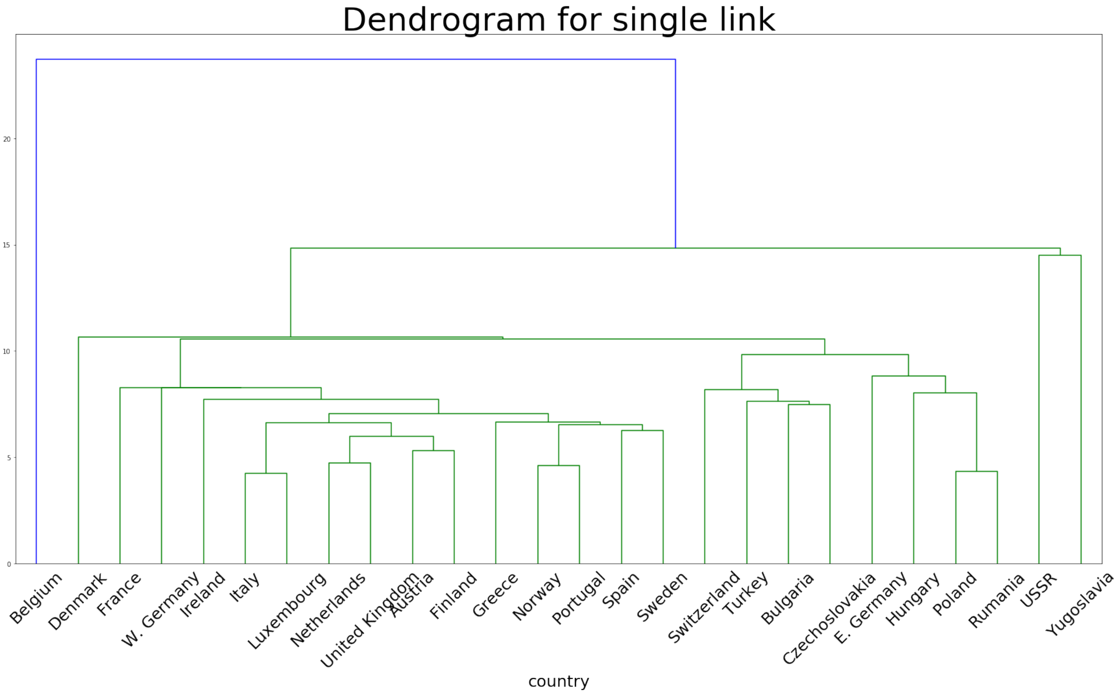
upload

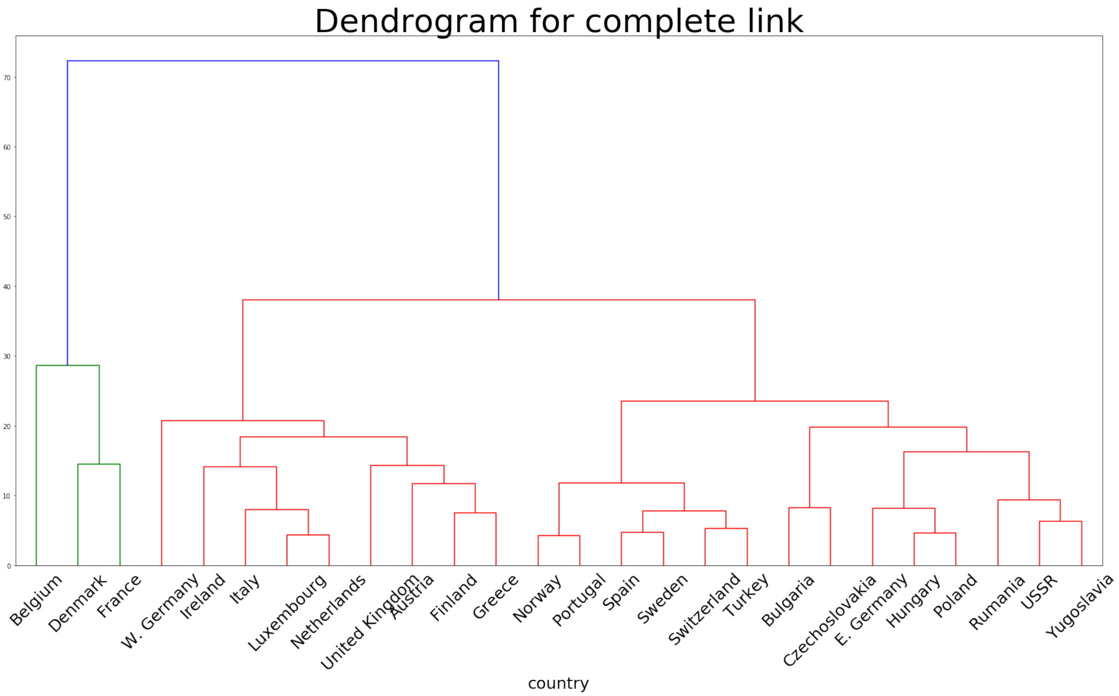
