In [151]:
```
#Importing required libraries
import os
import pandas as pd
from itertools import combinations
import datetime as dt
```

Main aim of this is to delete .asm file and keep .byte files for th given task

In [152]:
```
source="C:\\Users\\konda\\Desktop\\MIcrosoft_data\\train"
if os.path.isdir(source):
    data_list=os.listdir(source)
    for file in data_list:
        if file.endswith("asm"):
            os.remove(source+'\\'+file)
```

Convert byte Files to text files

In [8]:
```
#No need to run this code keep in mind.
#Run just once to convert byte to text file
'''source="C:\\Users\\konda\\Desktop\\MIcrosoft_data\\train"
files=os.listdir(source)
print('Total byte files are' ,len(files))
filenames=[]
array=[]
temp=len(files)
for file in files:
    if(file.endswith('bytes')):
        temp=temp-1
        if temp%500==0:
            print('Remaning files ',temp,' and Remaning percentage is ', ((temp*100)/len(files)))
        file=file.split('.')[0]
        text_file=open('C:\\Users\\konda\\Desktop\\MIcrosoft_data\\text_files\\'+file+'.txt','w+')
        with open(source+'\\'+file+'.bytes','r') as fp:
            lines=""
            for line in fp:
                a=line.rstrip().split(" ")[1:]
                b=' '.join(a)
                b=b+'\n'
                text_file.write(b)
            fp.close()
            os.remove('C:\\Users\\konda\\Desktop\\MIcrosoft_data\\train\\'+file+'.bytes')
        text_file.close()
'''
```

```
Total byte files are 10868
Remaning files  10800  and Remaning percentage is  99.37430990062569
Remaning files  10700  and Remaning percentage is  98.45417740154582
Remaning files  10600  and Remaning percentage is  97.53404490246595
Remaning files  10500  and Remaning percentage is  96.61391240338608
Remaning files  10400  and Remaning percentage is  95.69377990430623
Remaning files  10300  and Remaning percentage is  94.77364740522636
Remaning files  10200  and Remaning percentage is  93.85351490614649
Remaning files  10100  and Remaning percentage is  92.93338240706662
Remaning files  10000  and Remaning percentage is  92.01324990798675
Remaning files  9900  and Remaning percentage is  91.09311740890688
Remaning files  9800  and Remaning percentage is  90.17298490982702
Remaning files  9700  and Remaning percentage is  89.25285241074715
Remaning files  9600  and Remaning percentage is  88.33271991166728
Remaning files  9500  and Remaning percentage is  87.41258741258741
Remaning files  9400  and Remaning percentage is  86.49245491350754
Remaning files  9300  and Remaning percentage is  85.57232241442767
Remaning files  9200  and Remaning percentage is  84.65218991533478
Remaning files  9100  and Remaning percentage is  83.73205741626795
Remaning files  9000  and Remaning percentage is  82.81192491718808
Remaning files  8900  and Remaning percentage is  81.89179241810821
Remaning files  8800  and Remaning percentage is  80.97165991902834
Remaning files  8700  and Remaning percentage is  80.05152741994847
Remaning files  8600  and Remaning percentage is  79.13139492086686
Remaning files  8500  and Remaning percentage is  78.21126242178873
Remaning files  8400  and Remaning percentage is  77.29112992270888
Remaning files  8300  and Remaning percentage is  76.3709974236929
Remaning files  8200  and Remaning percentage is  75.45086492454914
Remaning files  8100  and Remaning percentage is  74.53073242546927
Remaning files  8000  and Remaning percentage is  73.6105999263894
Remaning files  7900  and Remaning percentage is  72.69046742730953
Remaning files  7800  and Remaning percentage is  71.77033492822966
Remaning files  7700  and Remaning percentage is  70.8502024291498
Remaning files  7600  and Remaning percentage is  69.9300693006993
Remaning files  7500  and Remaning percentage is  69.0099374099006
Remaning files  7400  and Remaning percentage is  68.0898049319102
Remaning files  7300  and Remaning percentage is  67.1696724328302
Remaning files  7200  and Remaning percentage is  66.2495399375046
Remaning files  7100  and Remaning percentage is  65.3294074346769
Remaning files  7000  and Remaning percentage is  64.4092749355973
Remaning files  6900  and Remaning percentage is  63.4891424365186
Remaning files  6800  and Remaning percentage is  62.5690099374399
Remaning files  6700  and Remaning percentage is  61.6488774383512
Remaning files  6600  and Remaning percentage is  60.728744939275
