1 Topic modelling with NMF

```
Ptilde <- as.matrix( read.csv("data/news.csv") )
P <- Ptilde/sum(Ptilde)
showcol(pmin(P,0.0001))
```

a)

```
r <- 4
lr.gkl <- lee01.gkl(P, r, reps=5)
with(lr.gkl,
     for (k in 1:nrow(R)) {
         print(rev(sort(R[k,]))[1:10])
         cat(strrep('-',130), "\n")
     })
```

```
ggplotm(lr.gkl$R[,1:10], format="", show.axis=FALSE, mid="black")
```

b)

```
Phat <- lr.gkl$L %*% lr.gkl$R
showcol(pmin(Phat,0.0001))
```

c)

ii)

```
Phat <- P.svd$u[,1:4] %*% diag(P.svd$d[1:4]) %*% t(P.svd$v[,1:4])
showcol(pmin(Phat,0.0001))
```

```
r <- 2
lr.gkl <- lee01.gkl(P, r, reps=5)
with(lr.gkl,
     for (k in 1:nrow(R)) {
         print(rev(sort(R[k,]))[1:10])
         cat(strrep('-',130), "\n")
     })
```

```
Phat <- lr.gkl$L %*% lr.gkl$R
showcol(pmin(Phat,0.0001))
```

ii)

```
r <- 8
lr.gkl <- lee01.gkl(P, r, reps=5)
with(lr.gkl,
     for (k in 1:nrow(R)) {
         print(rev(sort(R[k,]))[1:10])
         cat(strrep('-',130), "\n")
     })


Phat <- lr.gkl$L %*% lr.gkl$R
showcol(pmin(Phat,0.0001))
```

e)

```
r <- 4
lr.gnmf <- lee01.gnmf(P, r, reps=5)
with(lr.gnmf,
     for (k in 1:nrow(R)) {
         print(rev(sort(R[k,]))[1:10])
         cat(strrep('-',130), "\n")
     })


Phat <- lr.gnmf$L %*% lr.gnmf$R

showcol(pmin(Phat,0.0001))
```

2 PLSA

a)

```
lsr.gkl <- nmf.lsr(lr.gkl)
summary(lsr.gkl)
  Length Class     Mode
L 1600   -none-    numeric
S   16   ddiMatrix S4
R 3200   -none-    numeric
```

```
> apply(lsr.gkl$L,2,sum)
[1] 1 1 1 1
> sum(lsr.gkl$S)
[1] 0.999818
> apply(lsr.gkl$R,1,sum)
[1] 1 1 1 1
```

b)

```
slr.gkl <- nmf.slr(lr.gkl)
summary(slr.gkl)
  Length Class     Mode
S 160000 ddiMatrix S4
L   1600 -none-    numeric
R   3200 -none-    numeric
```

```
> sum(slr.gkl$S)
[1] 0.9998147
> apply(slr.gkl$L,1,sum)
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [44] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [87] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[130] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[173] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[216] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[259] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[302] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[345] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

```
[388] 1 1 1 1 1 1 1 1 1 1 1 1 1
> apply(slr.gkl$R,1,sum)
[1] 1 1 1 1
```

3 Clustering

a)

```
cluster <-kmeans(P, 4, nstart=100)$cluster
```

```
> cluster
  [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [65] 3 3 4 3 3 3 3 3 3 3 3 3 3 3 3 3 4 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[129] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[193] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[257] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[321] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[385] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

