SUMMARY

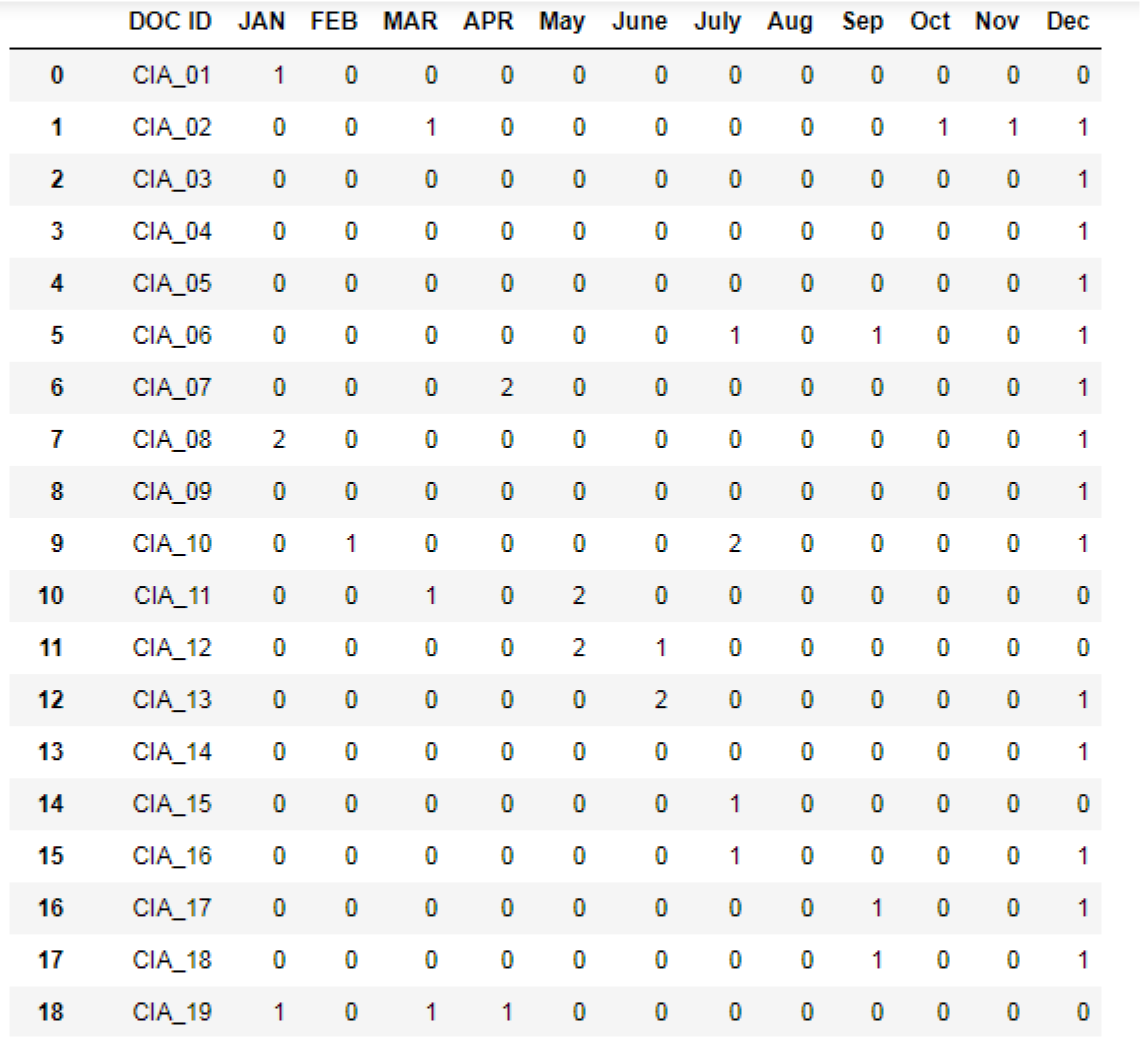
Design:

Our Design consists of, SQL database, python, and html. Database consists of document IDs and document text with ordering number which helps in ordering the document according to document category and cluster rules. Python is used to extract the dates from the content of each document by using datefinder library and the clustering is done based on the months.

We created a document entity matrix with documents as rows and months as a column and we applied Euclidean distance formula to obtain classical Multidimensional Scaling (MDS) for the matrix. We visualize third view above the workspace and list view, where each circle in the plot represents each document and pairwise distance shows similarity in documents.

