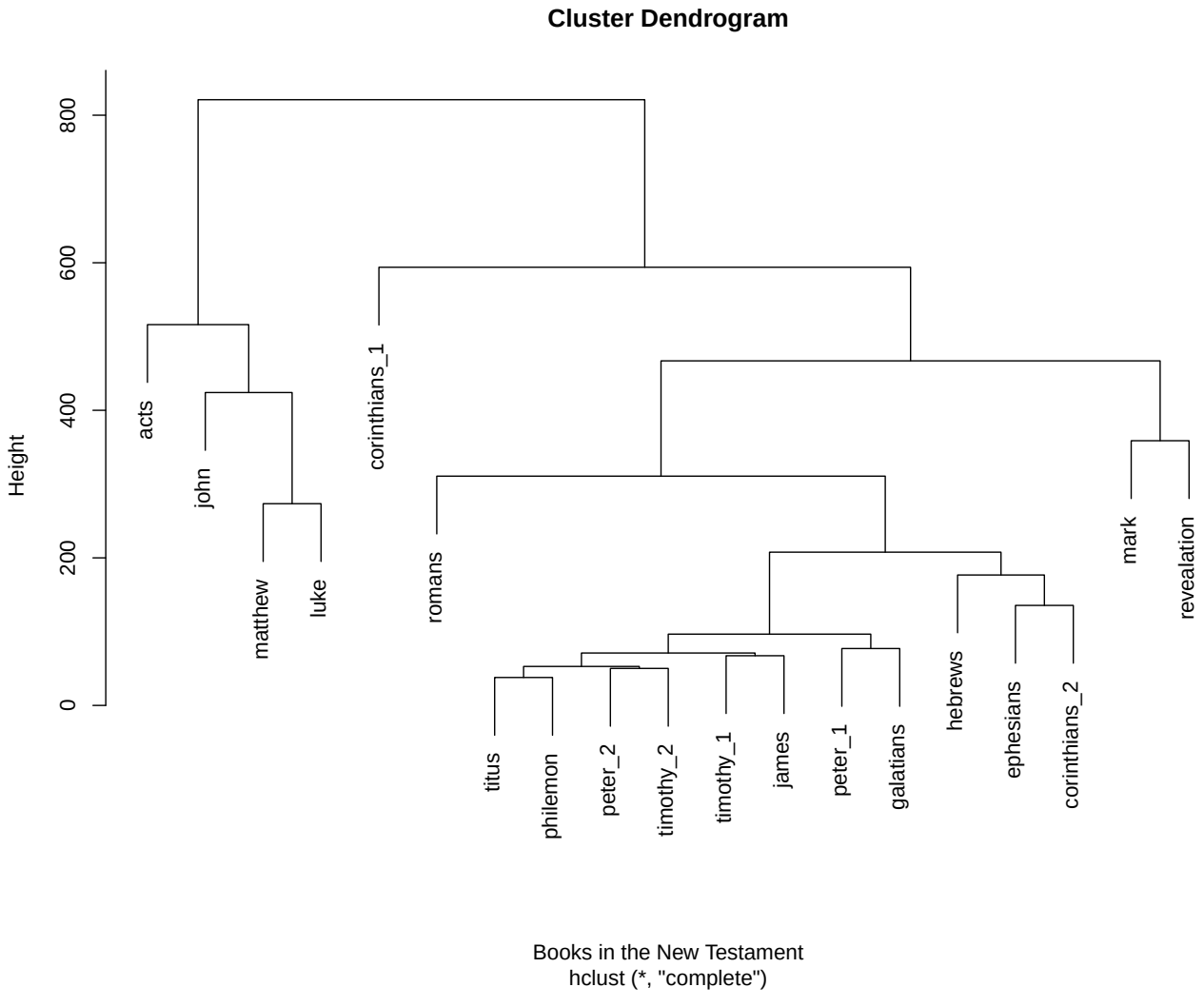
Hierarchical Clustering & PCA of New Testament Books

Steven Lam

```
library(SnowballC)
library(tm)
library(tidyr)
library(plyr)
library(dplyr)
library(MASS)
library(e1071)
setwd("/Users/Steven_Tom/Desktop/Hierarchical_Clustering_NT/NewTestament/Combined")

matthew <-  readChar('Matthew.txt', file.info('Matthew.txt')$size)
mark <-  readChar('Mark.txt', file.info('Mark.txt')$size)
luke <-  readChar('Luke.txt', file.info('Luke.txt')$size)
john <-  readChar('John.txt', file.info('John.txt')$size)

acts <-  readChar('Acts.txt', file.info('Acts.txt')$size)

peter_1 <-  readChar('1Peter.txt', file.info('1Peter.txt')$size)
peter_2 <-  readChar('2Peter.txt', file.info('2Peter.txt')$size)

hebrews <-  readChar('Hebrews.txt', file.info('Hebrews.txt')$size)

titus <-  readChar('Titus.txt', file.info('Titus.txt')$size)
timothy_1 <-  readChar('1Timothy.txt', file.info('1Timothy.txt')$size)
timothy_2 <-  readChar('2Timothy.txt', file.info('2Timothy.txt')$size)
philemon <-  readChar('Philemon.txt', file.info('Philemon.txt')$size)


ephesians <-  readChar('Ephesians.txt', file.info('Ephesians.txt')$size)
romans <- readChar('Romans.txt', file.info('Romans.txt')$size)

corinthians_1 <- readChar('1Corinthians.txt', file.info('1Corinthians.txt')$size)
