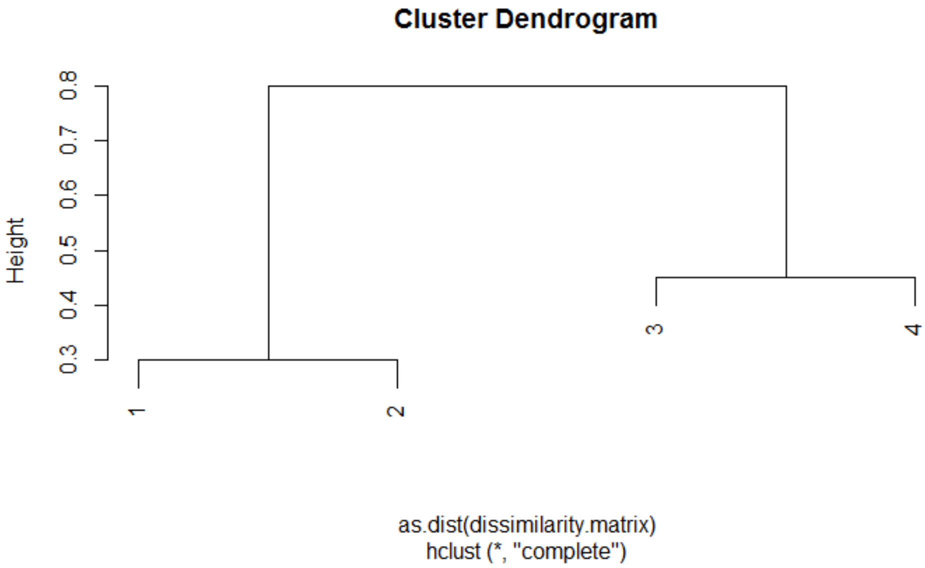
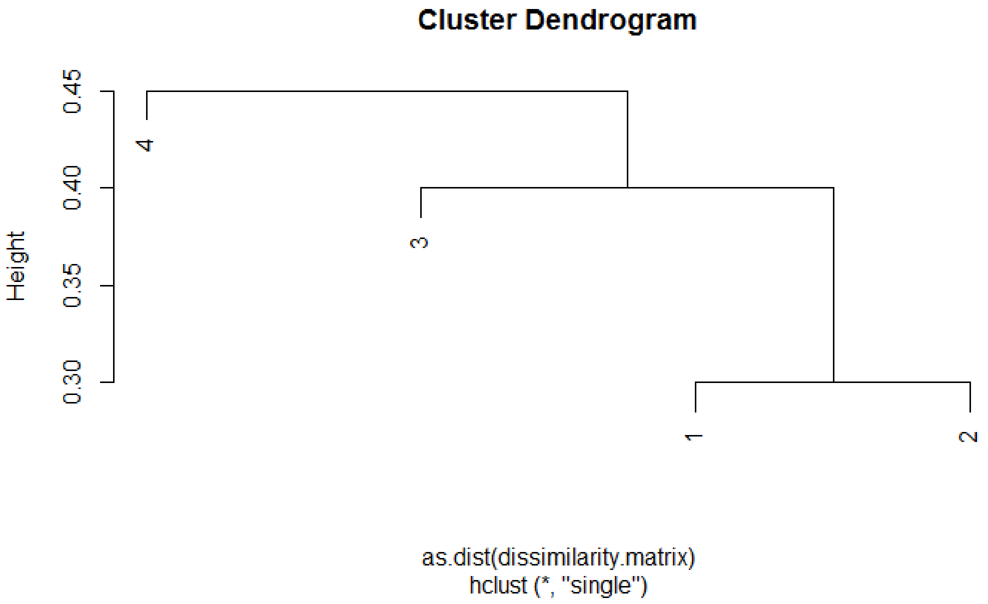
ISLR 10.7.2

```
# (a)

dissimilarity.matrix = matrix(c(0, 0.3, 0.4, 0.7,
                        0.3, 0, 0.5, 0.8,
                        0.4, 0.5, 0, 0.45,
                        0.7, 0.8, 0.45, 0), nrow = 4)

plot(hclust(as.dist(dissimilarity.matrix), method = "complete"))
```



```
# (b)

plot(hclust(as.dist(dissimilarity.matrix), method = "single"))
```



(c)

