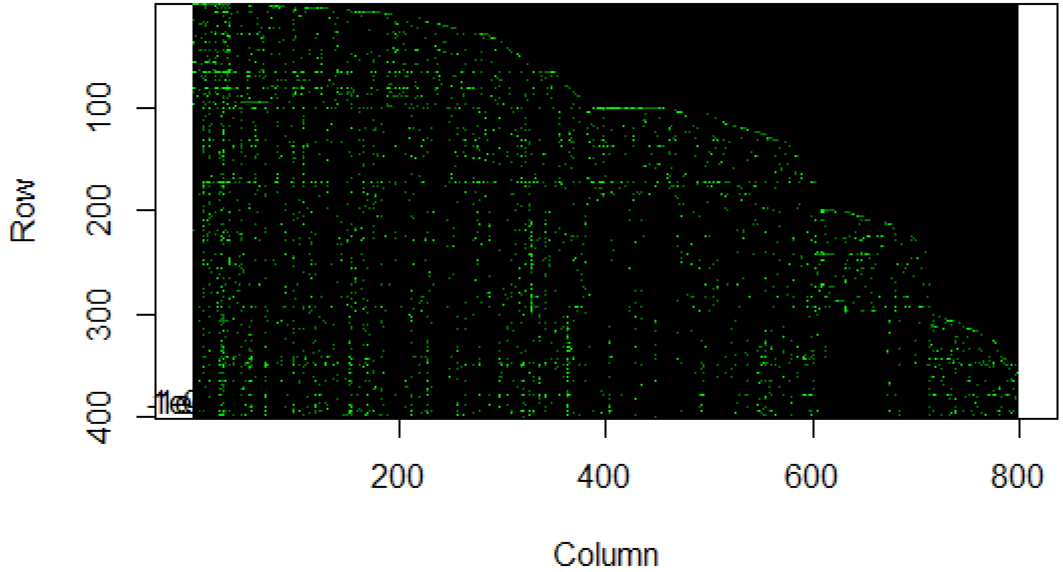
1 Topic modelling with NMF

I load the document-term matrix. And then as the instruction in the file, I replace word frequencies by "probabilities".

I present the matrix as a colored panel. And, as the instruction, I truncate values >0.0001 for imporoved visibility. It shows as below.

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



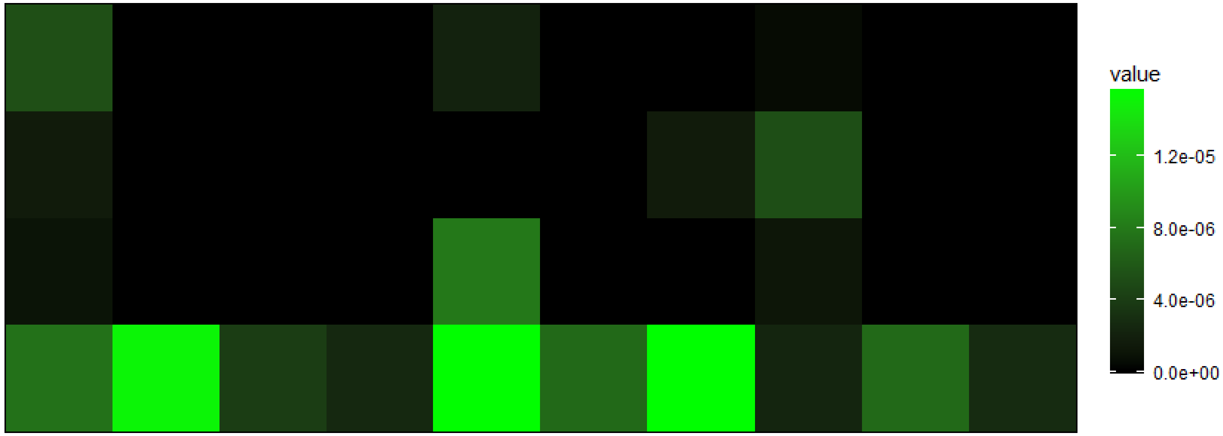
a)