Remaning files  6500  and Remaning percentage is  59.8086124401939
Remaning files  6400  and Remaning percentage is  58.888479941112
Remaning files  6300  and Remaning percentage is  57.968347442031065
Remaning files  6200  and Remaning percentage is  57.0482149429519
Remaning files  6100  and Remaning percentage is  56.1280824438719
Remaning files  6000  and Remaning percentage is  55.207499447920
Remaning files  5900  and Remaning percentage is  54.287817445712186
Remaning files  5800  and Remaning percentage is  53.367684946632316
Remaning files  5700  and Remaning percentage is  52.44755244755245
Remaning files  5600  and Remaning percentage is  51.527419484725
Remaning files  5500  and Remaning percentage is  50.607287449392715
Remaning files  5400  and Remaning percentage is  49.68715495031284
Remaning files  5300  and Remaning percentage is  48.767022451232975
Remaning files  5200  and Remaning percentage is  47.84688995215311
Remaning files  5100  and Remaning percentage is  46.926757453073246
Remaning files  5000  and Remaning percentage is  46.006624953993374
Remaning files  4900  and Remaning percentage is  45.086492549351
Remaning files  4800  and Remaning percentage is  44.16635995583364
Remaning files  4700  and Remaning percentage is  43.246227456757377
Remaning files  4600  and Remaning percentage is  42.3260949576739
Remaning files  4500  and Remaning percentage is  41.40596245859404
Remaning files  4400  and Remaning percentage is  40.485829595511417
Remaning files  4300  and Remaning percentage is  39.5656974604343
Remaning files  4200  and Remaning percentage is  38.645564961354444
Remaning files  4100  and Remaning percentage is  37.72541246227457
Remaning files  4000  and Remaning percentage is  36.8052999631947
Remaning files  3900  and Remaning percentage is  35.885167646411483
Remaning files  3800  and Remaning percentage is  34.9650349650349
Remaning files  3700  and Remaning percentage is  34.044902465955
Remaning files  3600  and Remaning percentage is  33.1247699668752
Remaning files  3500  and Remaning percentage is  32.204637467795365
Remaning files  3400  and Remaning percentage is  31.284504968715495
Remaning files  3300  and Remaning percentage is  30.364372469635626
Remaning files  3200  and Remaning percentage is  29.44423997055576
Remaning files  3100  and Remaning percentage is  28.524107471475894
Remaning files  3000  and Remaning percentage is  27.603974972396024
Remaning files  2900  and Remaning percentage is  26.683842473316158
Remaning files  2800  and Remaning percentage is  25.76370997423629
Remaning files  2700  and Remaning percentage is  24.8435774751564422
Remaning files  2600  and Remaning percentage is  23.923444976076556
Remaning files  2500  and Remaning percentage is  23.003312476996687
Remaning files  2400  and Remaning percentage is  22.083179977916682
Remaning files  2300  and Remaning percentage is  21.16304747883695
Remaning files  2200  and Remaning percentage is  20.242914979757085
Remaning files  2100  and Remaning percentage is  19.3227824806722
Remaning files  2000  and Remaning percentage is  18.402649981559735
Remaning files  1900  and Remaning percentage is  17.48251748251748
Remaning files  1800  and Remaning percentage is  16.562384983437614
Remaning files  1700  and Remaning percentage is  15.642252484357748
Remaning files  1600  and Remaning percentage is  14.722119985277788
Remaning files  1500  and Remaning percentage is  13.801987486198012
Remaning files  1400  and Remaning percentage is  12.881854987118144
Remaning files  1300  and Remaning percentage is  11.961722488038278
Remaning files  1200  and Remaning percentage is  11.041589988958411
Remaning files  1100  and Remaning percentage is  10.12145749878543
Remaning files  1000  and Remaning percentage is  9.201324999798675
Remaning files  900  and Remaning percentage is  8.281192491718807
Remaning files  800  and Remaning percentage is  7.361059992638894
Remaning files  700  and Remaning percentage is  6.4409274935590072
Remaning files  600  and Remaning percentage is  5.520749944479205
Remaning files  500  and Remaning percentage is  4.600662495399337
Remaning files  400  and Remaning percentage is  3.6805299963194
Remaning files  300  and Remaning percentage is  2.760397497239626
Remaning files  200  and Remaning percentage is  1.840264998159735
Remaning files  100  and Remaning percentage is  0.920132499079867
Remaning files  0  and Remaning percentage is  0.0
```