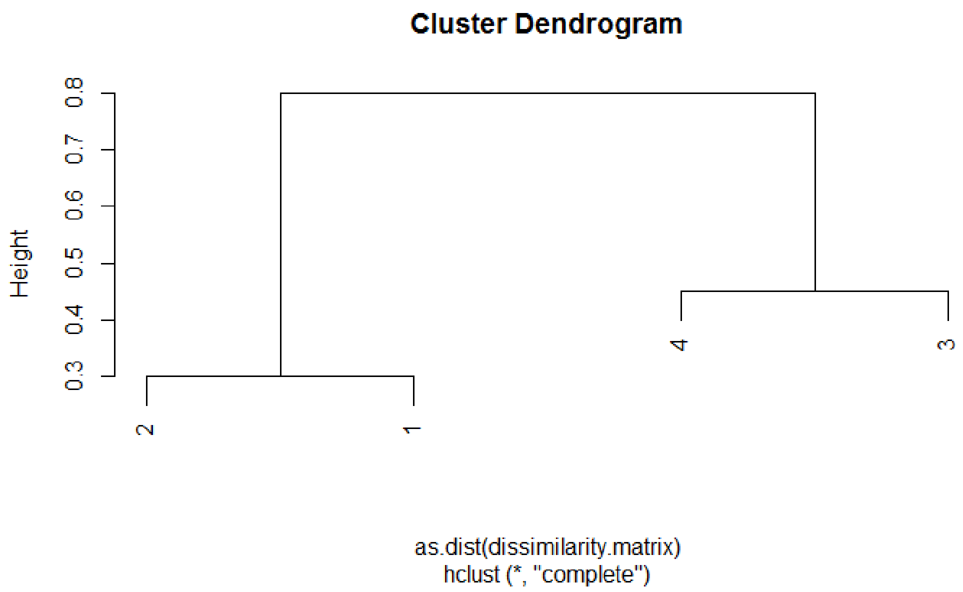
# If we cut the dendrogram in two clusters result, we will have clusters (1,2) and (3,4).

(d)

# If we cut the dendrogram in two clusters result, we will have clusters ((1,2),3) and (4).

(e)

plot(hclust(as.dist(dissimilarity.matrix), method = "complete"), labels = c(2,1,4,3))



10.7.4

(a)

There is not enough information to tell. For example, if d(4,1)=2, d(5,1)=1, d(4,2)=3, d(5,2)=4, d(4,3)=2 and d(5,3)=2, the single linkage dissimilarity between {1,2,3} and {4,5} would be equal to 1 and the complete linkage dissimilarity between {1,2,3} and {4,5} would be equal to 4. So, with single linkage, they would fuse at a height of 1, and with complete linkage, they would fuse at a height of 4.

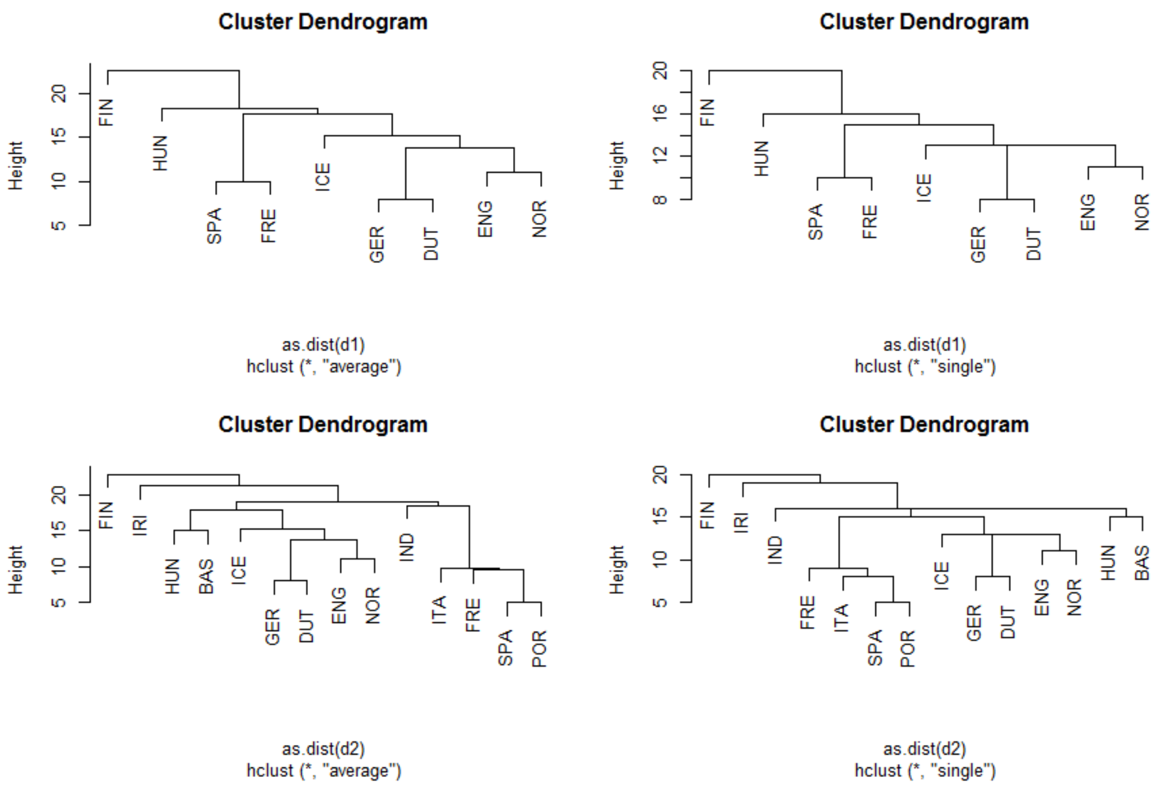
Thus, if the distances between all elements in {1,2,3} and {4,5} are not the same, the single linkage will choose the minimum, while the complete linkage will choose the maximum instead. However, if all inter-observations distances are equal, the single and complete linkage dissimilarities between {1,2,3} and {4,5} are equal, since the minimum equals to the maximum.