Look into the top-10 terms in the matrix R-hat. It is presented as below.

```
       space       launch        orbit      mission         nasa       shuttl         venu       system         year        earth 
4.675926e-05 2.791249e-05 2.266385e-05 2.242116e-05 2.123164e-05 1.980099e-05 1.765422e-05 1.507907e-05 1.399894e-05 1.380021e-05 
---------------------------------------------------------------------------------------------------------------------------------
       studi       diseas       effect       doctor        medic      candida      patient        peopl         drug         food 
1.854399e-05 1.851442e-05 1.810447e-05 1.806254e-05 1.783681e-05 1.670837e-05 1.603073e-05 1.566815e-05 1.456400e-05 1.354655e-05 
---------------------------------------------------------------------------------------------------------------------------------
```

```
        god     christian         peopl        church      homosexu          paul          jesu         faith         thing      question 
6.244223e-05 4.252302e-05 3.511040e-05 3.357088e-05 2.327599e-05 1.984683e-05 1.969462e-05 1.946757e-05 1.711896e-05 1.562345e-05 
--------------------------------------------------------------------------------------------------------------------------------------
        kei       encrypt        system         secur        govern          chip       clipper           law         peopl            de 
5.834261e-05 4.766727e-05 2.882645e-05 2.870541e-05 2.816826e-05 2.627861e-05 2.260502e-05 2.119552e-05 2.072403e-05 1.641549e-05 
--------------------------------------------------------------------------------------------------------------------------------------
```

I present it as a colored panel as below. I present the value for each row corresponding to the first 10 terms in different color. It is somehow clear to see the 4 X 10 squares with different level of darkness in color green.



And the top 10 terms for each row are as follow.

ROW 1	space	launch	orbit	mission	nasa	shuttl	venu	system	year	earth
	0.0000468	0.0000279	0.0000227	0.0000224	0.0000212	0.0000198	0.0000177	0.0000151	0.0000140	0.0000138
ROW 2	studi	diseas	effect	doctor	medic	candida	patient	peopl	drug	food
	0.0000185	0.0000185	0.0000181	0.0000181	0.0000178	0.0000167	0.0000160	0.0000157	0.0000146	0.0000135
ROW 3	god	christian	peopl	church	homosexu	paul	jesu	faith	thing	question
	0.0000624	0.0000425	0.0000351	0.0000336	0.0000233	0.0000198	0.0000197	0.0000195	0.0000171	0.0000156
ROW 4	kei	encrypt	system	secur	govern	chip	clipper	law	peopl	de
	0.0000583	0.0000477	0.0000288	0.0000287	0.0000282	0.0000263	0.0000226	0.0000212	0.0000207	0.0000164

As we know ahead that the topics are “sci.crypt”, “sci.med”, “sci.space”, and “soc.religion.christian”, it is easier to see the topic for each row. With the indication of <u>those words colored with orange cell</u>, it is easy to see that <u>for ROW 1, it’s</u>

about “sci.space”; for ROW 2, it’s about “sci.med”: for ROW 3, it’s about “soc.religion.christian”; for ROW 4, it’s about “sci.crypt”. So, in general, I think the subsets of each row constitute a meaningful topic respectively.

```
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")
```