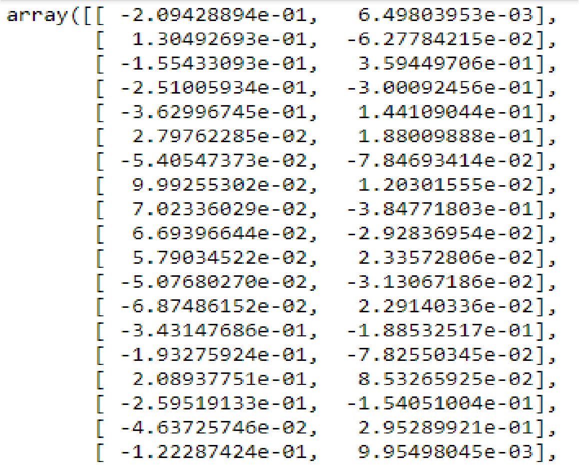
Referred to Scikit-learn Library,

Document- entity matrix looks like below:



	DOC ID	JAN	FEB	MAR	APR	May	June	July	Aug	Sep	Oct	Nov	Dec
0	CIA_01	1	0	0	0	0	0	0	0	0	0	0	0
1	CIA_02	0	0	1	0	0	0	0	0	0	1	1	1
2	CIA_03	0	0	0	0	0	0	0	0	0	0	0	1
3	CIA_04	0	0	0	0	0	0	0	0	0	0	0	1
4	CIA_05	0	0	0	0	0	0	0	0	0	0	0	1
5	CIA_06	0	0	0	0	0	0	1	0	1	0	0	1
6	CIA_07	0	0	0	2	0	0	0	0	0	0	0	1
7	CIA_08	2	0	0	0	0	0	0	0	0	0	0	1
8	CIA_09	0	0	0	0	0	0	0	0	0	0	0	1
9	CIA_10	0	1	0	0	0	0	2	0	0	0	0	1
10	CIA_11	0	0	1	0	2	0	0	0	0	0	0	0
11	CIA_12	0	0	0	0	2	1	0	0	0	0	0	0
12	CIA_13	0	0	0	0	0	2	0	0	0	0	0	1
13	CIA_14	0	0	0	0	0	0	0	0	0	0	0	1
14	CIA_15	0	0	0	0	0	0	1	0	0	0	0	0
15	CIA_16	0	0	0	0	0	0	1	0	0	0	0	1
16	CIA_17	0	0	0	0	0	0	0	0	1	0	0	1
17	CIA_18	0	0	0	0	0	0	0	0	1	0	0	1
18	CIA_19	1	0	1	1	0	0	0	0	0	0	0	0

The MDS manifold co-ordinates is as follows:



```
array([[ -2.09428894e-01,   6.49803953e-03],
       [  1.30492693e-01,  -6.27784215e-02],
       [ -1.55433093e-01,   3.59449706e-01],
       [ -2.51005934e-01,  -3.00092456e-01],
       [ -3.62996745e-01,   1.44109044e-01],
       [  2.79762285e-02,   1.88009888e-01],
       [ -5.40547373e-02,  -7.84693414e-02],
       [  9.99255302e-02,   1.20301555e-02],
       [  7.02336029e-02,  -3.84771803e-01],
       [  6.69396644e-02,  -2.92836954e-02],
       [  5.79034522e-02,   2.33572806e-02],
       [ -5.07680270e-02,  -3.13067186e-02],
       [ -6.87486152e-02,   2.29140336e-02],
       [ -3.43147686e-01,  -1.88532517e-01],
       [ -1.93275924e-01,  -7.82550345e-02],
       [  2.08937751e-01,   8.53265925e-02],
       [ -2.59519133e-01,  -1.54051004e-01],
       [ -4.63725746e-02,   2.95289921e-01],
       [ -1.22287424e-01,   9.95498045e-03],
```

Deploying the project:

1. Download Xampp Server.

2. Extract the zip file (DocumentClustering) in to htdocs folder in the xampp server folder in C drive.

3. Install python, and install the following libraries with the command (pip install lib_name).

a) Flask
b) Sql
c) Datefinder
d) Pandas
e) Numpy
f) Scipy
g) Matplotlib
h) sklearn

4. Open “PhpMyAdmin” (type “localhost/phpmyadmin” in browser), then create a database with name “visual” and import the file “visual.sql” (find it in DocumentClustering folder).

5. Install MySQL Workbench and create a connection called “visual”, and open the connection the database created will be shown in the workbench.

6. Open terminal and cd to DocumentClustering folder in htdocs and run the python scripts

1. python MDS.py
2. python app.py

5. Open browser, type “localhost:5000”.

6. Finally, you can view the project with document list and workspace.