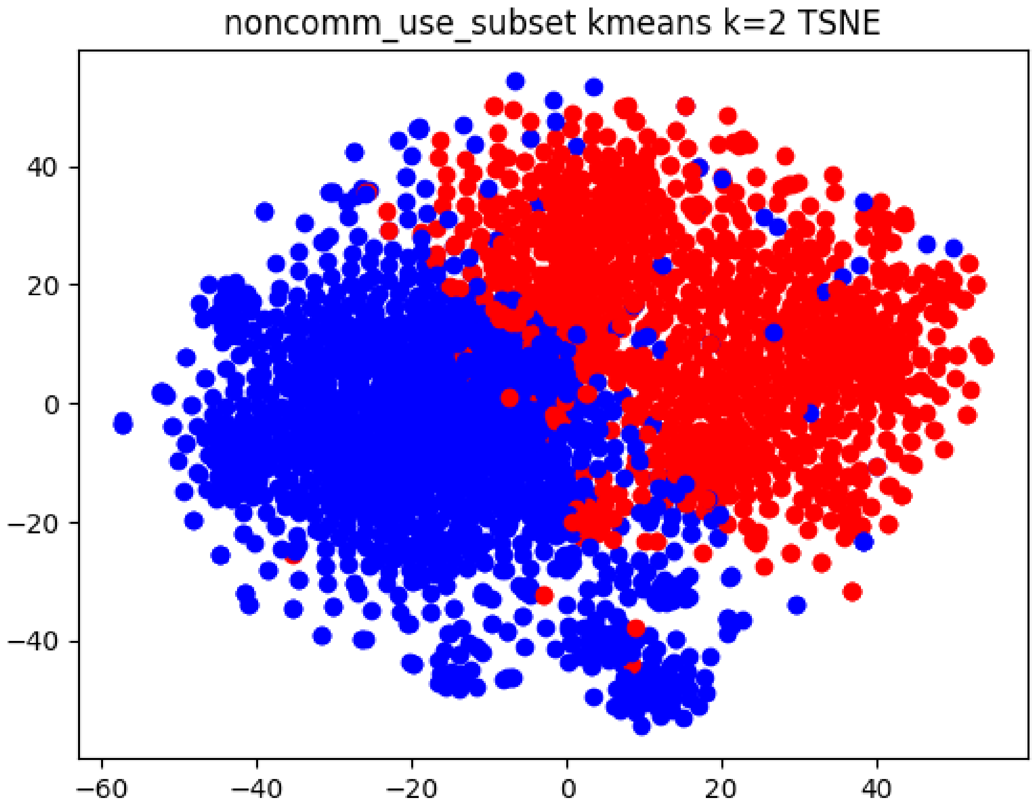
## noncomm_use_subeset

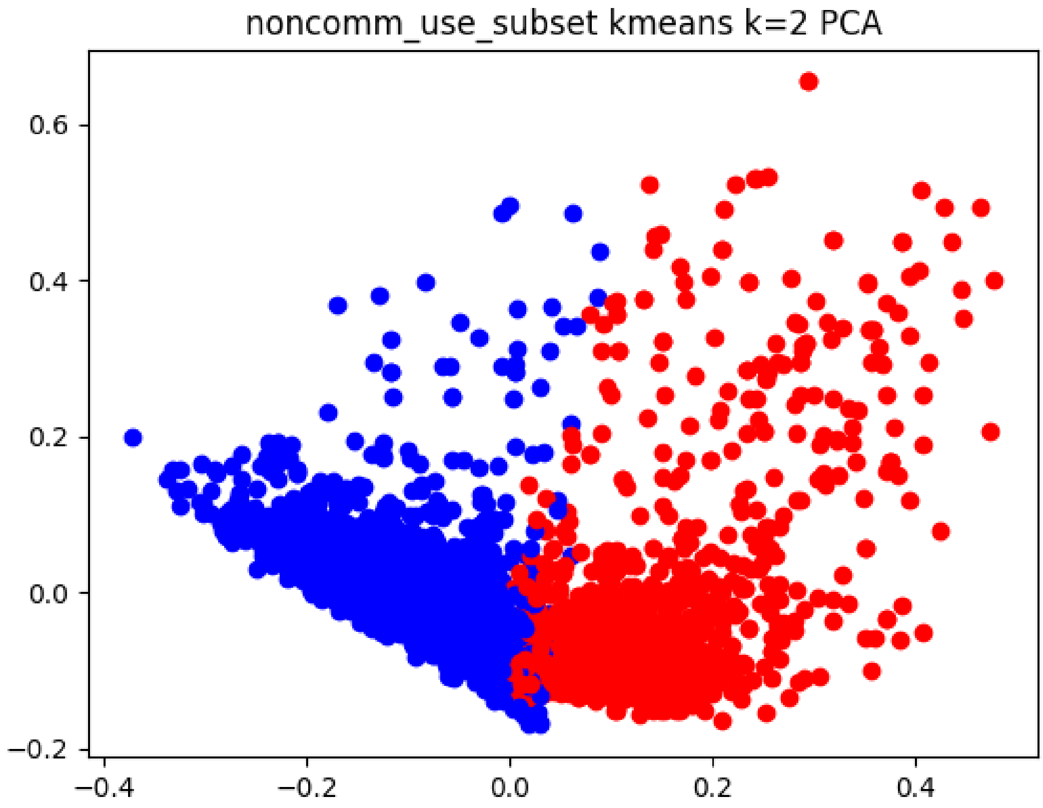

noncomm_use_subset kmeans k=1 TSNE
40
20
0
−20
−40
−60
−40
−20
0
20
40
60


noncomm_use_subset kmeans k=1 PCA
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
0.4

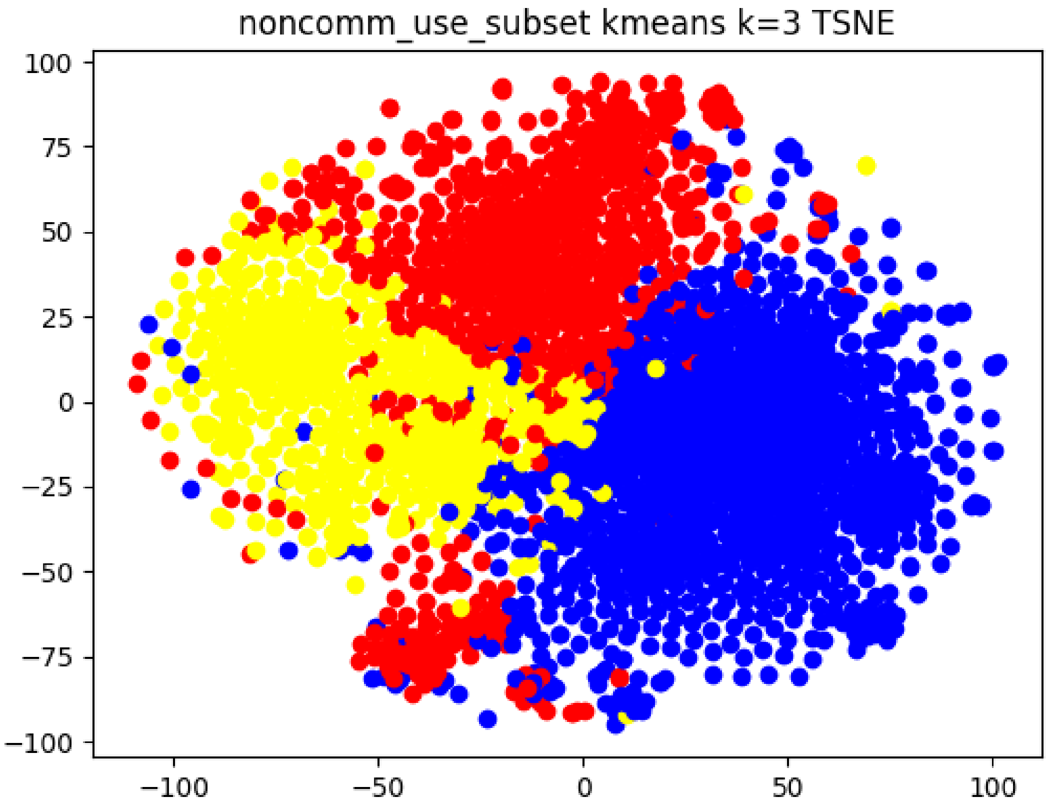


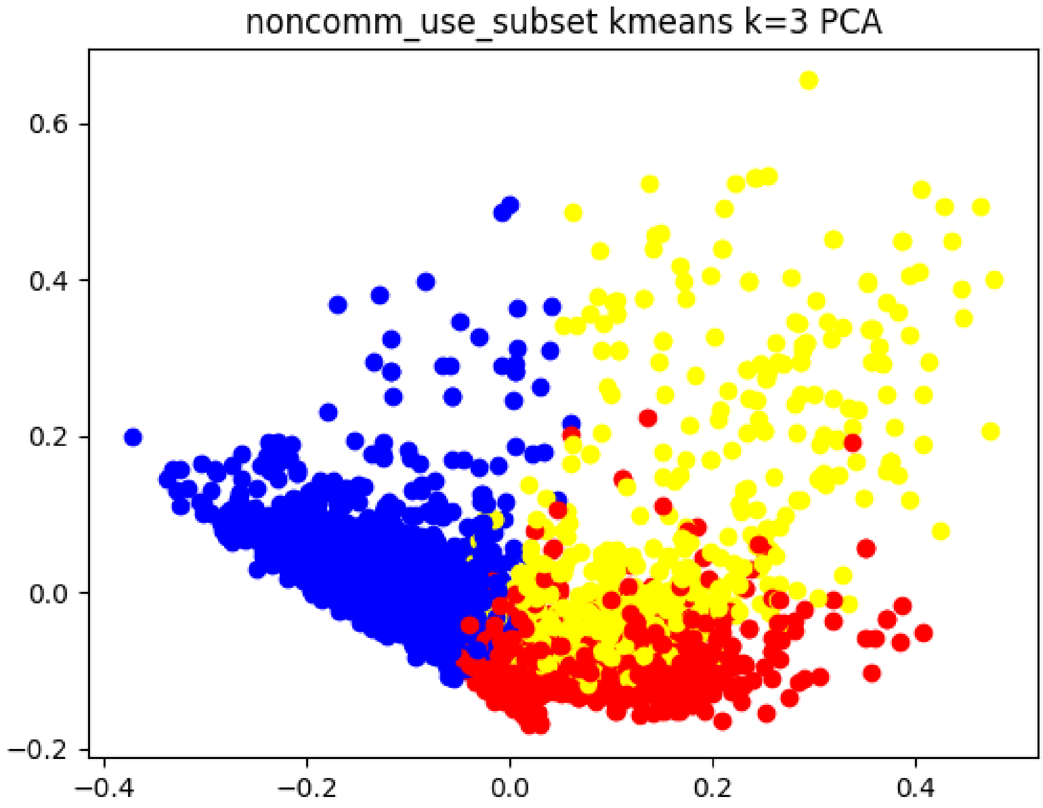
```
k = 1:
file counts in each cluster:
0    2584
Name: label, dtype: int64

cluster 0 top 10 words:['cell']  ['patient']  ['infect']  ['use']  ['studi']  ['protein']  ['viru']  ['case']  ['diseas']  ['merscov']
>>>
```

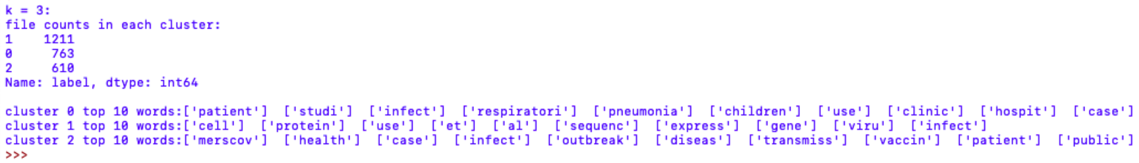




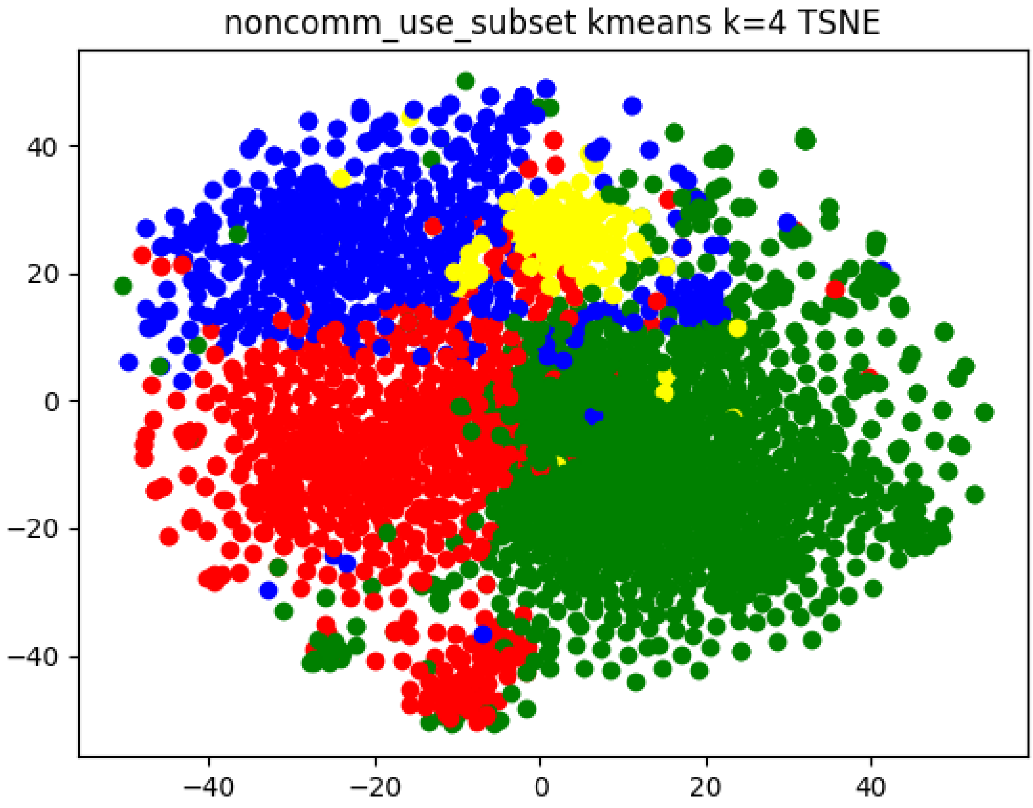
```
k = 2:
file counts in each cluster:
1    1424
0    1160
Name: label, dtype: int64

cluster 0 top 10 words:['patient']  ['infect']  ['case']  ['health']  ['merscov']  ['diseas']  ['studi']  ['respiratori']  ['hospit']  ['use']
cluster 1 top 10 words:['cell']  ['protein']  ['use']  ['et']  ['al']  ['sequenc']  ['infect']  ['express']  ['viru']  ['gene']
>>>
```