Converting text file to bigram

Finding all the possible combinations of unigrams to get bigrams

In [153]:
```
l="00,01,02,03,04,05,06,07,08,09,0a,0b,0c,0d,0e,0f,10,11,12,13,14,15,16,17,18,19,1a,1b,1c,1d,1e,1f,20,21,22,23,24,25,26,27,28,29,2a,2b,2c,2d,2e,2f,30,31,32,33,34,35,36,37,38,39,3a,
te=l.split(',')
total_bigramm= list(combinations(te, 2))
def big_gram(i):
    return i[0]+' '+i[1],i[1]+' '+i[0]
def gg(i):
    return i+' '+i
total_bigrams=[]
for i in total_bigramm:
    total_bigrams.extend(big_gram(i))
for i in te:
    total_bigrams.append(gg(i))
new=['FileName']
new.extend(total_bigrams)
```

In [154]:
```
gg('00')
```

Out[154]:
```
'00 00'
```

In [155]:
```
def dicg(fileName,total_bigrams):#return a dict of file name and biagrams in the form of dict\
    d={}
    d['FileName']=fileName
    for i in total_bigrams:
        d[i]=0
    return d
```

In [156]:
```
def find_in_it(bia,x):
    for i in bia:
        if i==x:
            return "Found "+i
    return -1
find_in_it(total_bigrams,'00 00')
```

Out[156]:
```
'Found 00 00'
```

Creating an empty pandas dataframe

In [157]:
```
#find the sequence combination of a list
#l=[1,2,3,4,5] ==> 1 2 , 2 3 , 3 4 , 4 5
def comb(lis,file,d):
    d=d.fromkeys(d,0)
    d['FileName']=file
    #d={'1 2':0,'2 3':0,'3 4':0,'4 5':0}
    #print('comb function is caled at ',dt.datetime.now())
    for i in range(1,len(lis)):
        tolis[i-1]+' '+lis[i]
        new_t=''
        for j in t:
            if j.isalpha():
                new_t=new_t+j.lower()
            else:
                new_t=new_t+j
        t=new_t
        d[t]=d[t]+1
    #print('comb function is end at ',dt.datetime.now())
    #t=lis[len(lis)-1]+' '+lis[len(lis)]
    #d[t]=d[t]+1
    return d
```

In [158]:
```
global df
df=pd.DataFrame(columns=new)
```

In [162]:
```
#Creating a function to convert text file data to bigrams vector
req=dicg('xxx',total_bigrams)
def read_text_files_to_vector(location):
    global df
    FileNames=os.listdir(location)
    all_files=len(FileNames)
    print('Total files are ',all_files)
    done=0
    for file in FileNames:
        total_lines=[]
        with open(location+'\\'+file) as lines:
            for line in lines:
                line=line.strip()
                total_lines.extend(line.split())
        df=df.append(comb(total_lines,file,req),ignore_index=True)
        os.remove(location+'\\'+file)

        if (done%500==0):
            print(f'Completed files are {done} and Remaning per is {(all_files-done)*100/all_files}')

        done=done+1
    return df
```

In [163]:
```
st=dt.datetime.now()

read_text_files_to_vector("C:\\Users\\konda\\Desktop\\MIcrosoft_data\\text_files")
print(f'Total time is {dt.datetime.now()-st}')
```

```
Total files are  10868
Completed files are 0 and Remaning per is 100.0
Completed files are 500 and Remaning per is 95.39933750460067
Completed files are 1000 and Remaning per is 90.79867500920132
Completed files are 1500 and Remaning per is 86.19801251380198
Completed files are 2000 and Remaning per is 81.59735001840265
Completed files are 2500 and Remaning per is 76.99668752300332
Completed files are 3000 and Remaning per is 72.39602502760397
Completed files are 3500 and Remaning per is 67.79536253220463
Completed files are 4000 and Remaning per is 63.194700038053
Completed files are 4500 and Remaning per is 58.594037541405596
Completed files are 5000 and Remaning per is 53.993375046006626
Completed files are 5500 and Remaning per is 49.392712550607285
Completed files are 6000 and Remaning per is 44.792050055207795
Completed files are 6500 and Remaning per is 40.19138755980861
Completed files are 7000 and Remaning per is 35.590725064409928
Completed files are 7500 and Remaning per is 30.990062569009936
Completed files are 8000 and Remaning per is 26.389400073610
Completed files are 8500 and Remaning per is 21.788737578721126
Completed files are 9000 and Remaning per is 17.188075082811924
Completed files are 9500 and Remaning per is 12.587412587412587
Completed files are 10000 and Remaning per is 7.986750092201325
Completed files are 10500 and Remaning per is 3.386087596613923
Total time is 1 day, 22:02:39.509819
```

In [165]:
```
df.to_csv('new_features_biagram.csv')
```