we are assuming that we don’t know the targeted subsegments. We work only with the message field of the input data, so we preprocess it; then we try to find the optimal number of subsegments using the clustering algorithm and mark the “train data” with the cluster id, which is then used to train the classification algorithm; then we will use  the classification algo to fit the “test data”; train data set and test data sets are the subsets of the given input data

1)How to use it?
 Run
 `jupyter notebook HW2.ipynb`
 Tested on macOS, with python 3.7.2 and jupyter 4.4.0; .csv, the input data file, and .ipynb files should be in the same folder. Jupyter notebook generates an output file “predicted_test_data.csv” that has cluster ids for the “test data”

2)summary?
   we use Silhousette coefficient to measure clustering algo and F1 score to measure classification algo

3) What you would do given more time?
–As part of preprocessing the input data, combine 2 other unused fields (id and subject) into one message, assuming id field can be mapped into some from of segment/security related text. Mark the segments manually (in my approach I used clustering algorithm) and train the classification algorithm; then use this classification algo on test data or any future data

-In general, I use multiple classification/ clustering techniques before narrowing down to one technique
Because manually marking the input data with a subsegment is time consuming task, I used clustering technique for marking which makes it hard to give intuitive reason what each cluster number is