(b)

They would fuse at the same height because the distance between {5} and {6} is the only number that will be considered for further fusion. Both the single linkage and the complete linkage will fuse at the same height.

```
### A set of languages represented by spelling the numbers 1-5
lang1 <- c("one two three four five",
      "uno dos tres cuatro cinco",
      "ein zwei drei vier funf",
      "un deux trois quatre cinq",
      "en to tre fire fem",
      "einn tveir thrir fjogur fimm",
      "een twee drie vier vijf",
      "egy ket harom negy ot",
      "yksi kaksi kolme nelja viisi")
lb1 <- c("ENG","SPA","GER","FRE","NOR","ICE","DUT","HUN","FIN")

lang2 <- c("one two three four five",
      "uno dos tres cuatro cinco",
      "ein zwei drei vier funf",
      "un deux trois quatre cinq",
      "en to tre fire fem",
      "einn tveir thrir fjogur fimm",
      "een twee drie vier vijf",
      "egy ket harom negy ot",
      "yksi kaksi kolme nelja viisi",
      "um dois três quatro cinco",
      "uno due tre quattro cinque",
      "amháin dhá trí ceathair cúig",
      "bat bi hiru lau bost",
      "satu dua tiga empat lima")
lb2 <- c("ENG","SPA","GER","FRE","NOR","ICE","DUT","HUN","FIN","POR","ITA","IRI","BAS","IND")

#Calculate Levenshtein distances
d1 <- stringdistmatrix(lang1, lang1, method = "lv")
```

```
d2 <- stringdistmatrix(lang2, lang2, method = "lv")



#Hierarchical cluster analysis on a set of dissimilarities and methods for analyzing it.
cluster1 <- hclust(as.dist(d1),method="average")
cluster2 <- hclust(as.dist(d1),method="single")
cluster3 <- hclust(as.dist(d2),method="average")
cluster4 <- hclust(as.dist(d2),method="single")

#Plot Dendrogram
par(mfrow=c(2,2))
plot(cluster1,labels=lb1)
plot(cluster2,labels=lb1)
plot(cluster3,labels=lb2)
plot(cluster4,labels=lb2)
```



From the comparison between 4 dendrograms above, we can see that the tree structures of method “average” and “single” look similar before adding new languages. For the original languages dataset, switching method from average linkage to single average does not make obvious difference. The main difference between the two methods for the original dataset is the fusion height. The scope of average linkage heights is greater than the scope of single linkage heights, since single linkage selects the minimum distance while average linkage calculates all distances and selects the mean one. There is a slightly different fusion among language ICE, with GER and DUT, and with ENG and NOR for the two methods for the original dataset.

After adding 5 more languages, the tree structure among language ICE, with GER and DUT, and with ENG and NOR is retained for both average and single linkages compared to their corresponding original datasets. That means the languages I added do not affect the fusion between those languages. Besides, for the comparison between the two methods for the same modified dataset, the dendrograms look obviously different. Therefore, the linkage methods do matter for the modified languages dataset. The scope of fusion height of average linkage is a bit larger than that of single linkage for the modified dataset.