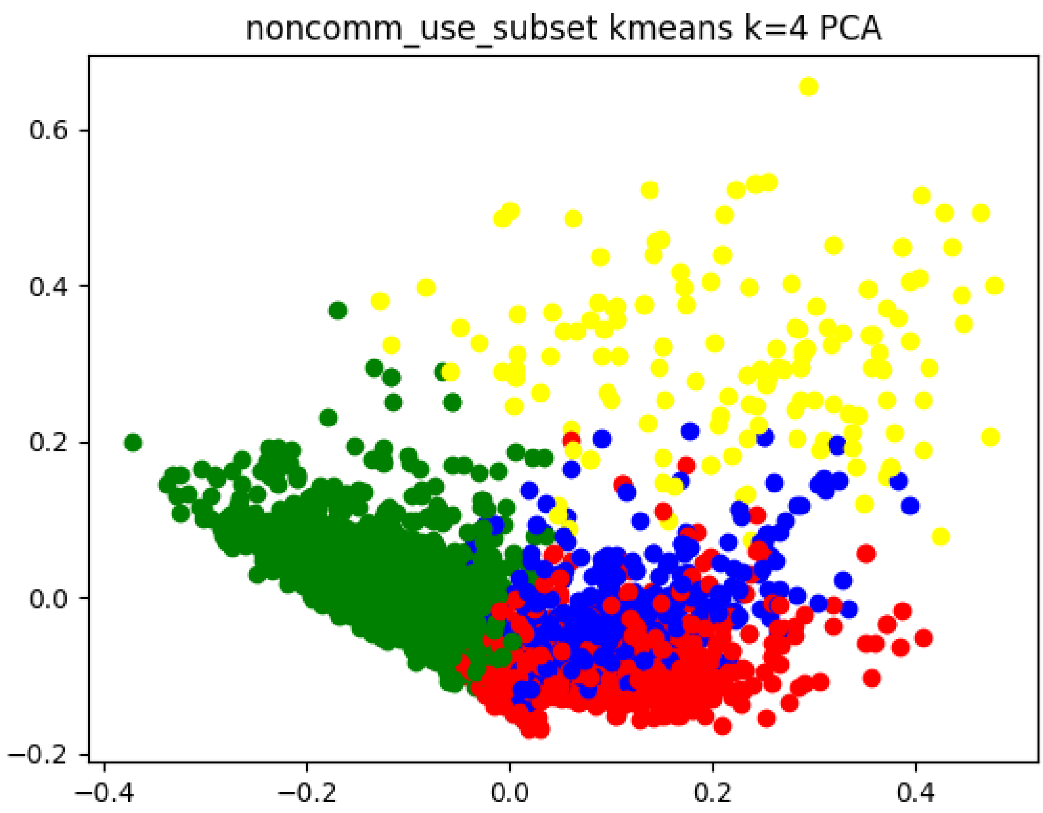

noncomm_use_subset kmeans k=3 TSNE
100
75
50
25
0
−25
−50
−75
−100
−100
−50
0
50
100


noncomm_use_subset kmeans k=3 PCA
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
0.4

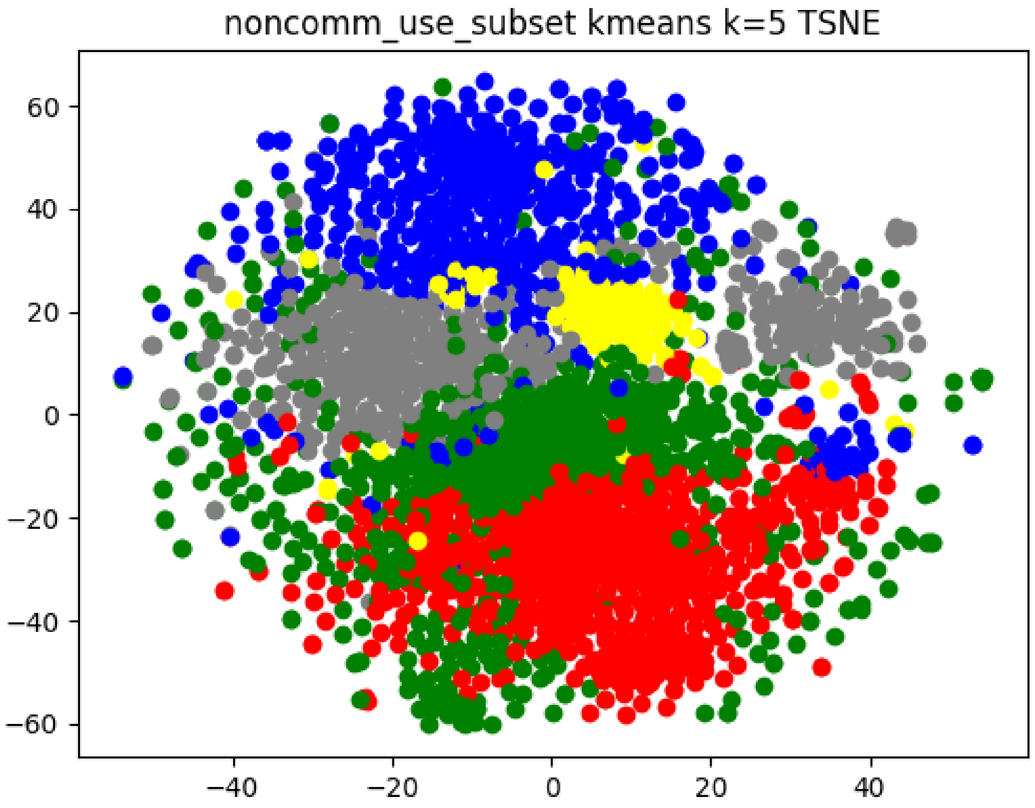


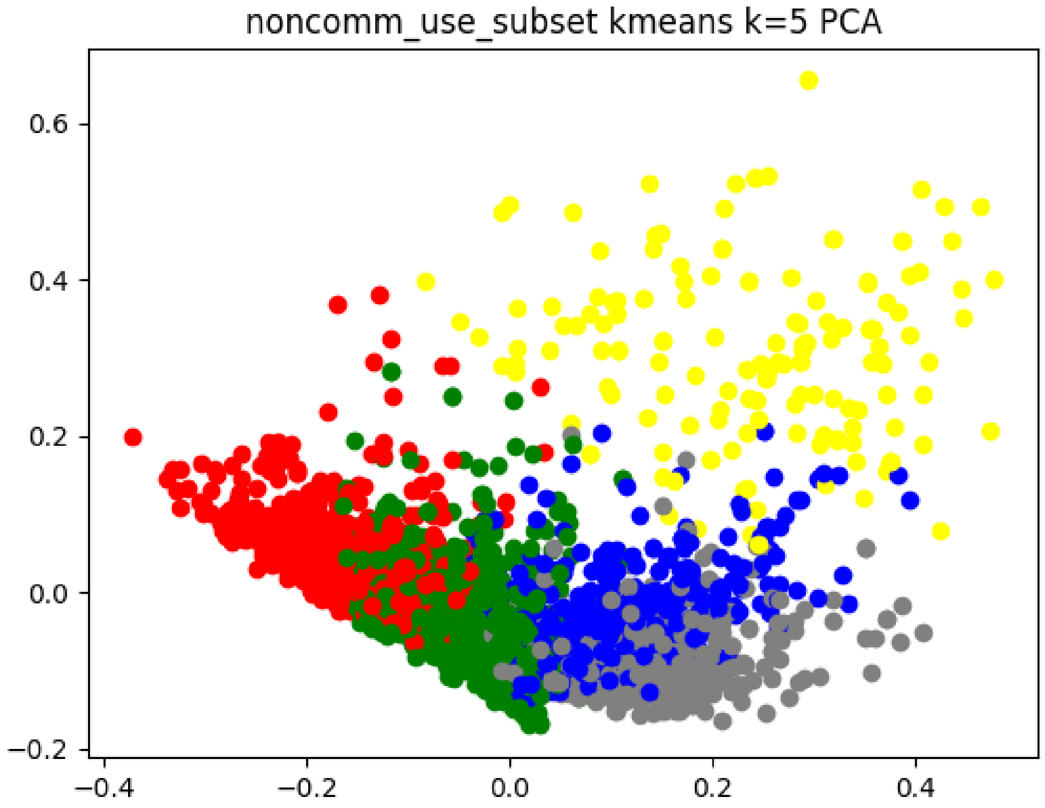
```
k = 3:
file counts in each cluster:
1    1211
0     763
2     610
Name: label, dtype: int64

cluster 0 top 10 words:['patient']  ['studi']  ['infect']  ['respiratori']  ['pneumonia']  ['children']  ['use']  ['clinic']  ['hospit']  ['case']
cluster 1 top 10 words:['cell']  ['protein']  ['use']  ['et']  ['al']  ['sequenc']  ['express']  ['gene']  ['viru']  ['infect']
cluster 2 top 10 words:['merscov']  ['health']  ['case']  ['infect']  ['outbreak']  ['diseas']  ['transmiss']  ['vaccin']  ['patient']  ['public']
>>>
```

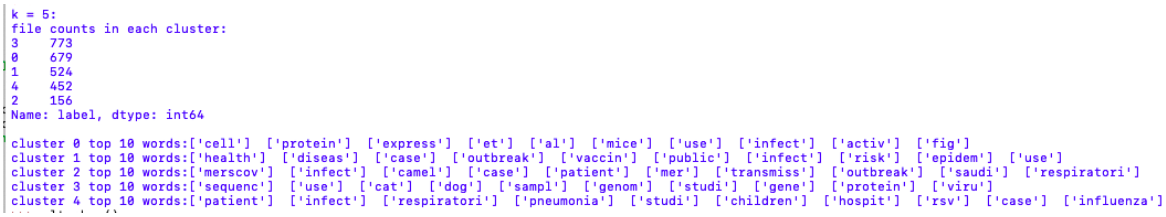




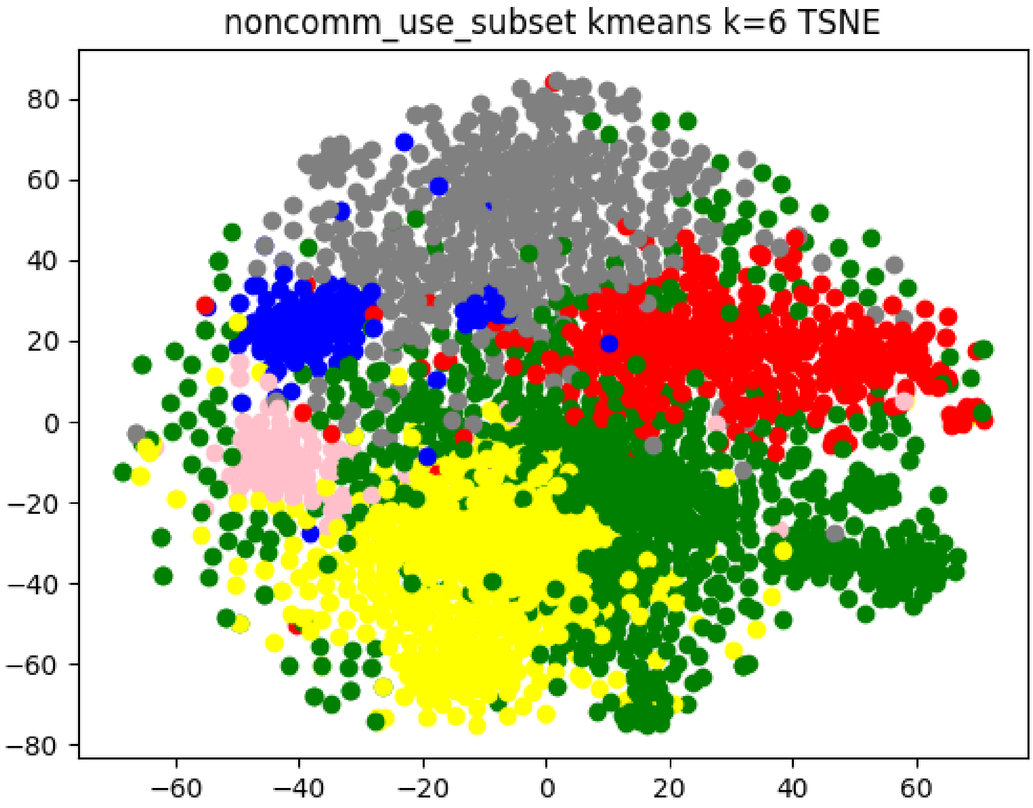
```
k = 4:
file counts in each cluster:
3    1217
0     697
1     512
2     158
Name: label, dtype: int64

cluster 0 top 10 words:['patient']  ['studi']  ['infect']  ['respiratori']  ['pneumonia']  ['children']  ['case']  ['hospit']  ['use']  ['clinic']
cluster 1 top 10 words:['health']  ['diseas']  ['case']  ['outbreak']  ['vaccin']  ['public']  ['infect']  ['epidem']  ['risk']  ['use']
cluster 2 top 10 words:['merscov']  ['infect']  ['camel']  ['case']  ['patient']  ['mer']  ['transmiss']  ['outbreak']  ['saudi']  ['respiratori']
cluster 3 top 10 words:['cell']  ['protein']  ['use']  ['et']  ['al']  ['sequenc']  ['express']  ['gene']  ['viru']  ['infect']
>>>
```