March 3, 2018

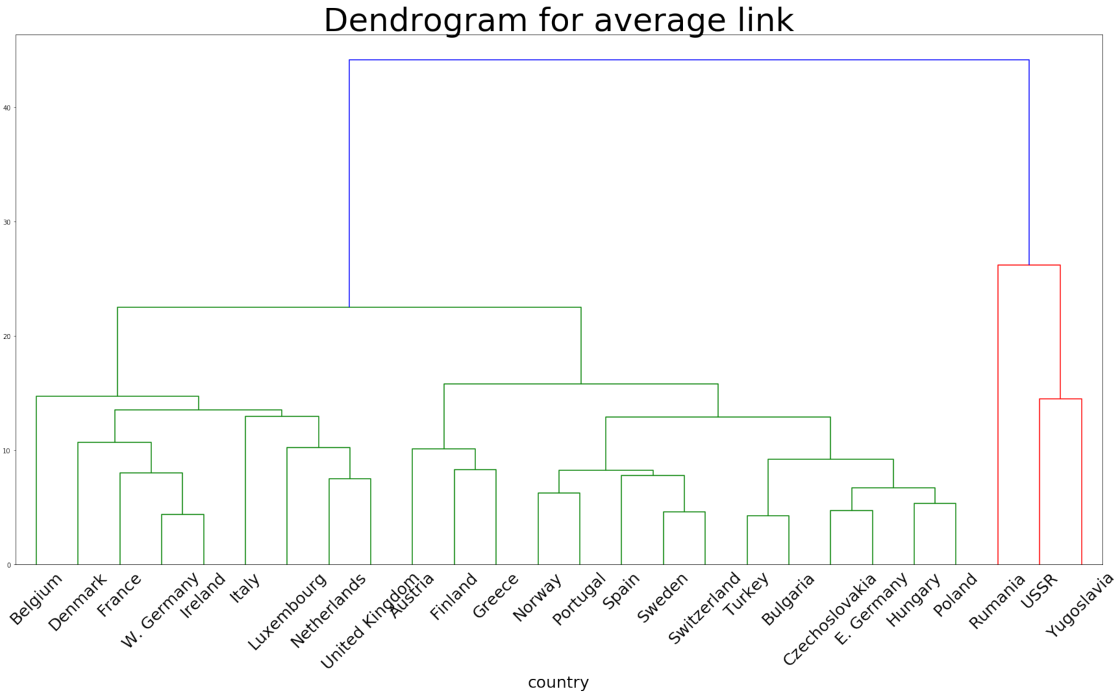
0.1 HW4

0.1.1 Problem1 Part A

In [1]:




