Spectral Clustering Preprocessing

Pipline:

1) Translate-Convert-Extract-1: Translate process

   Input: DB + ID

```
#This part is for saving the file loacation as csv file.
DB="covidpublished"
ID=18
```

   Output:

   CovidPublishedID18ALL.csv, AA_CovidPublishedID18ALL.csv, CovidPublishedID18_seqK.csv

```
#This part is for saving the file loacation as csv file.
toFile = "CovidPublishedID18/CovidPublishedID18ALL.csv"
toFile1 = "CovidPublishedID18/AA_CovidPublishedID18ALL.csv"
toFile2 = "CovidPublishedID18/CovidPublishedID18_seqK.csv"
```

2) Find_Unique_By_Score-2:

   Input: CovidPublishedID18_seqK.csv

```
kmers=pd.read_csv(r"CovidPublishedID18/CovidPublishedID18_seqK.csv",usecols=col_list)
```

   Output: CovidPublishedID18_byScore.csv

```
df.to_csv("CovidPublishedID18/CovidPublishedID18_byScore.csv",index=False)
```

3) Reduce_Using_Variance_Map-3:

   Input: CovidPublishedID18/CovidPublishedID18_byScore.csv

```
path="CovidPublishedID18/CovidPublishedID18_byScore.csv"
```

   Output: CovidPublishedID18_VarRemain.csv

```
final.to_csv('CovidPublishedID18/CovidPublishedID18_VarRemain.csv',index=False)
```

4) SlidingWindow-4:

   Input: CovidPublishedID18_VarRemain.csv

```
df = pd.read_csv("CovidPublishedID18/CovidPublishedID18_VarRemain.csv",usecols=["kmer"])
```

   Output: CovidPublishedID18_slidingwindow_Var.csv

```
df.to_csv(r"CovidPublishedID18/CovidPublishedID18_slidingwindow_Var.csv",index=False)
```

5) Trimers-Filter-5:

   Input: CovidPublishedID18_slidingwindow_Var.csv

```
df = pd.read_csv("CovidPublishedID18/CovidPublishedID18_slidingwindow_Var.csv")
```

trimersDict.csv (Added it to GitHub in the Pipline folder)

```
vocab = pd.read_csv(r"trimersDict.csv",index_col=False)
```

Output: filtered-trimers-CovidPublishedID18_VarRemain.csv

```
result.to_csv("CovidPublishedID18/filtered-trimers-CovidPublishedID18_VarRemain.csv")
```

6) SlidingWindow-Filter-6

Input: CovidPublishedID18_slidingwindow_Var.csv

```
df = pd.read_csv("CovidPublishedID18/CovidPublishedID18_slidingwindow_Var.csv",usecols=["SlidingWindow"])
```

filtered-trimers-CovidPublishedID18_VarRemain.csv

```
vocab = pd.read_csv(r"CovidPublishedID18/filtered-trimers-CovidPublishedID18_VarRemain.csv",index_col=False)
```

Output: CovidPublishedID18_var_SlidingWindow_filter.csv

```
df_new.to_csv("CovidPublishedID18/CovidPublishedID18_var_SlidingWindow_filter.csv",index=False)
```

7) Finding-Trimers-Weights-7

Input: trimersDict.csv

```
#Trimers Table
TriMers = pd.read_csv(r"trimersDict.csv")
```

CovidPublishedID18_var_SlidingWindow_filter.csv

```
#Read CSV File
df = pd.read_csv(r"CovidPublishedID18/CovidPublishedID18_var_SlidingWindow_filter.csv")
```

Output: CovidPublishedID18_VarRemain_trimer_weights.p

```
pickle.dump(result, open(r"CovidPublishedID18/CovidPublishedID18_VarRemain_trimer_weights.p","wb"))
```

MatrixMarket:

1) Matrix-Builder:
   Input: CovidPublishedID18_var_SlidingWindow_filter.csv

```
df = pd.read_csv("CovidPublishedID18/CovidPublishedID18_var_SlidingWindow_filter.csv",usecols=["SlidingWindow"])
```

   CovidPublishedID18_VarRemain_trimer_weights.p

```
trimer_weights = pd.read_pickle(r"CovidPublishedID18/CovidPublishedID18_VarRemain_trimer_weights.p")
```

   filtered-trimers-CovidPublishedID18_VarRemain.csv

```
vocab = pd.read_csv(r"CovidPublishedID18/filtered-trimers-CovidPublishedID18_VarRemain.csv",index_col=False)
```

   Output: matrix (mtx file)

```
scipy.io.mmwrite("CovidPublishedID18/matrix", sparse_matrix)
```

2) Barcodes_Creator:
   Input: NO INPUT
   Output: barcodes.tsv

```
with open('CovidPublishedID18/barcodes.tsv','wt') as out_file:
```

   *** Need to change Manually the number of kmers (Be attentive to the Range function; the second number is not contained):
   For example the below range is: 1 ----→ 679320

```
with open('/home/bshara/trimer_project/Clustering/CovidPublishedID10/barcodes.tsv','wt') as out_file:
    tsv_writer=csv.writer(out_file,delimiter='\t')
    for i in range(1,679321):
            tsv_writer.writerow([i])
```

3) Features_Creator:
   Input: NO INPUT
   Output: genes.tsv

```
with open('CovidPublishedID18/genes.tsv','wt') as out_file:
```

   *** Need to change Manually the number of trimers (Be attentive to the Range function; the second number is not contained):
   For example, the below range is: 1 ----→ 8994

```
with open('CovidPublishedID18/genes.tsv','wt') as out_file:
    tsv_writer=csv.writer(out_file,delimiter='\t')
    for i in range(1,8995):
            tsv_writer.writerow([i])
```

Now You are ready to use the Too-Many-Cells tool.

Please read the full Too-Many-Cells documentation at the following link:
https://gregoryschwartz.github.io/too-many-cells/

GitHub code + commands:
https://github.com/GregorySchwartz/too-many-cells

To add labels to the clustered data, please check the folder Coloring:
Spectral Clustering Preprocessing/Coloring

Working the same technique as the previous scripts, inputs from the output of the too-many-cells tool with some files that were extracted before. (Nothing special or new).

**You will face some data that has been filled manually. You need to change it depending on your desire. (Usually noticeable) like:

```
binding_sample_id = [9,10,11,12,13,14,25,26,27,28,29,30,37,38,39,40,41,42,49,50,51,52,53,54,61,62,63,64,65,66]
```

Wishing you all the best with the project