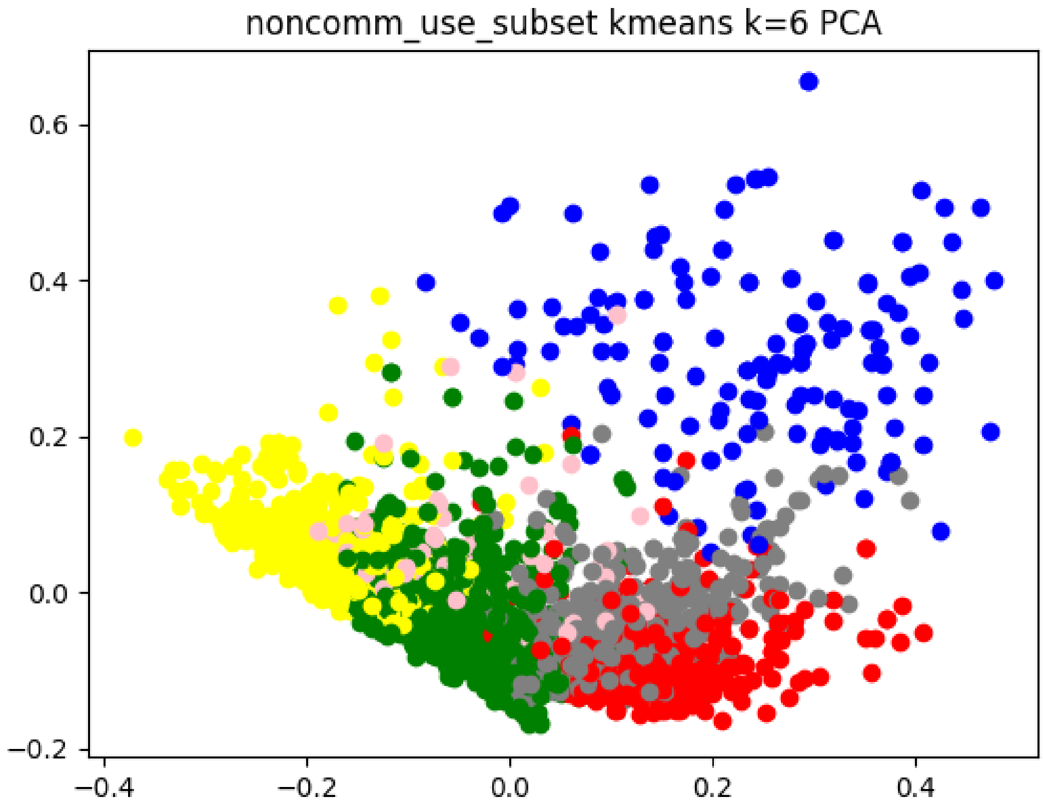

noncomm_use_subset kmeans k=5 TSNE
60
40
20
0
−20
−40
−60
−40
−20
0
20
40


noncomm_use_subset kmeans k=5 PCA
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
0.4

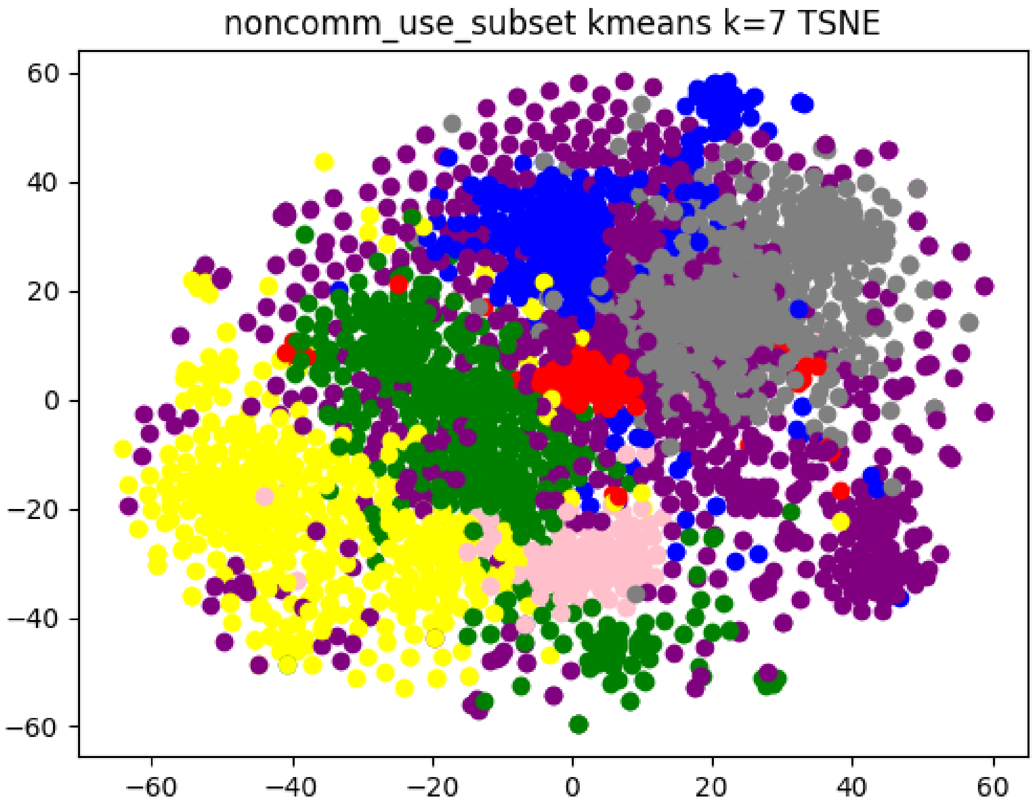


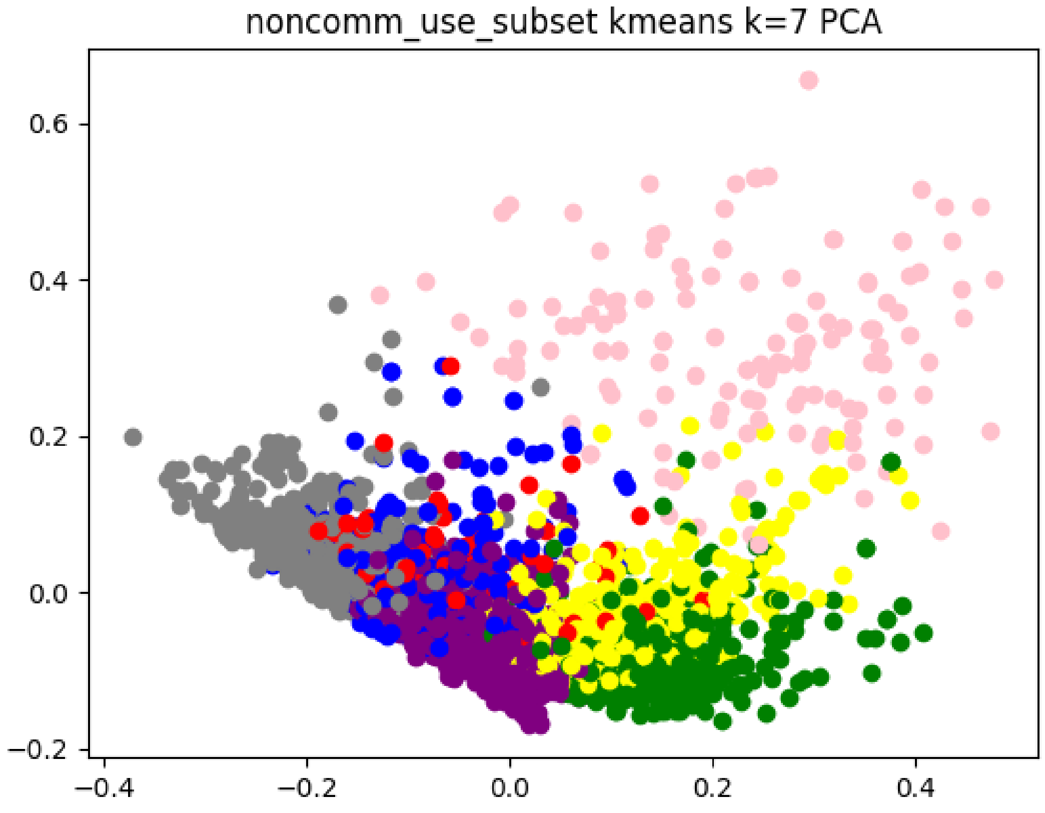
```
k = 5:
file counts in each cluster:
3    773
0    679
1    524
4    452
2    156
Name: label, dtype: int64

cluster 0 top 10 words:['cell']  ['protein']  ['express']  ['et']  ['al']  ['mice']  ['use']  ['infect']  ['activ']  ['fig']
cluster 1 top 10 words:['health']  ['diseas']  ['case']  ['outbreak']  ['vaccin']  ['public']  ['infect']  ['risk']  ['epidem']  ['use']
cluster 2 top 10 words:['merscov']  ['infect']  ['camel']  ['case']  ['patient']  ['mer']  ['transmiss']  ['outbreak']  ['saudi']  ['respiratori']
cluster 3 top 10 words:['sequenc']  ['use']  ['cat']  ['dog']  ['sampl']  ['genom']  ['studi']  ['gene']  ['protein']  ['viru']
cluster 4 top 10 words:['patient']  ['infect']  ['respiratori']  ['pneumonia']  ['studi']  ['children']  ['hospit']  ['rsv']  ['case']  ['influenza']
```

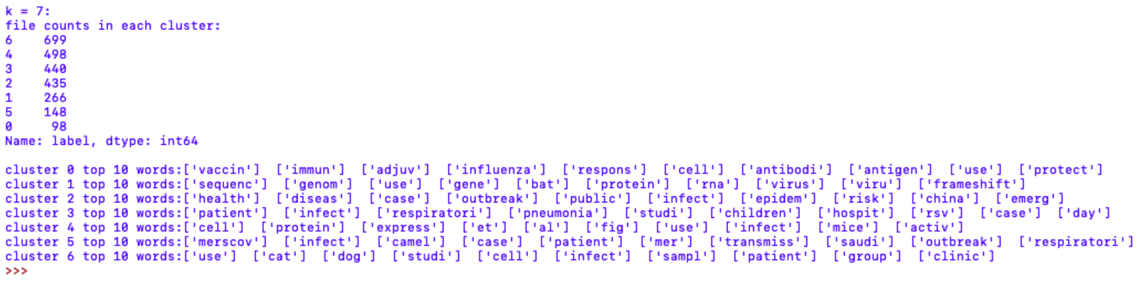




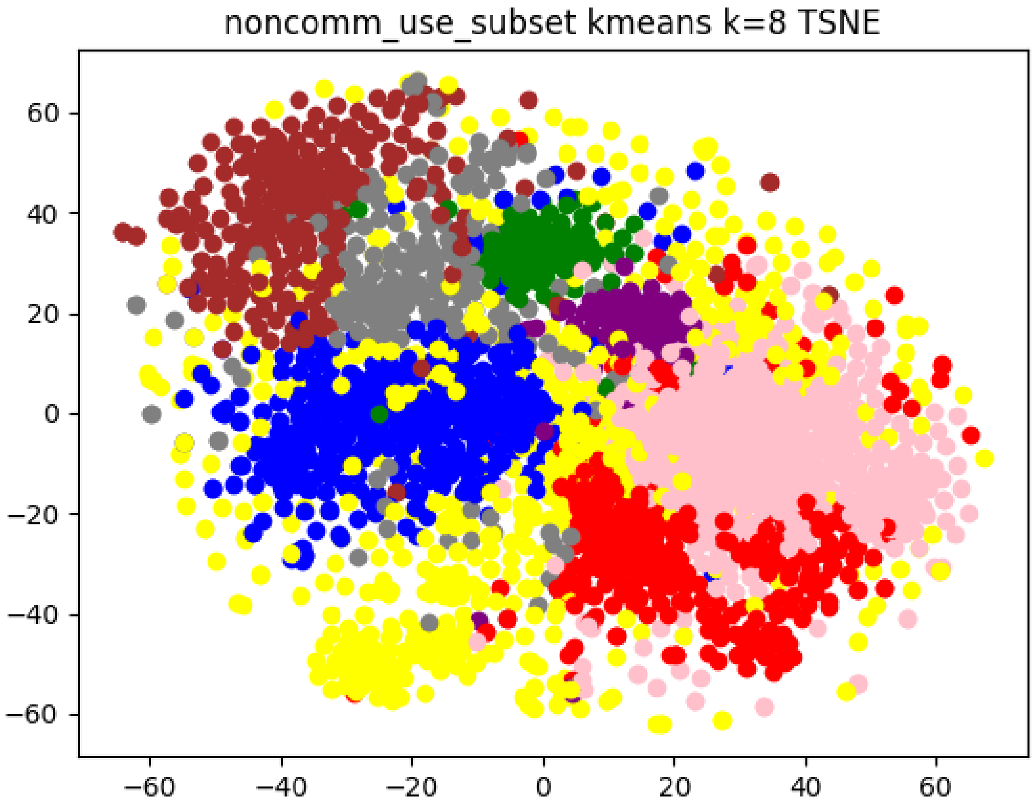
```
k = 6:
file counts in each cluster:
3    855
2    575
4    495
0    406
1    153
5    100
Name: label, dtype: int64

cluster 0 top 10 words:['patient']  ['infect']  ['respiratori']  ['pneumonia']  ['children']  ['studi']  ['rsv']  ['hospit']  ['influenza']  ['case']
cluster 1 top 10 words:['merscov']  ['infect']  ['camel']  ['case']  ['patient']  ['mer']  ['transmiss']  ['outbreak']  ['saudi']  ['respiratori']
cluster 2 top 10 words:['cell']  ['protein']  ['express']  ['et']  ['al']  ['mice']  ['use']  ['infect']  ['fig']  ['activ']
cluster 3 top 10 words:['sequenc']  ['use']  ['cat']  ['dog']  ['studi']  ['sampl']  ['genom']  ['gene']  ['protein']  ['viru']
cluster 4 top 10 words:['health']  ['diseas']  ['case']  ['outbreak']  ['public']  ['infect']  ['risk']  ['patient']  ['epidem']  ['use']
cluster 5 top 10 words:['vaccin']  ['immun']  ['adjuv']  ['respons']  ['antibodi']  ['cell']  ['antigen']  ['protect']  ['use']  ['epitop']
>>>
```