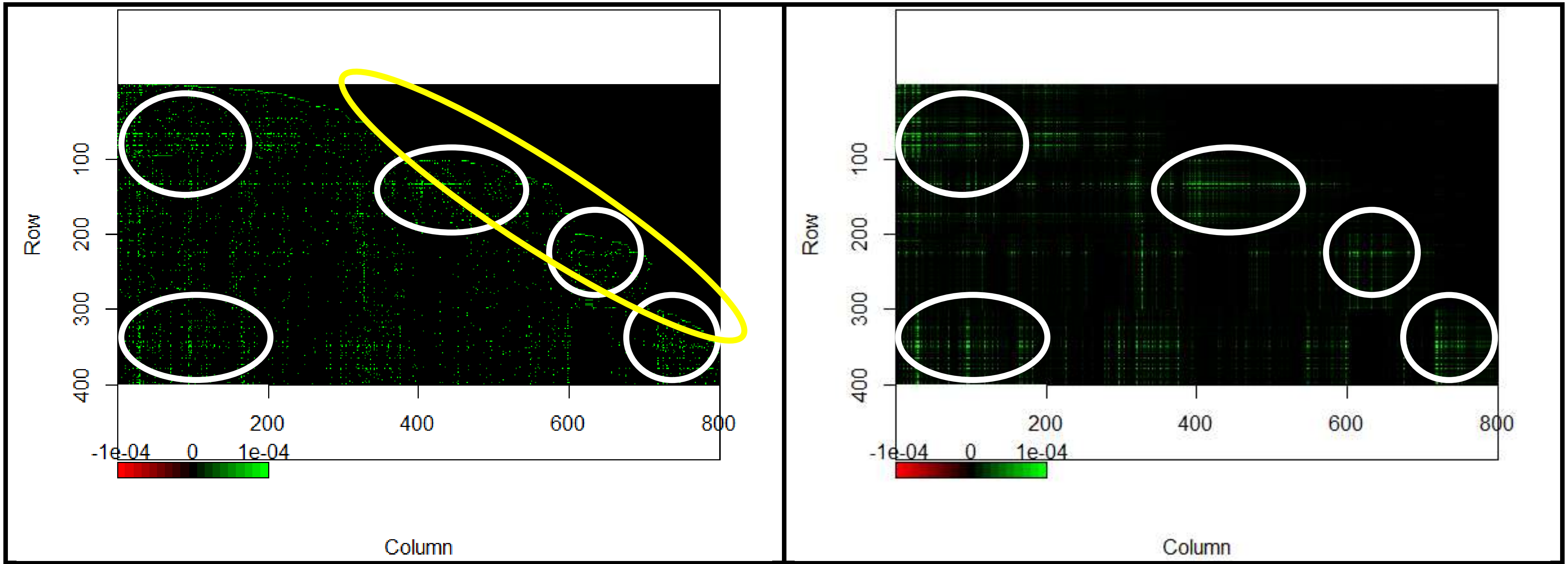
b)

I use the code provided in the file to reconstruct the matrix.

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

I compare both of them as below.

Original matrix	Reconstructed matrix (Using KL divergence with r = 4)



After the observation, the result is quite as what I expected, which is **i)** the dense (high-value) part will be presented, and also that **ii)** for some cells in the reconstructed matrix, the value will be a bit larger or smaller comparing with the corresponding cells in the original matrix.

i) This point is showed as the white circles, which points out the dense (high-value) part.

ii) This point is showed as the yellow circle. For the original matrix, almost all of the cells at the right-up side are zero(empty), and the yellow circle is pointing out the clear line distinguishing two area. On the opposite, the line between two area is much more blurry.

c)

i)

Use SVD to redo the process as in a) and b).

Look into the top-10 terms(using the absolutes of each cell) in the matrix V. It is presented as below.

```
      venu      space     soviet      probe     system    mission       year      studi    program      earth
-0.3467093 -0.2408041 -0.2279326 -0.2041376 -0.1901447 -0.1813604 -0.1564498 -0.1438261 -0.1335832 -0.1298500
---------------------------------------------------------------------------------------------------------------
      venu     soviet      probe      space     cancer    mission       drug     diseas       diet      peopl
-0.3843787 -0.2514368 -0.2209769 -0.2148817  0.1793495 -0.1781779  0.1768813  0.1715696  0.1685100  0.1492283
---------------------------------------------------------------------------------------------------------------
   encrypt        kei        law     cancer       diet     diseas      secur      devic       chip     health
 0.3817868  0.2849342  0.2344941 -0.2042601 -0.1884720 -0.1883962  0.1843405  0.1754694  0.1627822 -0.1616573
---------------------------------------------------------------------------------------------------------------
       god   homosexu  christian    encrypt       paul        sin      peopl        kei       jesu     church
-0.5239122 -0.2421823 -0.2380392  0.2224059 -0.2172073 -0.1901181 -0.1890682  0.1737290 -0.1469805 -0.1180113
---------------------------------------------------------------------------------------------------------------
```