corinthians_2 <- readChar('2Corinthians.txt', file.info('2Corinthians.txt')$size)

galatians <- readChar('Galatians.txt', file.info('Galatians.txt')$size)
james <- readChar('James.txt', file.info('James.txt')$size)

revealation <- readChar('Revelation.txt', file.info('Revelation.txt')$size)


df1 <-  data.frame(rbind(matthew, mark, luke, john,acts, peter_1 ,peter_2,hebrews ,titus,timothy_1,timo

data_corpus <- Corpus(DataframeSource(df1))
data_corpus <- tm_map(data_corpus, content_transformer(tolower))
data_corpus <- tm_map(data_corpus, removePunctuation)
data_corpus <- tm_map(data_corpus, removeNumbers)
data_corpus <- tm_map(data_corpus, removeWords, stopwords("en"))

data_corpus <- tm_map(data_corpus,stemDocument)
data_corpus <-tm_map(data_corpus,stripWhitespace)
```

```
data_corpus <- tm_map(data_corpus, PlainTextDocument)


tdm <- DocumentTermMatrix(data_corpus)
train <- as.matrix(tdm) %>%
  as.data.frame()

row.names(train) <- c('matthew','mark','luke','john','acts' , 'peter_1' ,'peter_2' ,'hebrews', 'titus',


clusters <- hclust(dist(train))
plot(clusters, xlab='Books in the New Testament')
```

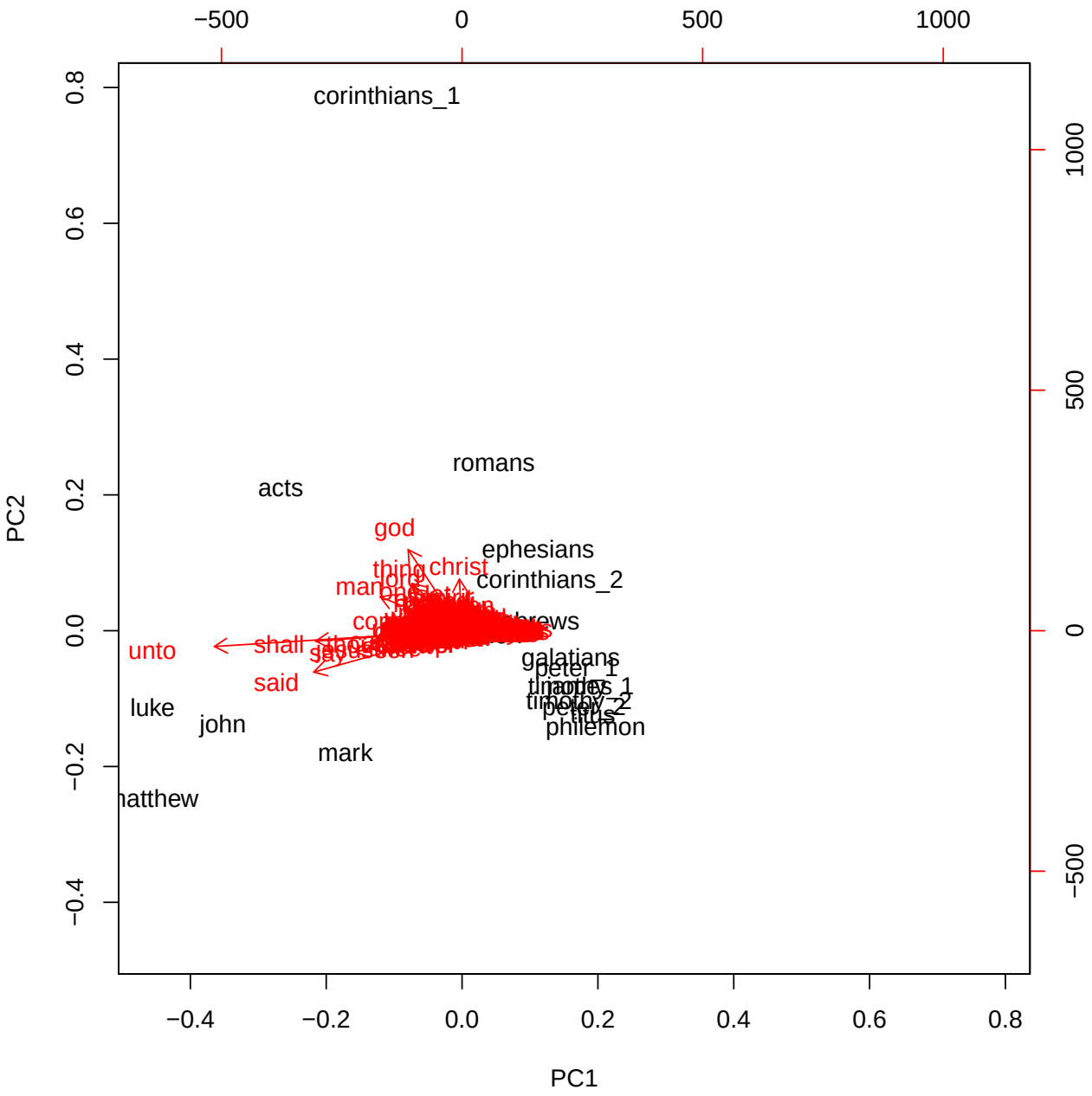



```
PCA_NT <-  prcomp(train)
summary(PCA_NT)
```

```