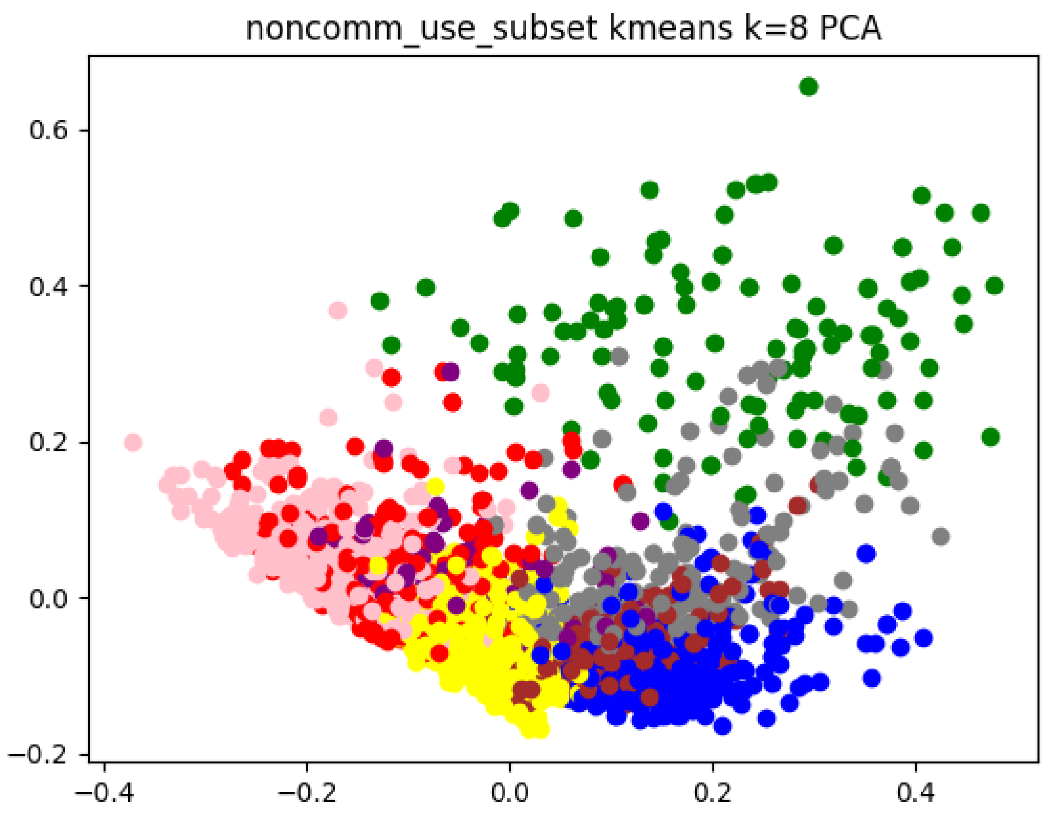

noncomm_use_subset kmeans k=7 TSNE
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60

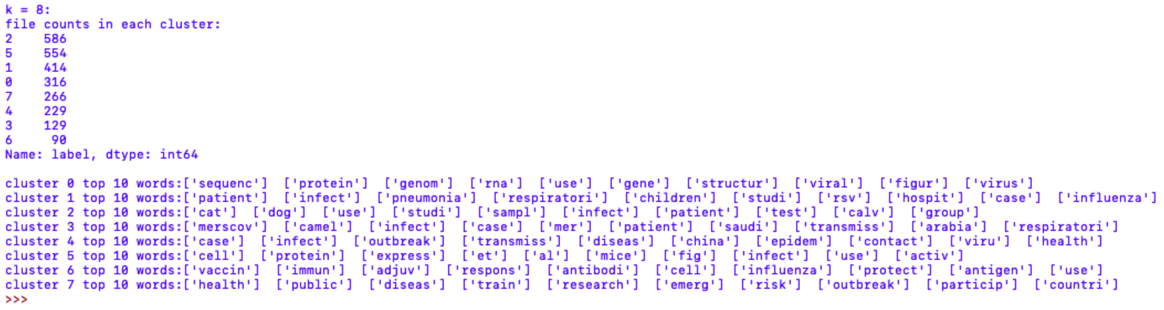

noncomm_use_subset kmeans k=7 PCA
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
0.4



```
k = 7:
file counts in each cluster:
6    699
4    498
3    440
2    435
1    266
5    148
0     98
Name: label, dtype: int64

cluster 0 top 10 words:['vaccin']  ['immun']  ['adjuv']  ['influenza']  ['respons']  ['cell']  ['antibodi']  ['antigen']  ['use']  ['protect']
cluster 1 top 10 words:['sequenc']  ['genom']  ['use']  ['gene']  ['bat']  ['protein']  ['rna']  ['virus']  ['viru']  ['frameshift']
cluster 2 top 10 words:['health']  ['diseas']  ['case']  ['outbreak']  ['public']  ['infect']  ['epidem']  ['risk']  ['china']  ['emerg']
cluster 3 top 10 words:['patient']  ['infect']  ['respiratori']  ['pneumonia']  ['studi']  ['children']  ['hospit']  ['rsv']  ['case']  ['day']
cluster 4 top 10 words:['cell']  ['protein']  ['express']  ['et']  ['al']  ['fig']  ['use']  ['infect']  ['mice']  ['activ']
cluster 5 top 10 words:['merscov']  ['infect']  ['camel']  ['case']  ['patient']  ['mer']  ['transmiss']  ['saudi']  ['outbreak']  ['respiratori']
cluster 6 top 10 words:['use']  ['cat']  ['dog']  ['studi']  ['cell']  ['infect']  ['sampl']  ['patient']  ['group']  ['clinic']
>>>
```







```
k = 8:
file counts in each cluster:
2    586
5    554
1    414
0    316
7    266
4    229
3    129
6     90
Name: label, dtype: int64

cluster 0 top 10 words:['sequenc']  ['protein']  ['genom']  ['rna']  ['use']  ['gene']  ['structur']  ['viral']  ['figur']  ['virus']
cluster 1 top 10 words:['patient']  ['infect']  ['pneumonia']  ['respiratori']  ['children']  ['studi']  ['rsv']  ['hospit']  ['case']  ['influenza']
cluster 2 top 10 words:['cat']  ['dog']  ['use']  ['studi']  ['sampl']  ['infect']  ['patient']  ['test']  ['calv']  ['group']
cluster 3 top 10 words:['merscov']  ['camel']  ['infect']  ['case']  ['mer']  ['patient']  ['saudi']  ['transmiss']  ['arabia']  ['respiratori']
cluster 4 top 10 words:['case']  ['infect']  ['outbreak']  ['transmiss']  ['diseas']  ['china']  ['epidem']  ['contact']  ['viru']  ['health']
cluster 5 top 10 words:['cell']  ['protein']  ['express']  ['et']  ['al']  ['mice']  ['fig']  ['infect']  ['use']  ['activ']
cluster 6 top 10 words:['vaccin']  ['immun']  ['adjuv']  ['respons']  ['antibodi']  ['cell']  ['influenza']  ['protect']  ['antigen']  ['use']
cluster 7 top 10 words:['health']  ['public']  ['diseas']  ['train']  ['research']  ['emerg']  ['risk']  ['outbreak']  ['particip']  ['countri']
>>>
```

In above graphs, K = 4 appears to be optimal, and the following words are recovered:

Cluster 0: patient, study, infect, respiratory, pneumonia, children, case, hospital, use, clinic

Cluster 1: health, disease, case, outbreak, vaccine, public, infect, epidemic, risk, use

Cluster 2: Merscov, infect, camel, case, patient, mer, transmission, outbreak, Saudi, respiratory

Cluster 3: cell, protein, use, et, al, sequence, express, gene, virus, infect

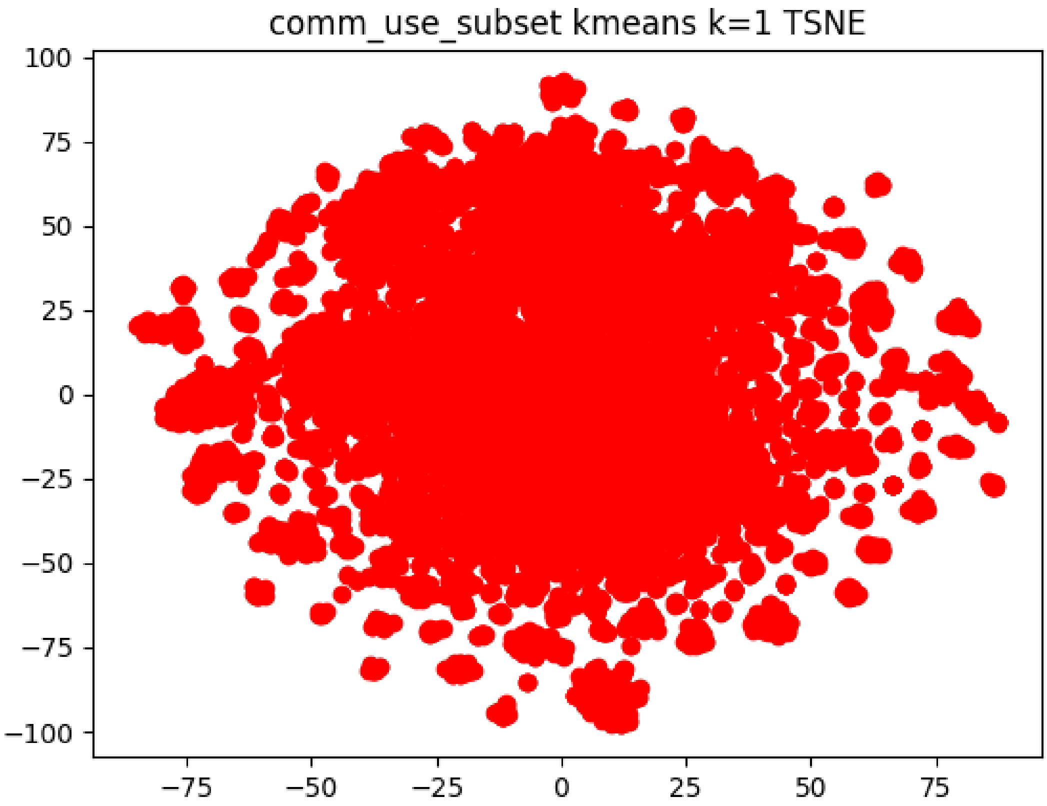
Cluster0 outlines concerns about the risk of novel Coronavirus infection in a pneumonia patient, with concern for children cases among them.

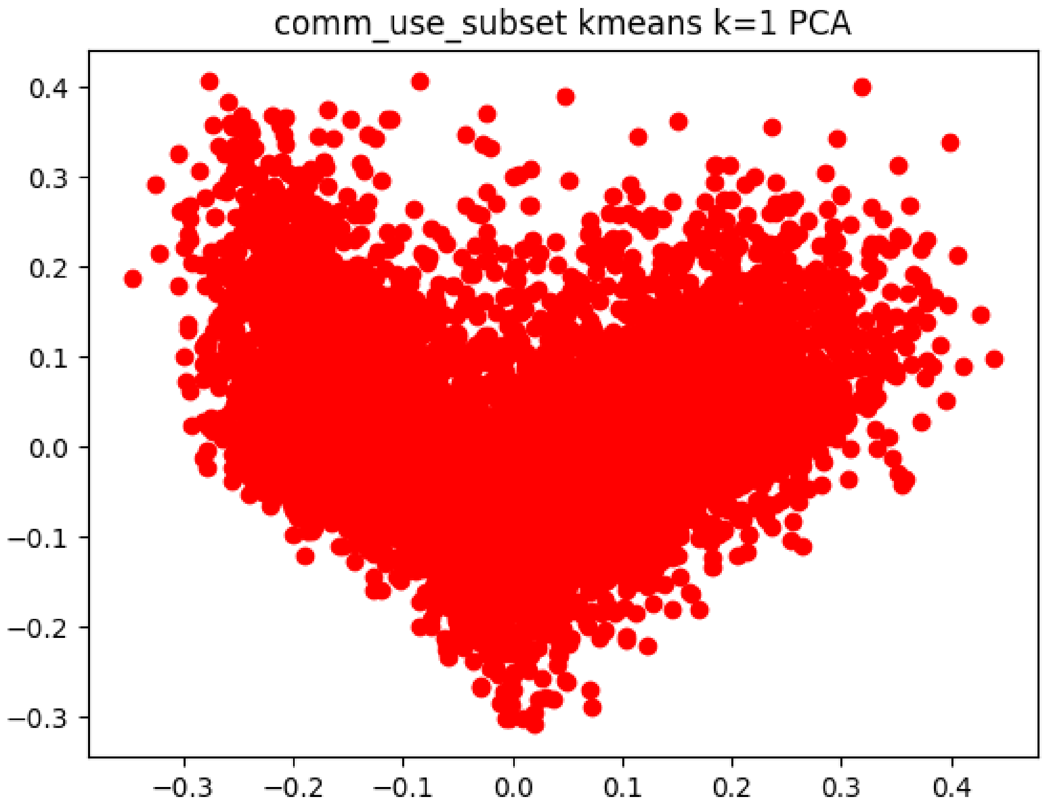
Cluster1 outlines the public health risks of an outbreak and sets out to develop a vaccine.