I adjust the presentation as the form below.

ROW 1	venu	space	soviet	probe	system	mission	year	studi	program	earth
	(0.35)	(0.24)	(0.23)	(0.20)	(0.19)	(0.18)	(0.16)	(0.14)	(0.13)	(0.13)
ROW 2	venu	soviet	probe	space	cancer	mission	drug	diseas	diet	peopl
	(0.38)	(0.25)	(0.22)	(0.21)	0.18	(0.18)	0.18	0.17	0.17	0.15
ROW 3	encrypt	kei	law	cancer	diet	diseas	secur	devic	chip	health
	0.38	0.28	0.23	(0.20)	(0.19)	(0.19)	0.18	0.18	0.16	(0.16)
ROW 4	god	homosexu	christian	encrypt	paul	sin	peopl	kei	jesu	church
	(0.52)	(0.24)	(0.24)	0.22	(0.22)	(0.19)	(0.19)	0.17	(0.15)	(0.12)

As the information above, the negative values are telling that the related topic is “NOT” about certain term, and the positive values are telling that the related topic is “HAVING SOMETHING TO DO” with certain term. By this way, we can see that for ROW 2, it’s about “sci.med” ; and for ROW 3, it’s about “sci.crypt”. But it’s hard to tell what topics for the rest two row are related to.

So, I try to “not to use absolutes”, and I use the original value for the selection of the top 10 terms for each row. And the result is presented below.

ROW 1	omiss	strnlghtc	suno	kipl	geb	collision	metzger	pmetzger	idealist	ec
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
ROW 2	cancer	drug	diseas	diet	peopl	encrypt	health	studi	kei	effect
	0.18	0.18	0.17	0.17	0.15	0.15	0.14	0.13	0.12	0.12

ROW 3	encrypt	kei	law	secur	devic	chip	protect	govern	clipper	god
	0.38	0.28	0.23	0.18	0.18	0.16	0.16	0.16	0.13	0.13
ROW 4	encrypt	kei	devic	secur	protect	govern	chip	cancer	drug	privaci
	0.22	0.17	0.10	0.10	0.10	0.09	0.09	0.09	0.08	0.07

For this trial, it is clear that for ROW 2, it's about "sci.med". But some terms are causing problem while allocating the topics. For example, the term "encrypt" appears in both row 3 and 4 with a "not so low" value. And also there is still no telling that what topic it is for ROW 1; though there is term "collision" in the top 10, but the value is too low to have a efficient indication.