Dendrogram for complete link


Dendrogram for average link

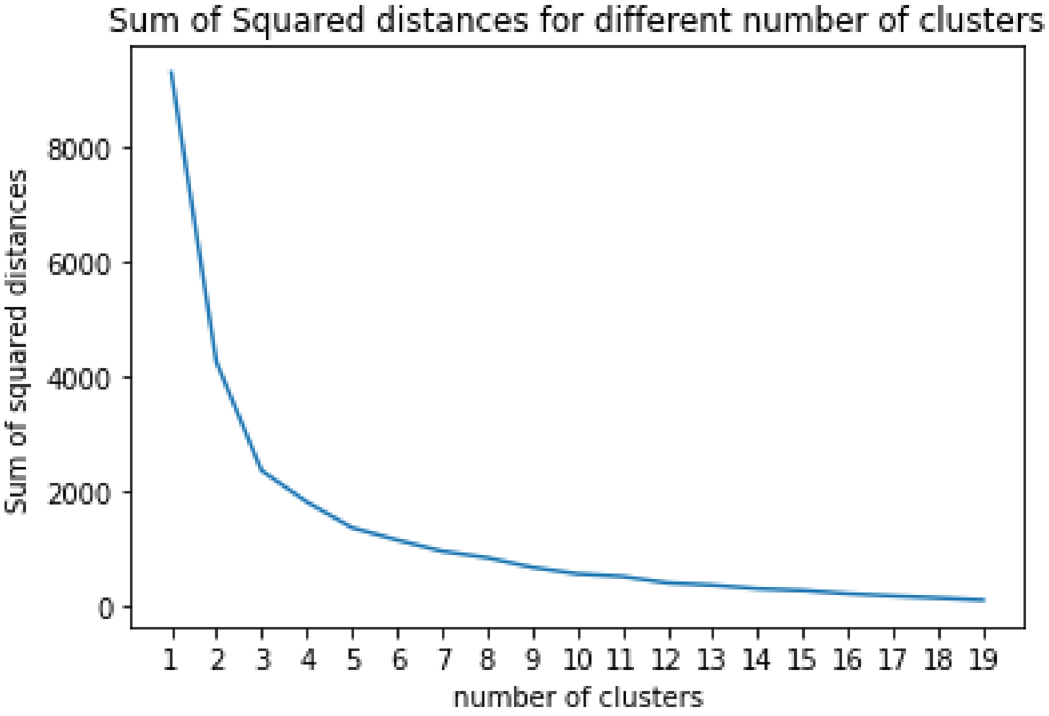
Complete line and average link generates very likely graph. In the middle of the graph, you can find the patterns are the same. Also you can find countries freed from Soviet (Bulgaria,

E.Germany etc.)tend to cluster together.

0.1.2 Problem 1 Part b

```
In [2]:
```

```
Out[2]: Text(0.5,1,'Sum of Squared distances for different number of clusters')
```



0.1.3 This plot shows a sudden decrease from k =1 to k=3, after k =7, cost drops slowly. So k=7,8,9 is good choices

0.1.4 Problem 2 Part A

0.1.5 Using segment length = 32 and following textbook using first level= 40 clusters and second level 12 clusters. I get below results from random forest classifier(trees=50,depth=30). Error rate is about 26% . Also confusion matrix is shown below