Relable:

```
i<-1
while(i<101){

  if(cluster[i]==1){cluster[i]<-4

  }else if(cluster[i]==2){cluster[i]<-2

  }else if(cluster[i]==3){cluster[i]<-3

  }else{cluster[i]<-1

  }

  i<-i+1
}
trail_a<- cm(cluster)

trail_a$overall["Accuracy"]
```

b)

```
P.svd <- svd(P)

test2<- P.svd$u[,1:4]%*%diag(P.svd$d[1:4])

cluster <-kmeans(test2, 4, nstart=100)$cluster
```

```
> cluster
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 4 1 1 1 1 1 1 1 1 1 1 1
 [65] 1 1 4 1 1 1 1 1 1 1 1 1 1 1 1 1 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[129] 1 1 1 1 1 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1
[193] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[257] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[321] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 4 1 4 1 1 4 1 4 4 1 1 1 1 1 1 4 4 1 1 1 1 1 4 1 1 1 1 1 1 1 4 1 1 1 1 4 1 1 1 1 1
[385] 1 1 1 1 1 1 1 1 1 1 4 1 1 1 1 1
```

Relabel:

```
i<-1
while(i<101){

  if(cluster[i]==1){cluster[i]<-4

  }else if(cluster[i]==2){cluster[i]<-2

  }else if(cluster[i]==3){cluster[i]<-3

  }else{cluster[i]<-1

  }

  i<-i+1
}


trail_b<- cm(cluster)

trail_b$overall["Accuracy"]
```

c)

```
r <- 4

lr.gkl <- lee01.gkl(P, r, reps=5)

cluster <-kmeans(lr.gkl$L, 4, nstart=100)$cluster
```

```
> cluster
  [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [65] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 4 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[129] 2 2 2 2 2 4 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 4 2 2 2 2 2 2 2 2 2 2 2
[193] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[257] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

```
[321] 2 2 2 1 2 1 2 2 2 2 2 1 2 2 2 1 1 2 2 2 2 2 1 1 1 2 2 1 2 1 1 2 2 2 1 2 2 1 1 2 1 2 2 2 1 2 2 2 2 2 2 2 2 1 2 2 2 2 1 2 2 2 2 2
[385] 2 2 2 2 1 2 2 2 2 2 1 2 2 1 1 2
```

Relabel:

```
i<-1
while(i<101){

  if(cluster[i]==1){cluster[i]<-4

  }else if(cluster[i]==2){cluster[i]<-2

  }else if(cluster[i]==3){cluster[i]<-3

  }else{cluster[i]<-1

  }

  i<-i+1
}


trail_c<- cm(cluster)

trail_c$overall["Accuracy"]
```

d)

```
lsr.gkl <- nmf.lsr(lr.gkl)

cluster <-kmeans(lsr.gkl$L, 4, nstart=100)$cluster
```

```
> cluster
  [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 3 3 3 3 3 3 3 3 3 3
 [65] 3 3 4 3 3 3 3 3 3 3 3 3 3 3 3 3 4 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[129] 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3
[193] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[257] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[321] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[385] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

Relabel:

```
i<-1
while(i<101){

  if(cluster[i]==1){cluster[i]<-3

  }else if(cluster[i]==2){cluster[i]<-2

  }else if(cluster[i]==3){cluster[i]<-4

  }else{cluster[i]<-1

  }

  i<-i+1
}


trail_d<- cm(cluster)

trail_d$overall["Accuracy"]
```

e)

```
slr.gkl <- nmf.slr(lr.gkl)

cluster <-kmeans(slr.gkl$L, 4, nstart=100)$cluster
```

```
> cluster
  [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 3 2 2 2 2 2 4 2 2 2 2 2 2 2 2 2 2 2 2 2 2 4 2 2 2 2 2 2
 [65] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 4 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[129] 1 1 1 1 1 1 1 1 1 3 1 1 1 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 4
[193] 1 1 1 1 1 1 1 1 4 4 4 4 4 4 4 2 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[257] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[321] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3
[385] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

Relabel:

```
i<-1
while(i<101){

  if(cluster[i]==1){cluster[i]<-2

  }else if(cluster[i]==2){cluster[i]<-1

  }else if(cluster[i]==3){cluster[i]<-4
```

```
  }else{cluster[i]<-3
  }
  i<-i+1
}

trail_e<- cm(cluster)
trail_e$overall["Accuracy"]
```