## Importance of components:
##                             PC1      PC2      PC3      PC4      PC5
## Standard deviation     280.4326 109.4305 77.47429 76.19246 55.73029
```

```
## Proportion of Variance   0.7028   0.1070  0.05364  0.05188  0.02776
## Cumulative Proportion    0.7028   0.8099  0.86350  0.91538  0.94314
##                            PC6      PC7      PC8      PC9     PC10
## Standard deviation     42.4452 38.54732 28.67757 26.37052 23.63955
## Proportion of Variance  0.0161  0.01328  0.00735  0.00621  0.00499
## Cumulative Proportion   0.9592  0.97252  0.97987  0.98608  0.99108
##                           PC11     PC12     PC13     PC14    PC15    PC16
## Standard deviation     19.17353 13.29522 11.04855 10.63136 9.73517 7.41624
## Proportion of Variance  0.00329  0.00158  0.00109  0.00101 0.00085 0.00049
## Cumulative Proportion   0.99436  0.99594  0.99704  0.99805 0.99889 0.99938
##                           PC17    PC18      PC19
## Standard deviation     6.39028 5.30040 2.184e-12
## Proportion of Variance 0.00036 0.00025 0.000e+00
## Cumulative Proportion  0.99975 1.00000 1.000e+00
```

```r
biplot(PCA_NT)
```


−500
0
500
1000
corinthians_1
romans
acts
god
ephesians
christ
corinthians_2
hebrews
unto
shall
said
galatians
peter_1
luke
john
mark
philemon
matthew
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
PC2
PC1
1000
500
0
−500