Cluster2 outlines the outbreak and spread of pneumonia and the virus in the Saudi region, with camels as a possible cause.

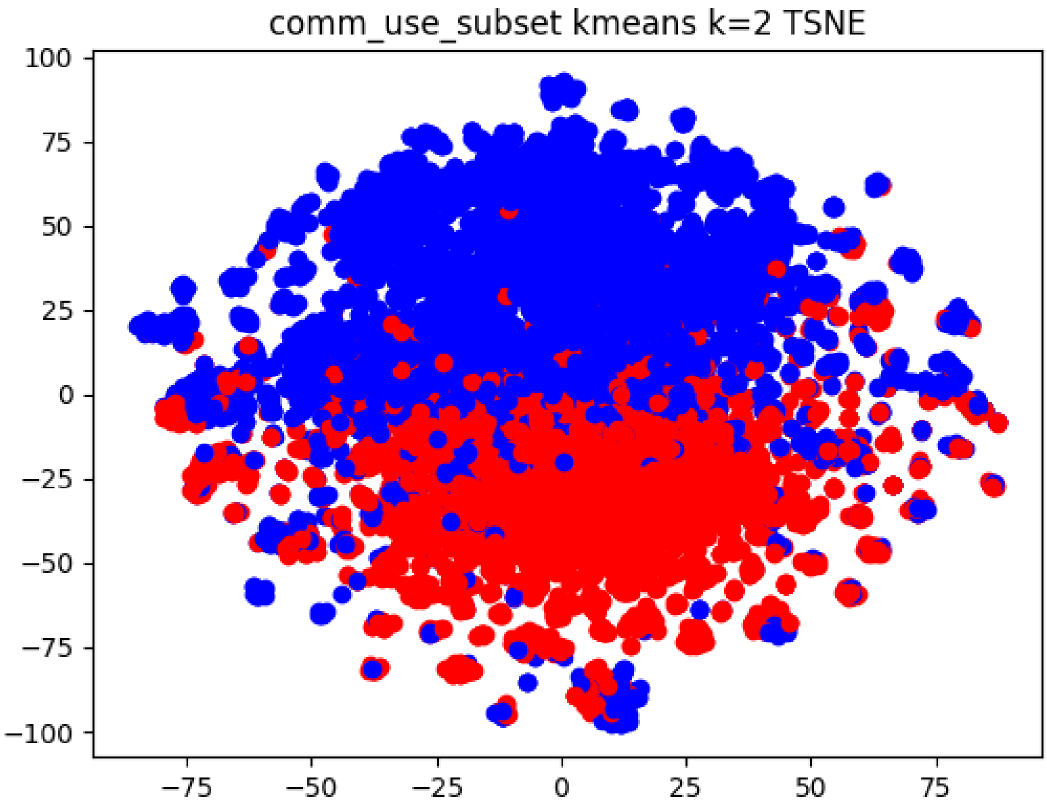
Cluster3 outlines the novel Coronavirus biological study, including the analysis of gene chains, proteins and DNA in VIUR region.

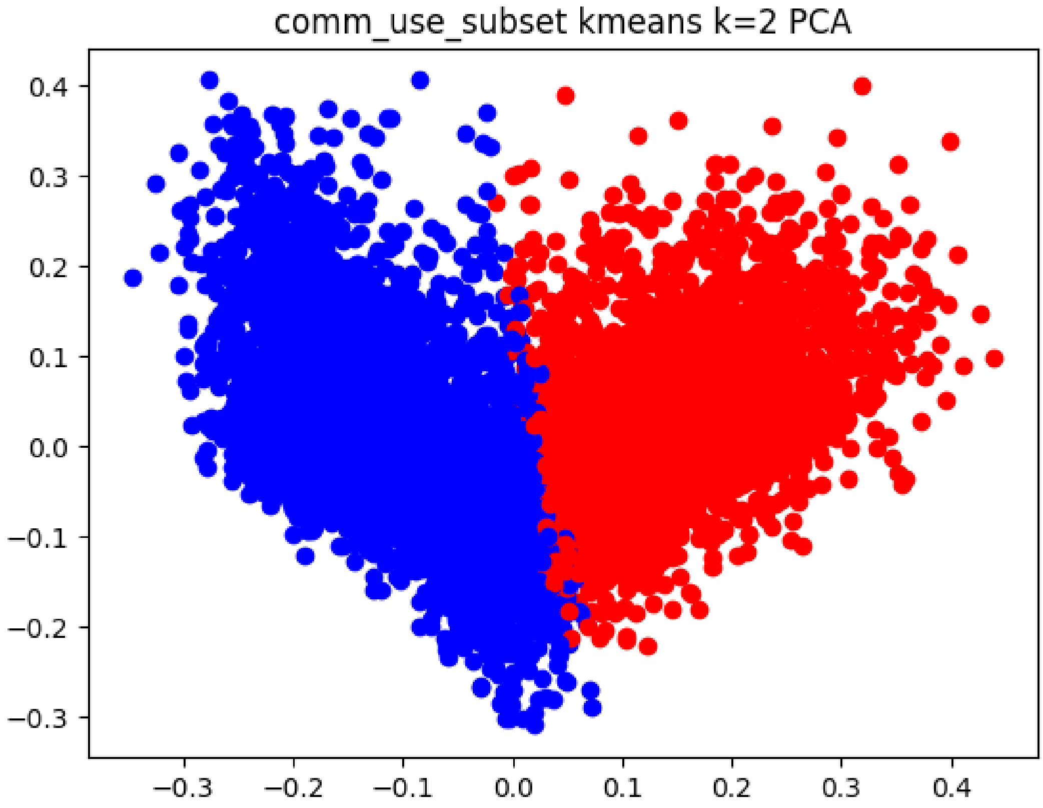
## comm_use_subset



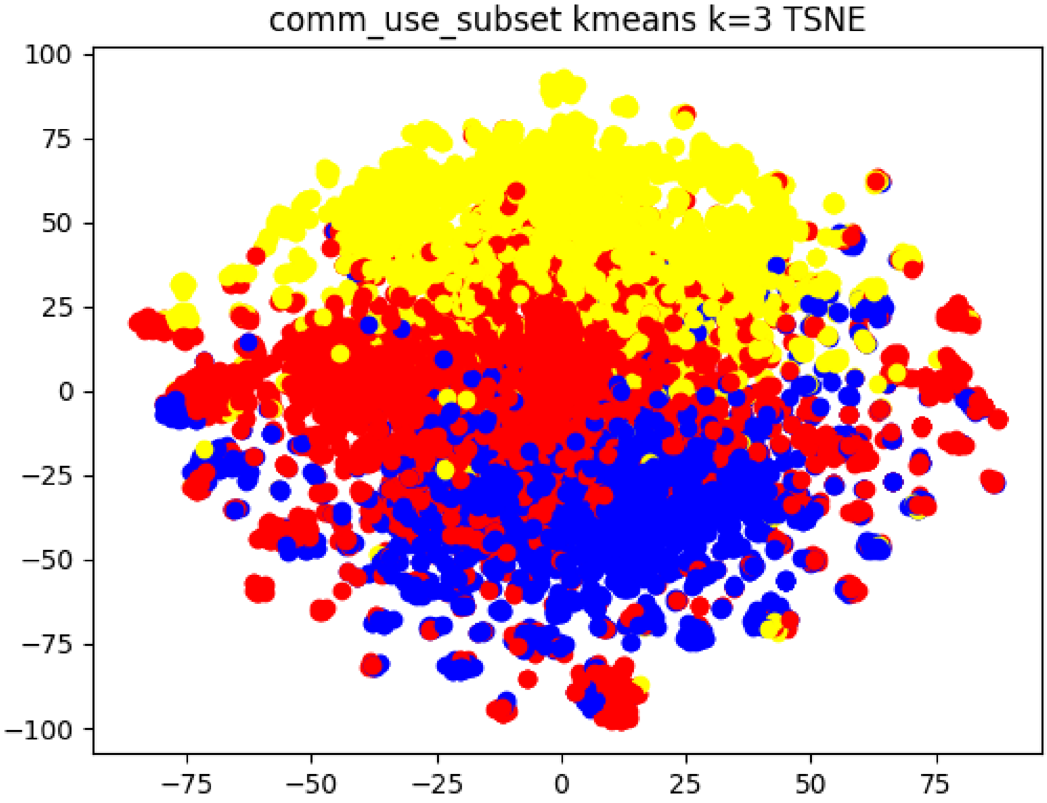


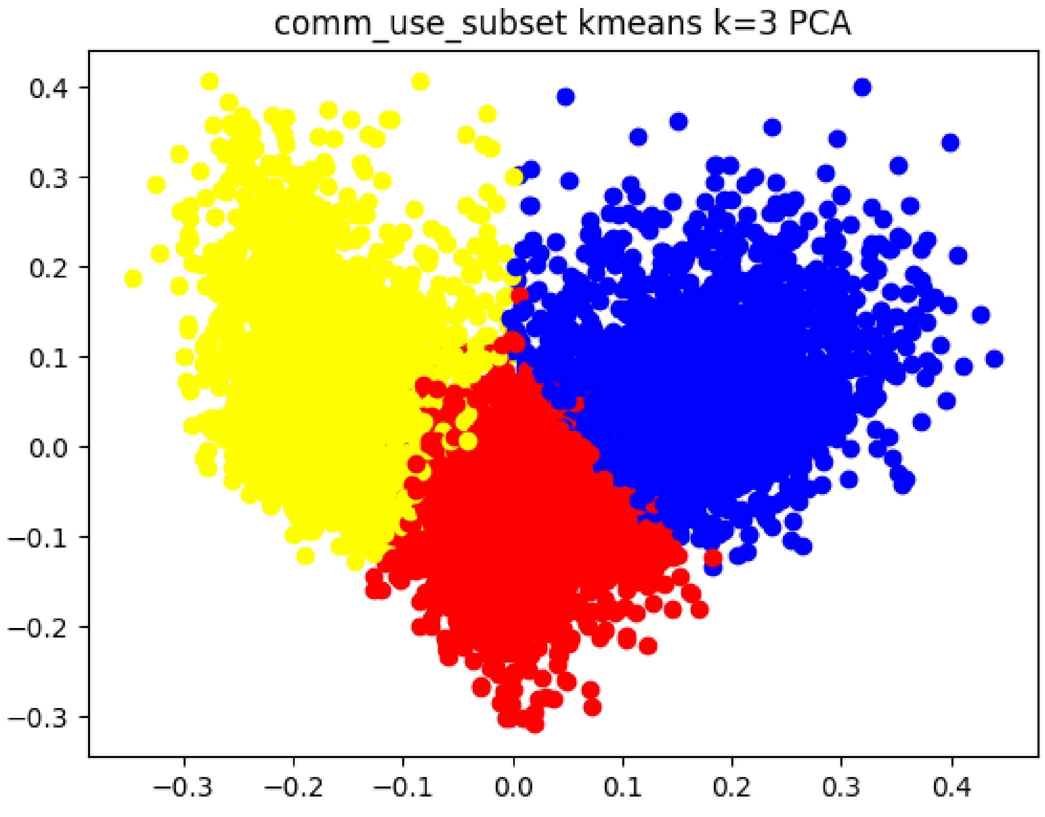
```
k = 1:
file counts in each cluster:
0    9918
Name: label, dtype: int64

cluster 0 top 10 words:['cell']  ['infect']  ['use']  ['protein']  ['viru']  ['patient']  ['studi']  ['viral']  ['sequenc']  ['express']
```

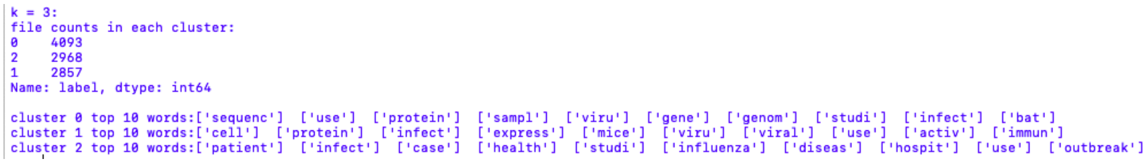

comm_use_subset kmeans k=2 TSNE
100
75
50
25
0
−25
−50
−75
−100
−75
−50
−25
0
25
50
75


comm_use_subset kmeans k=2 PCA
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4