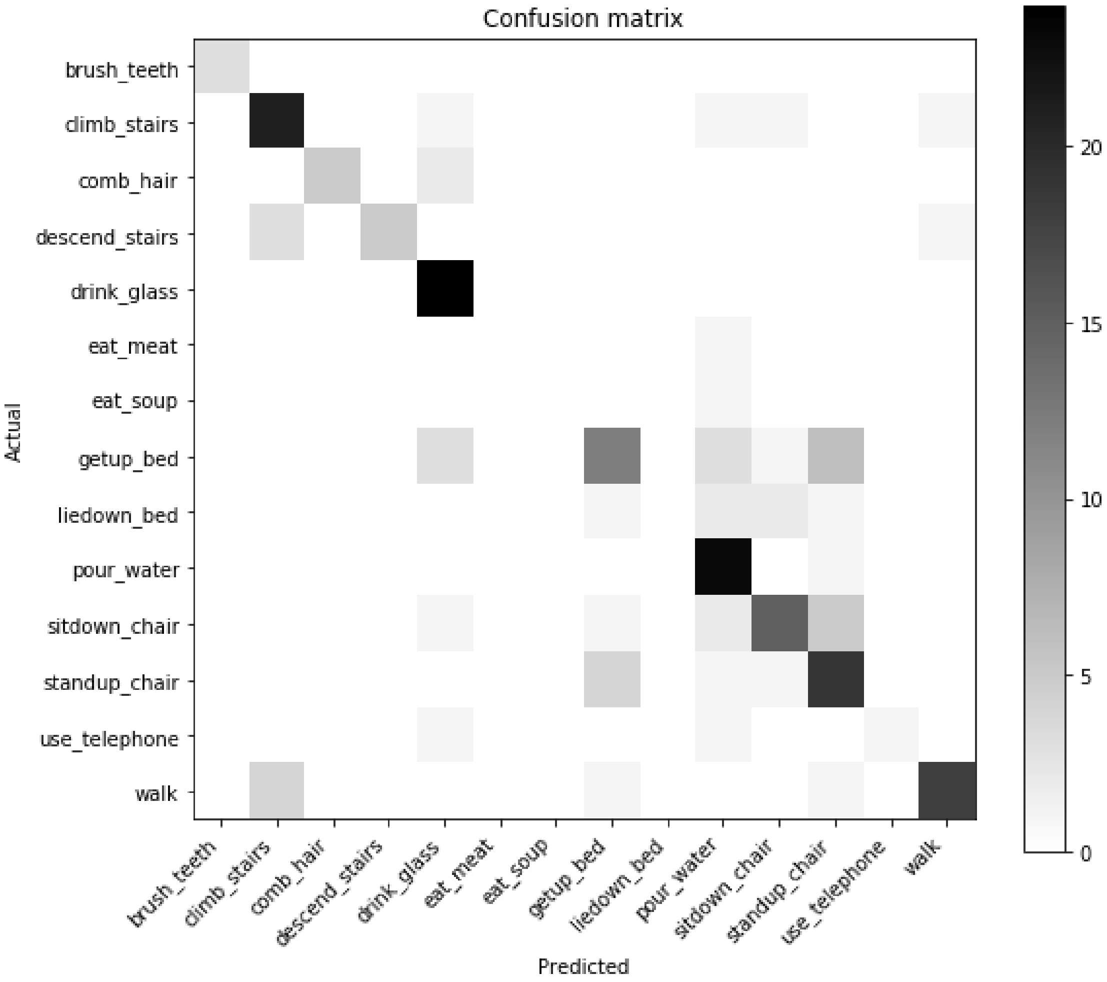
```
In [2]: clf(seg_len=32,k1=40,k2=12)
```

```
0.736318407960199
```

```
Out[2]: array([[ 3,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0],
               [ 0, 21,  0,  1,  0,  0,  0,  0,  0,  1,  1,  1,  0,  0],
               [ 0,  0,  3,  0,  4,  0,  0,  0,  0,  0,  0,  0,  0,  0],
               [ 0,  3,  0,  6,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0],
               [ 0,  0,  0,  0, 24,  0,  0,  0,  0,  0,  0,  0,  0,  0],
               [ 0,  0,  0,  0,  0,  1,  0,  0,  0,  0,  0,  0,  0,  0],
               [ 0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,  0],
               [ 0,  0,  0,  0,  1,  0,  0, 13,  0,  1,  4,  6,  0,  0],
               [ 0,  0,  0,  0,  0,  0,  0,  2,  0,  0,  4,  0,  0,  0],
               [ 0,  0,  0,  0,  0,  0,  0,  0,  0, 23,  1,  0,  0,  0],
               [ 0,  0,  0,  0,  1,  0,  0,  2,  0,  2, 15,  4,  0,  0],
               [ 0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  3, 21,  0,  0],
               [ 0,  0,  0,  0,  2,  0,  0,  0,  0,  1,  0,  0,  0,  0],
               [ 0,  2,  0,  0,  0,  0,  0,  1,  0,  0,  1,  2,  0, 18]],
              dtype=int64)
```

In [24]:

Out[24]: <matplotlib.axes._subplots.AxesSubplot at 0x24f9fcabb38>



0.1.6 Problem 2 Part B

0.1.7 Under the same clusters, choose segment from [4,8,16,32,64] to check if which is the best length. From the result we can see shorter length gives better accuracy. So I choose length 4 to test clusters numbers. By tuning around (slightly below or above 40 with 12), I found out the accuracy dropped as the clusters goes up. So I select first level from[12,16,20,24,28,32], second level from[6,8,10]. This produces more than 80% accuracy. Also with segment length 4, first level 16, second level 6. This is so far I see the highest, about 86.5%(shown below)

0.1.8 I also tried very higher clusters 60,80,100, second 20,30. Sometimes there aren't enough samples than the cluster numbers.

```
In [17]: for k in [4]:
             for i in [12,16,20,24,28,32]:
                 for j in [6,8,10]:
                      clf(seg_len=k,k1=i,k2=j)
```

```
0.8606965174129353
0.835820895522388
0.8258706467661692
0.8656716417910447
0.845771144278607
0.8308457711442786
0.8557213930348259
0.835820895522388
0.8507462686567164
0.8308457711442786
0.8656716417910447
0.8208955223880597
0.845771144278607
0.8059701492537313
0.8258706467661692
0.8159203980099502
0.8308457711442786
0.8258706467661692
```

```
In [18]: clf(seg_len=4,k1=16,k2=6)
```

```
0.8656716417910447
```