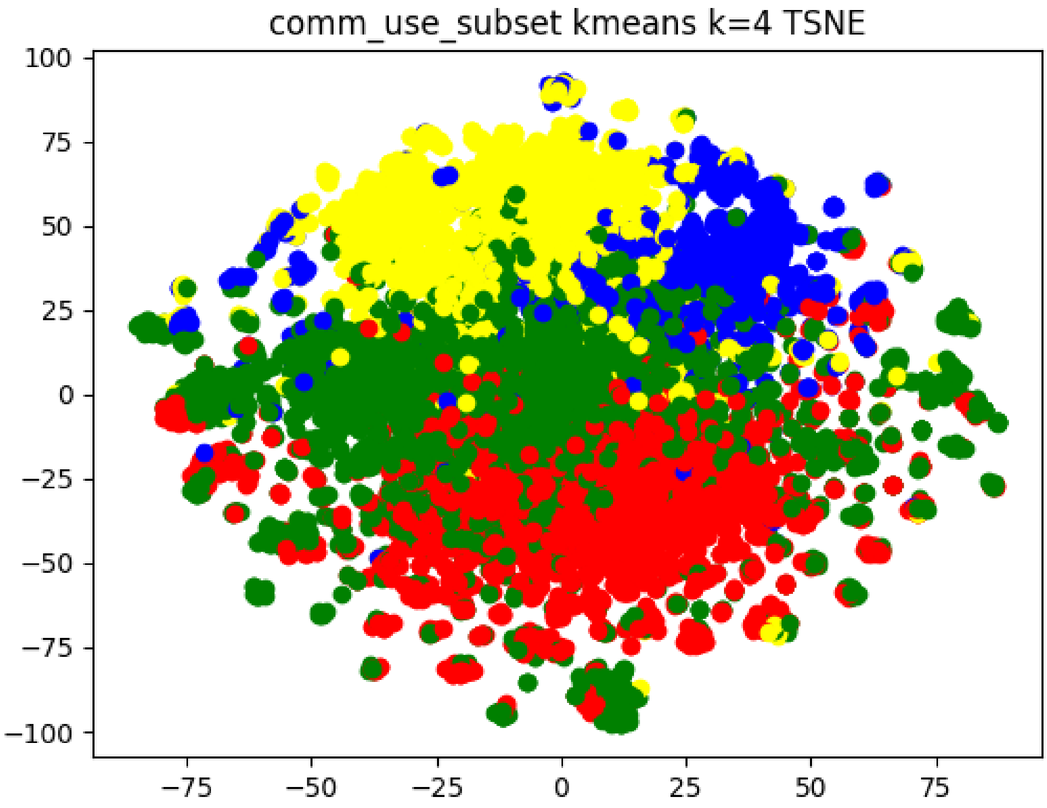
```
k = 2:
file counts in each cluster:
1    5851
0    4067
Name: label, dtype: int64

cluster 0 top 10 words:['cell']  ['protein']  ['infect']  ['express']  ['use']  ['viru']  ['mice']  ['viral']  ['activ']  ['fig']
cluster 1 top 10 words:['patient']  ['use']  ['infect']  ['studi']  ['case']  ['sequenc']  ['diseas']  ['health']  ['sampl']  ['viru']
```

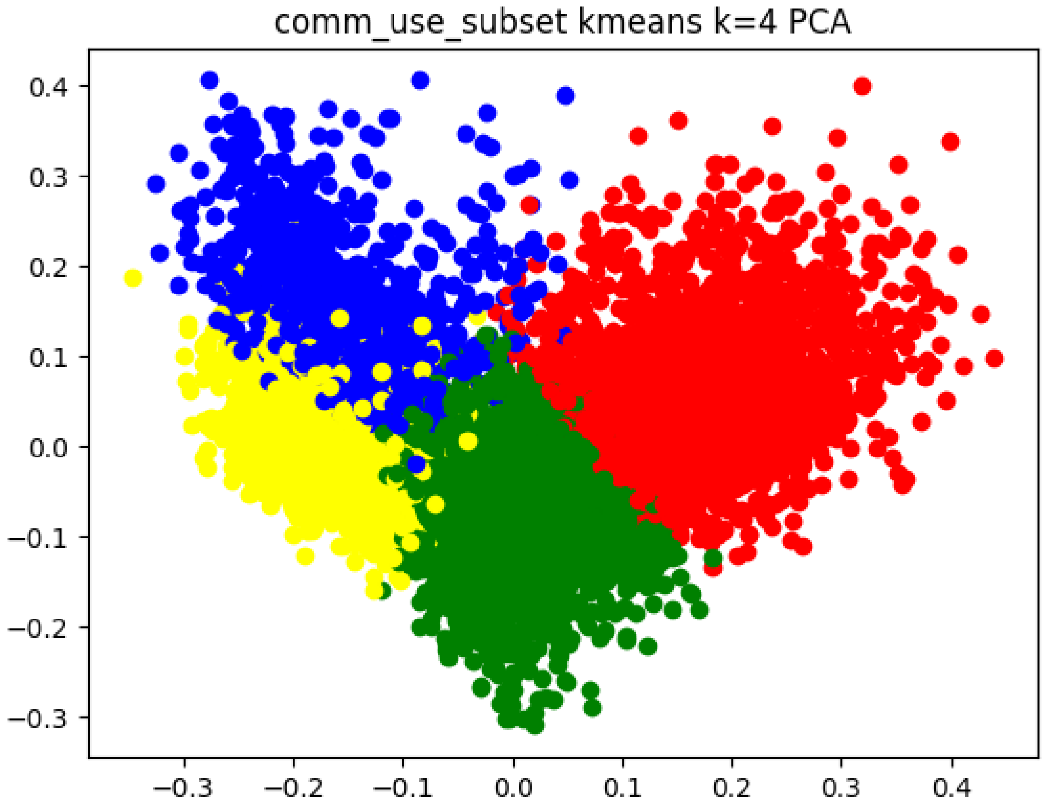


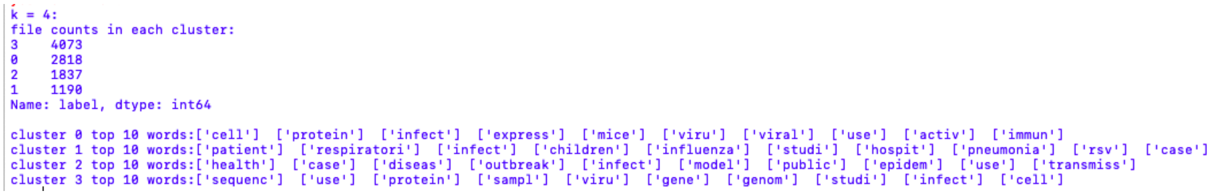




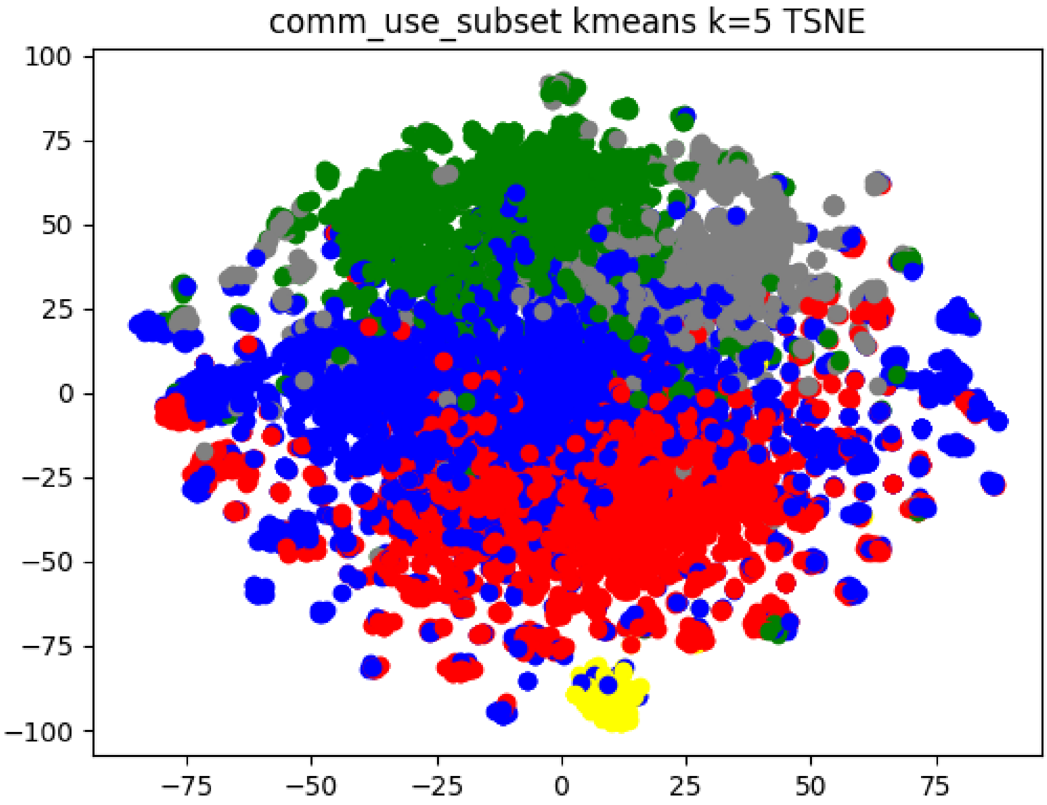
```
k = 3:
file counts in each cluster:
0    4093
2    2968
1    2857
Name: label, dtype: int64

cluster 0 top 10 words:['sequenc']  ['use']  ['protein']  ['sampl']  ['viru']  ['gene']  ['genom']  ['studi']  ['infect']  ['bat']
cluster 1 top 10 words:['cell']  ['protein']  ['infect']  ['express']  ['mice']  ['viru']  ['viral']  ['use']  ['activ']  ['immun']
cluster 2 top 10 words:['patient']  ['infect']  ['case']  ['health']  ['studi']  ['influenza']  ['diseas']  ['hospit']  ['use']  ['outbreak']
```

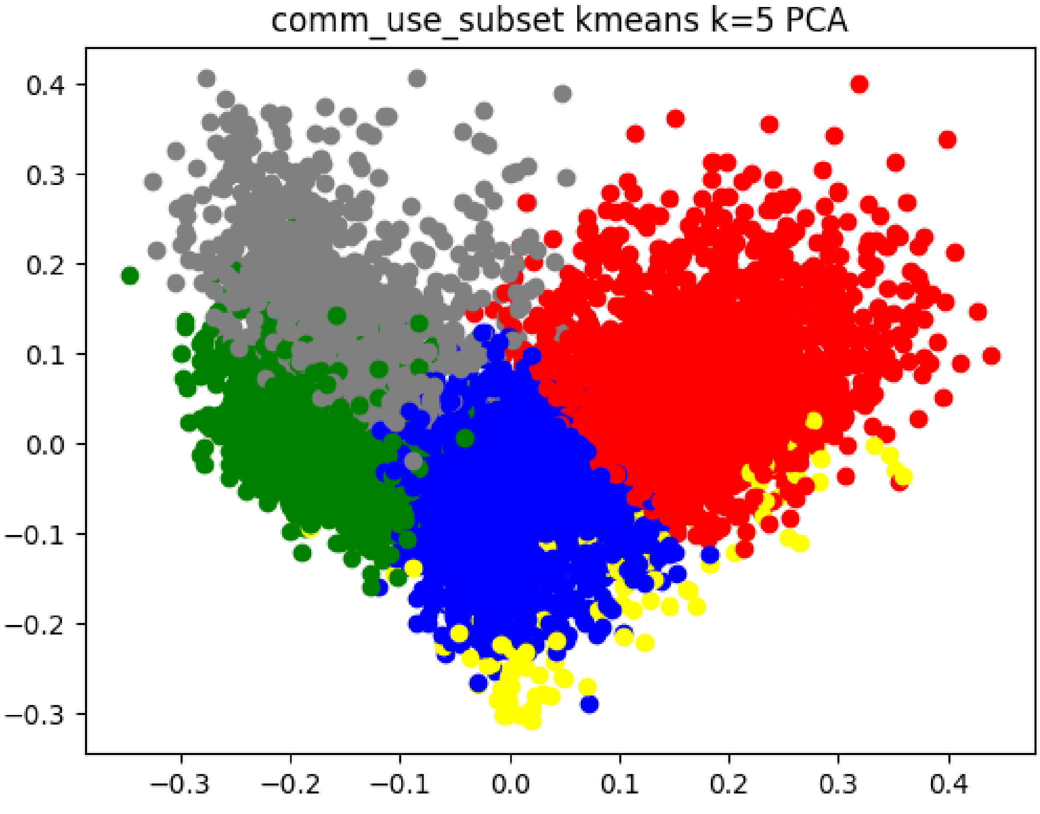

comm_use_subset kmeans k=4 TSNE
100
75
50
25
0
−25
−50
−75
−100
−75
−50
−25
0
25
50
75

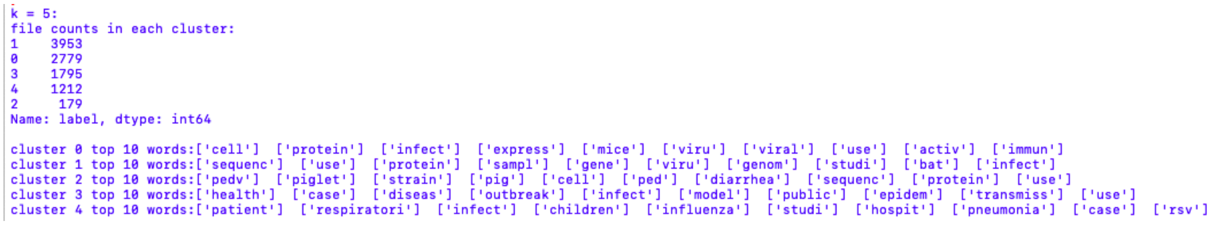

comm_use_subset kmeans k=4 PCA
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4



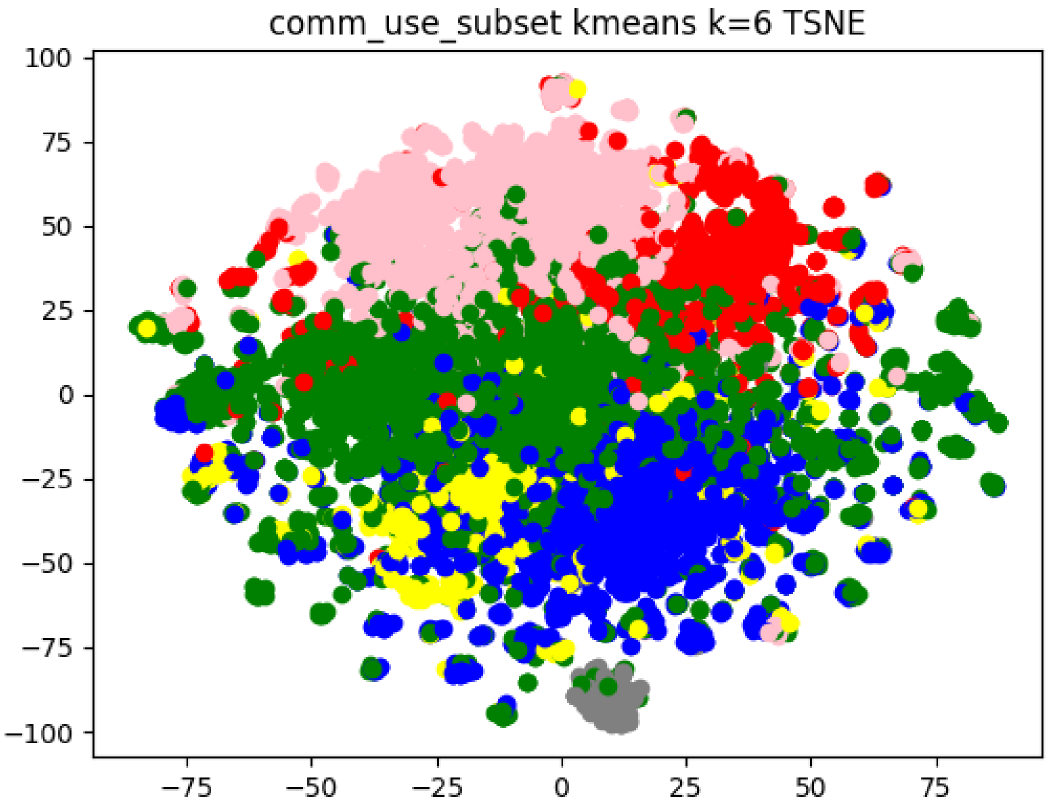
```
k = 4:
file counts in each cluster:
3    4073
0    2818
2    1837
1    1190
Name: label, dtype: int64

cluster 0 top 10 words:['cell']  ['protein']  ['infect']  ['express']  ['mice']  ['viru']  ['viral']  ['use']  ['activ']  ['immun']
cluster 1 top 10 words:['patient']  ['respiratori']  ['infect']  ['children']  ['influenza']  ['studi']  ['hospit']  ['pneumonia']  ['rsv']  ['case']
cluster 2 top 10 words:['health']  ['case']  ['diseas']  ['outbreak']  ['infect']  ['model']  ['public']  ['epidem']  ['use']  ['transmiss']
cluster 3 top 10 words:['sequenc']  ['use']  ['protein']  ['sampl']  ['viru']  ['gene']  ['genom']  ['studi']  ['infect']  ['cell']
```

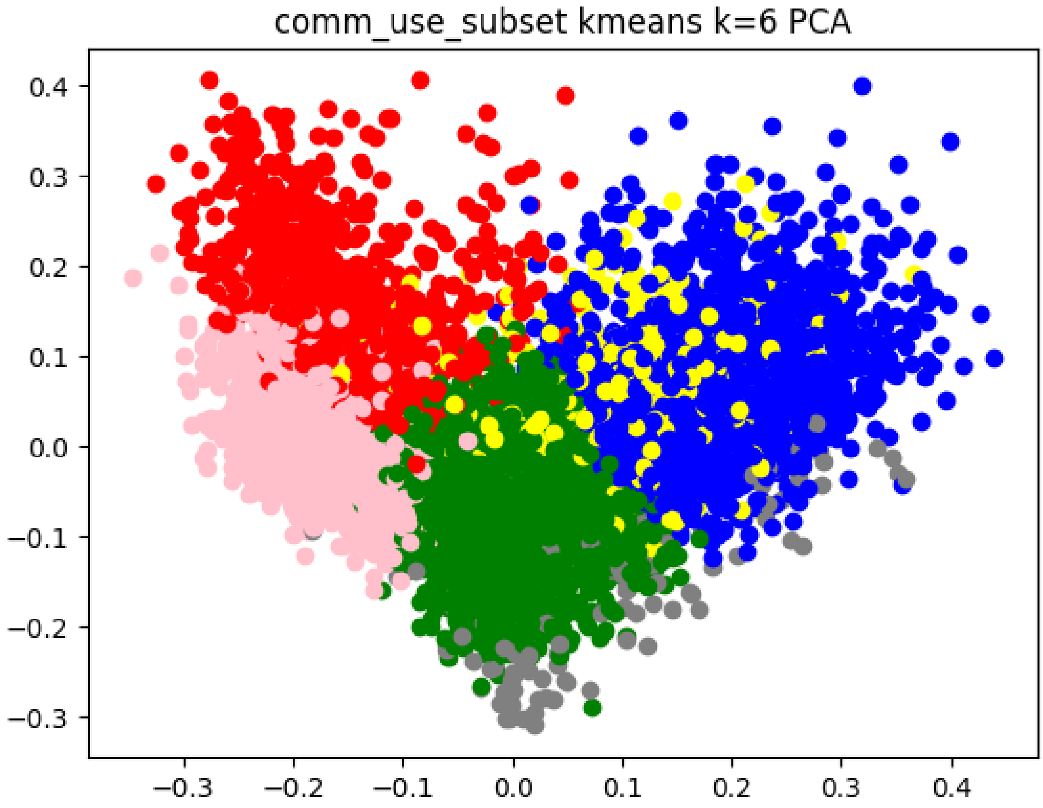


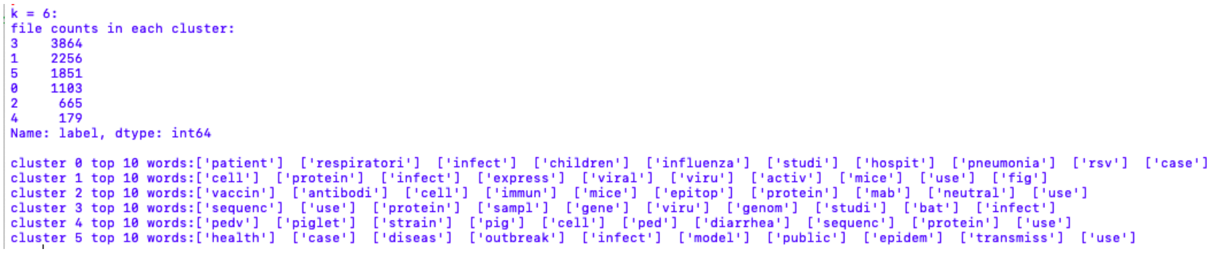




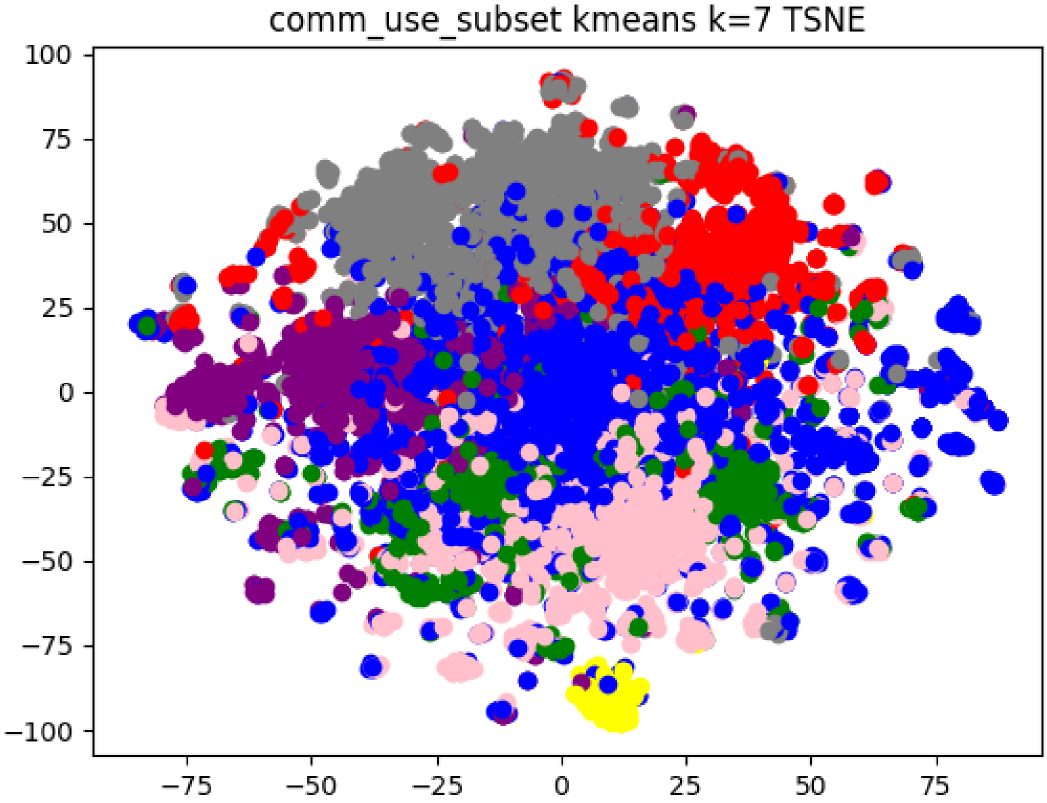
```
k = 5:
file counts in each cluster:
1    3953
0    2779
3    1795
4    1212
2     179
Name: label, dtype: int64

cluster 0 top 10 words:['cell']  ['protein']  ['infect']  ['express']  ['mice']  ['viru']  ['viral']  ['use']  ['activ']  ['immun']
cluster 1 top 10 words:['sequenc']  ['use']  ['protein']  ['sampl']  ['gene']  ['viru']  ['genom']  ['studi']  ['bat']  ['infect']
cluster 2 top 10 words:['pedv']  ['piglet']  ['strain']  ['pig']  ['cell']  ['ped']  ['diarrhea']  ['sequenc']  ['protein']  ['use']
cluster 3 top 10 words:['health']  ['case']  ['diseas']  ['outbreak']  ['infect']  ['model']  ['public']  ['epidem']  ['transmiss']  ['use']
cluster 4 top 10 words:['patient']  ['respiratori']  ['infect']  ['children']  ['influenza']  ['studi']  ['hospit']  ['pneumonia']  ['case']  ['rsv']
```

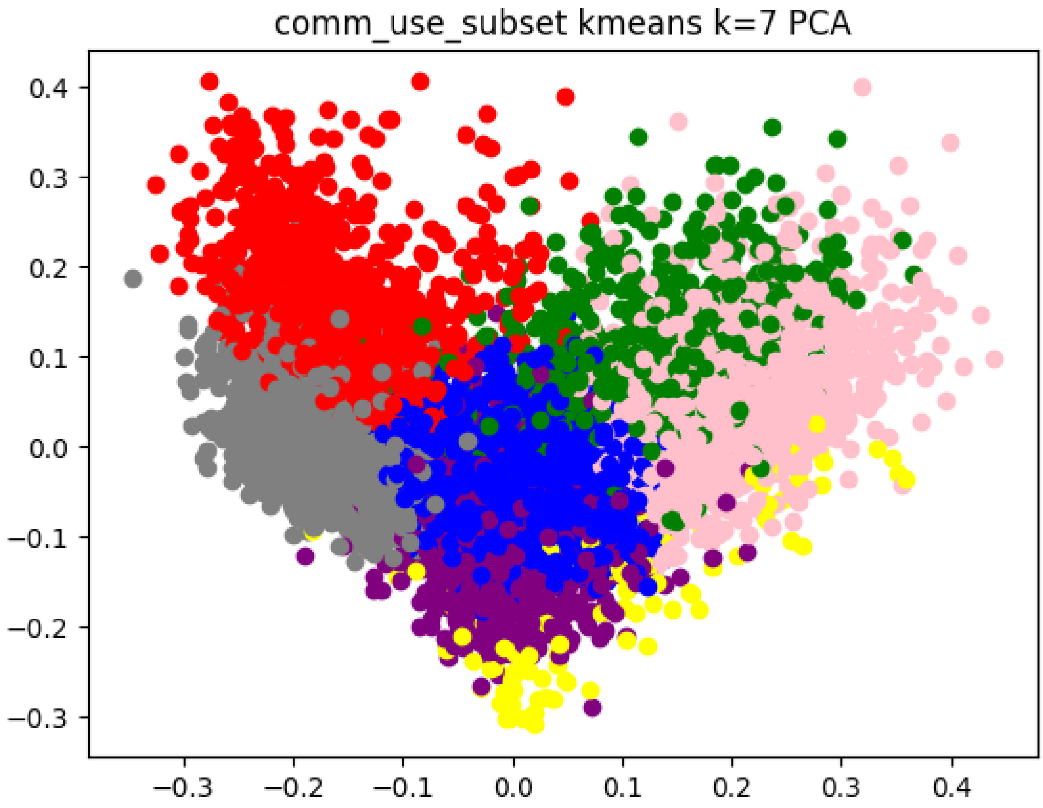

comm_use_subset kmeans k=6 TSNE
100
75
50
25
0
−25
−50
−75
−100
−75
−50
−25
0
25
50
75

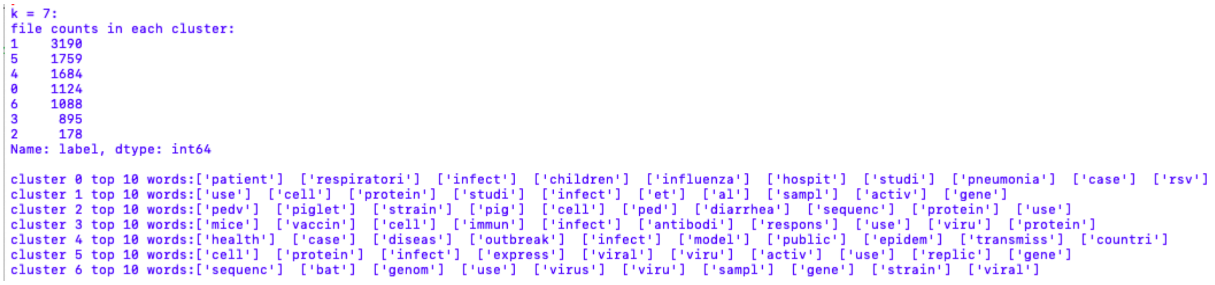

comm_use_subset kmeans k=6 PCA
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4

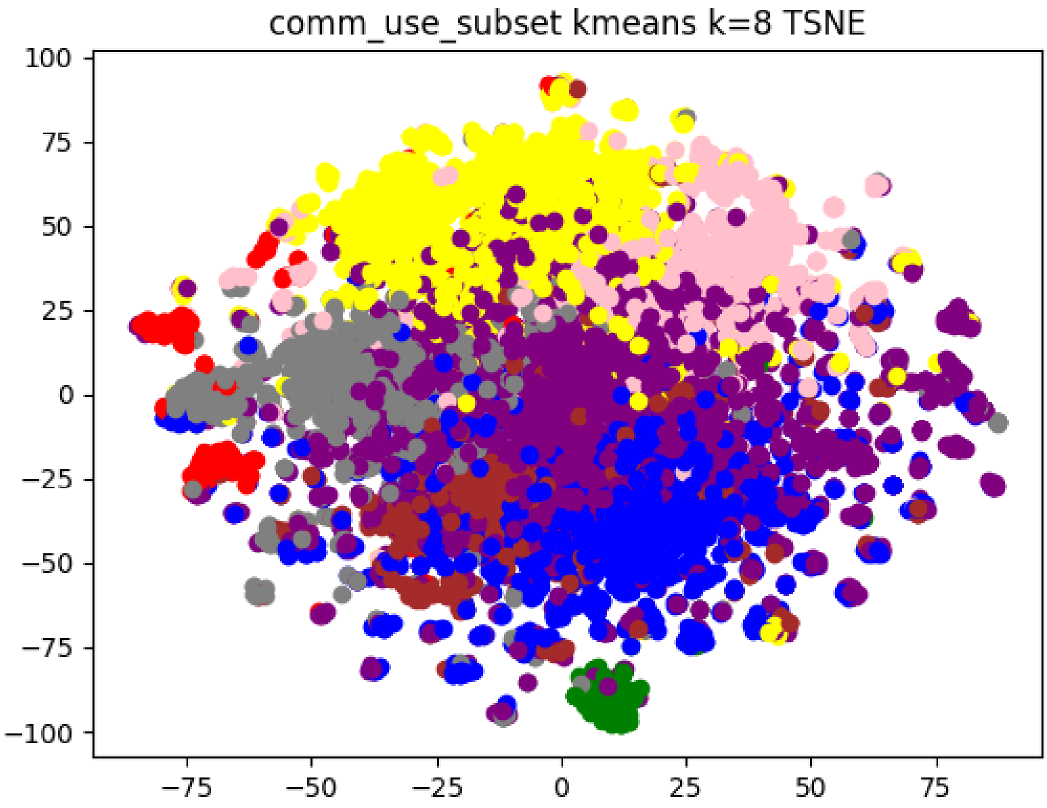


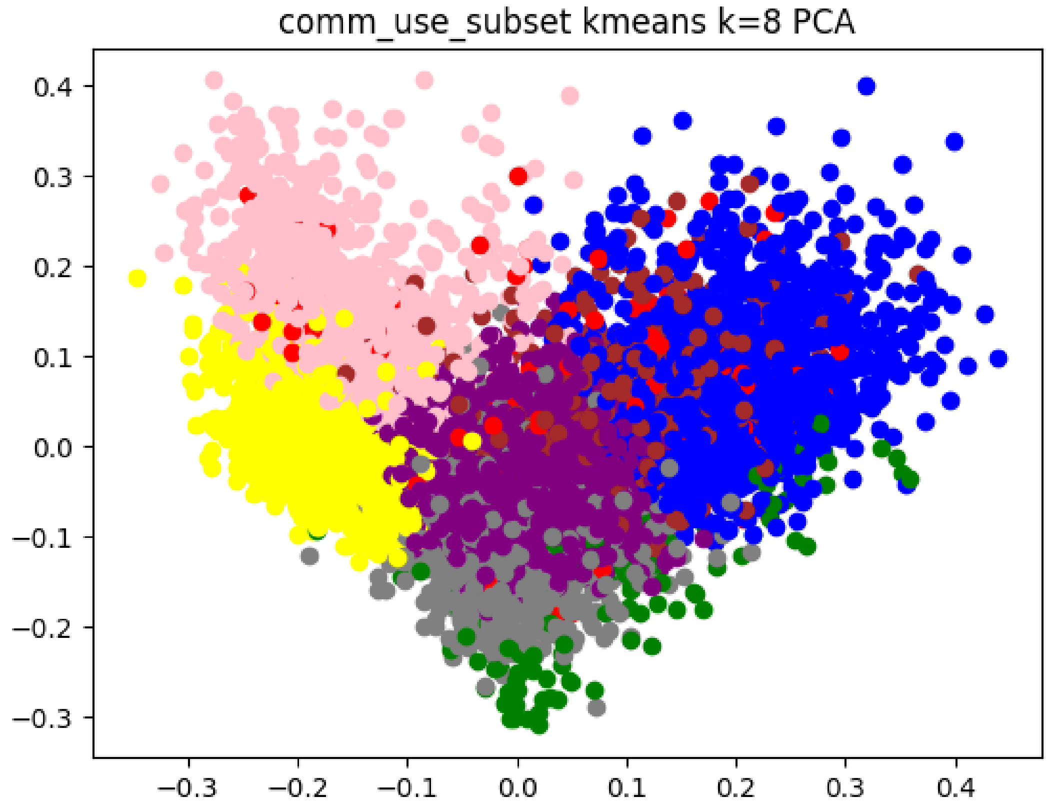
```
k = 6:
file counts in each cluster:
3    3864
1    2256
5    1851
0    1103
2     665
4     179
Name: label, dtype: int64

cluster 0 top 10 words:['patient']  ['respiratori']  ['infect']  ['children']  ['influenza']  ['studi']  ['hospit']  ['pneumonia']  ['rsv']  ['case']
cluster 1 top 10 words:['cell']  ['protein']  ['infect']  ['express']  ['viral']  ['viru']  ['activ']  ['mice']  ['use']  ['fig']
cluster 2 top 10 words:['vaccin']  ['antibodi']  ['cell']  ['immun']  ['mice']  ['epitop']  ['protein']  ['mab']  ['neutral']  ['use']
cluster 3 top 10 words:['sequenc']  ['use']  ['protein']  ['sampl']  ['gene']  ['viru']  ['genom']  ['studi']  ['bat']  ['infect']
cluster 4 top 10 words:['pedv']  ['piglet']  ['strain']  ['pig']  ['cell']  ['ped']  ['diarrhea']  ['sequenc']  ['protein']  ['use']
cluster 5 top 10 words:['health']  ['case']  ['diseas']  ['outbreak']  ['infect']  ['model']  ['public']  ['epidem']  ['transmiss']  ['use']
```

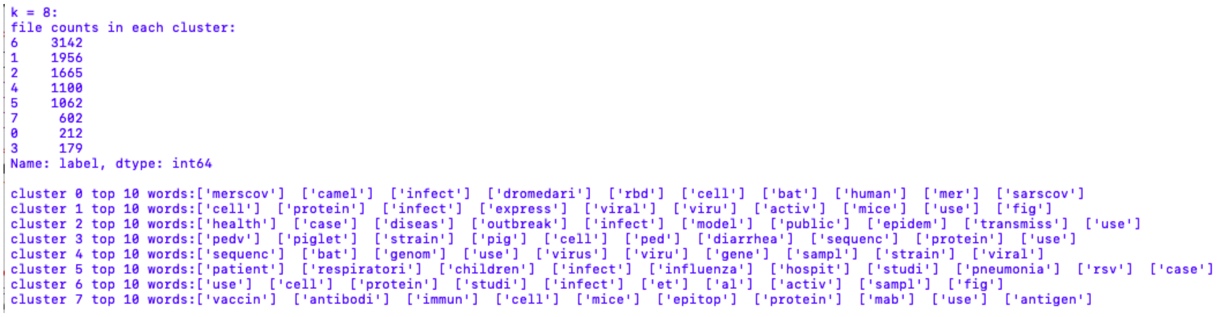





```
k = 7:
file counts in each cluster:
1    3190
5    1759
4    1684
0    1124
6    1088
3     895
2     178
Name: label, dtype: int64

cluster 0 top 10 words:['patient']  ['respiratori']  ['infect']  ['children']  ['influenza']  ['hospit']  ['studi']  ['pneumonia']  ['case']  ['rsv']
cluster 1 top 10 words:['use']  ['cell']  ['protein']  ['studi']  ['infect']  ['et']  ['al']  ['sampl']  ['activ']  ['gene']
cluster 2 top 10 words:['pedv']  ['piglet']  ['strain']  ['pig']  ['cell']  ['ped']  ['diarrhea']  ['sequenc']  ['protein']  ['use']
cluster 3 top 10 words:['mice']  ['vaccin']  ['cell']  ['immun']  ['infect']  ['antibodi']  ['respons']  ['use']  ['viru']  ['protein']
cluster 4 top 10 words:['health']  ['case']  ['diseas']  ['outbreak']  ['infect']  ['model']  ['public']  ['epidem']  ['transmiss']  ['countri']
cluster 5 top 10 words:['cell']  ['protein']  ['infect']  ['express']  ['viral']  ['viru']  ['activ']  ['use']  ['replic']  ['gene']
cluster 6 top 10 words:['sequenc']  ['bat']  ['genom']  ['use']  ['virus']  ['viru']  ['sampl']  ['gene']  ['strain']  ['viral']
```


comm_use_subset kmeans k=8 TSNE
100
75
50
25
0
−25
−50
−75
−100
−75
−50
−25
0
25
50
75


comm_use_subset kmeans k=8 PCA
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4



```
k = 8:
file counts in each cluster:
6    3142
1    1956
2    1665
4    1100
5    1062
7     602
0     212
3     179
Name: label, dtype: int64

cluster 0 top 10 words:['merscov']  ['camel']  ['infect']  ['dromedari']  ['rbd']  ['cell']  ['bat']  ['human']  ['mer']  ['sarscov']
cluster 1 top 10 words:['cell']  ['protein']  ['infect']  ['express']  ['viral']  ['viru']  ['activ']  ['mice']  ['use']  ['fig']
cluster 2 top 10 words:['health']  ['case']  ['diseas']  ['outbreak']  ['infect']  ['model']  ['public']  ['epidem']  ['transmiss']  ['use']
cluster 3 top 10 words:['pedv']  ['piglet']  ['strain']  ['pig']  ['cell']  ['ped']  ['diarrhea']  ['sequenc']  ['protein']  ['use']
cluster 4 top 10 words:['sequenc']  ['bat']  ['genom']  ['use']  ['virus']  ['viru']  ['gene']  ['sampl']  ['strain']  ['viral']
cluster 5 top 10 words:['patient']  ['respiratori']  ['children']  ['infect']  ['influenza']  ['hospit']  ['studi']  ['pneumonia']  ['rsv']  ['case']
cluster 6 top 10 words:['use']  ['cell']  ['protein']  ['studi']  ['infect']  ['et']  ['al']  ['activ']  ['sampl']  ['fig']
cluster 7 top 10 words:['vaccin']  ['antibodi']  ['immun']  ['cell']  ['mice']  ['epitop']  ['protein']  ['mab']  ['use']  ['antigen']
```

From the graphs above, K = 4 appears to be optimal, and the following words are recovered:

Cluster 0: cell, protein, infect, express, mice, virus, viral, use, active, immunity

Cluster 1: patient, respiratory, infect, children, influenza, study, hospital, pneumonia, rsv, case

Cluster 2: health, case, disease, outbreak, infect, model, public, epidemic, use, transmission

Cluster 3: sequence, use, protein, sample, virus, gene, genome, study, infect, cell

Cluster 0 outlines the cellular mechanism of the virus, the signs of activation, the use of mice as experimental subjects for observation and evidence.

Cluster1 outlines concerns about the risk of novel Coronavirus infection in pneumonia patients, with concern for children cases among them.

Cluster2 outlines the outbreak and spread of the virus among the public, and attempts to use this information to establish a propagation model.

Cluster3 outlines biological research of viruses, such as analysis of viral cell components, protein structure, and gene sequences.

# Conclusion

From the above results, it can be seen that the academic theme for the COVID-19 epidemic revolves around the following aspects: virus biological analysis and vaccine production, information modeling and factor analysis of the spread of the epidemic, and investigation of the possibility of infection of the COVID-19 virus in patients with pneumonia.

```
In [ ]:
```