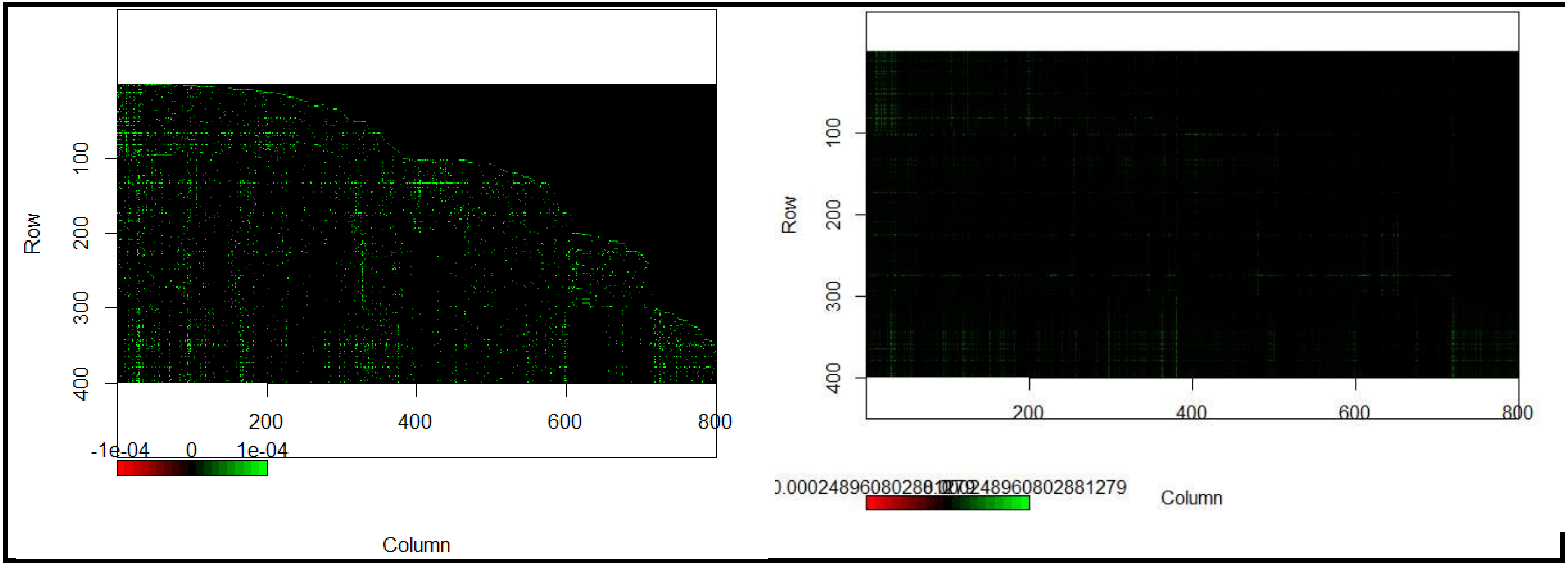
ii)

I use the code provided in the file to reconstruct the matrix.

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

I compare both of them as below.

Original matrix	Reconstructed matrix (Using SVD with k = 4)



It’s apparent that the reconstruction of the matrix using SVD is having less similarity comparing to the result using KL-divergence method.

d)

i)

Run the process with r =2.

```
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")
     })
```

```
        kei        space        system       encrypt        launch        govern        secur         data         chip       mission
```

```
5.642955e-05 4.850925e-05 4.576931e-05 4.479688e-05 2.895709e-05 2.722350e-05 2.697668e-05 2.513996e-05 2.500510e-05 2.351206e-05
------------------------------------------------------------------------------------------------------------------------------------
         god        peopl    christian       church     question     homosexu        thing         paul        studi       person
6.149216e-05 4.855624e-05 4.187606e-05 3.305997e-05 2.360997e-05 2.292185e-05 2.241282e-05 2.211637e-05 2.175841e-05 1.969726e-05
------------------------------------------------------------------------------------------------------------------------------------
```

Adjust the presentation as below.

ROW 1	kei	space	system	encrypt	launch	govern	secur	data	chip	mission
	0.000056	0.000049	0.000046	0.000045	0.000029	0.000027	0.000027	0.000025	0.000025	0.000024
ROW 2	god	peopl	christian	church	question	homosexu	thing	paul	studi	person
	0.000061	0.000049	0.000042	0.000033	0.000024	0.000023	0.000022	0.000022	0.000022	0.000020

It is seemingly that the “sci.crypt”(orange) and “sci.space”(purple) are becoming the mixed topic for ROW 1 collected terms; and the “sci.med”(green) and “soc.religion.christian”(red) are the mixed topic for ROW 2 collected terms.

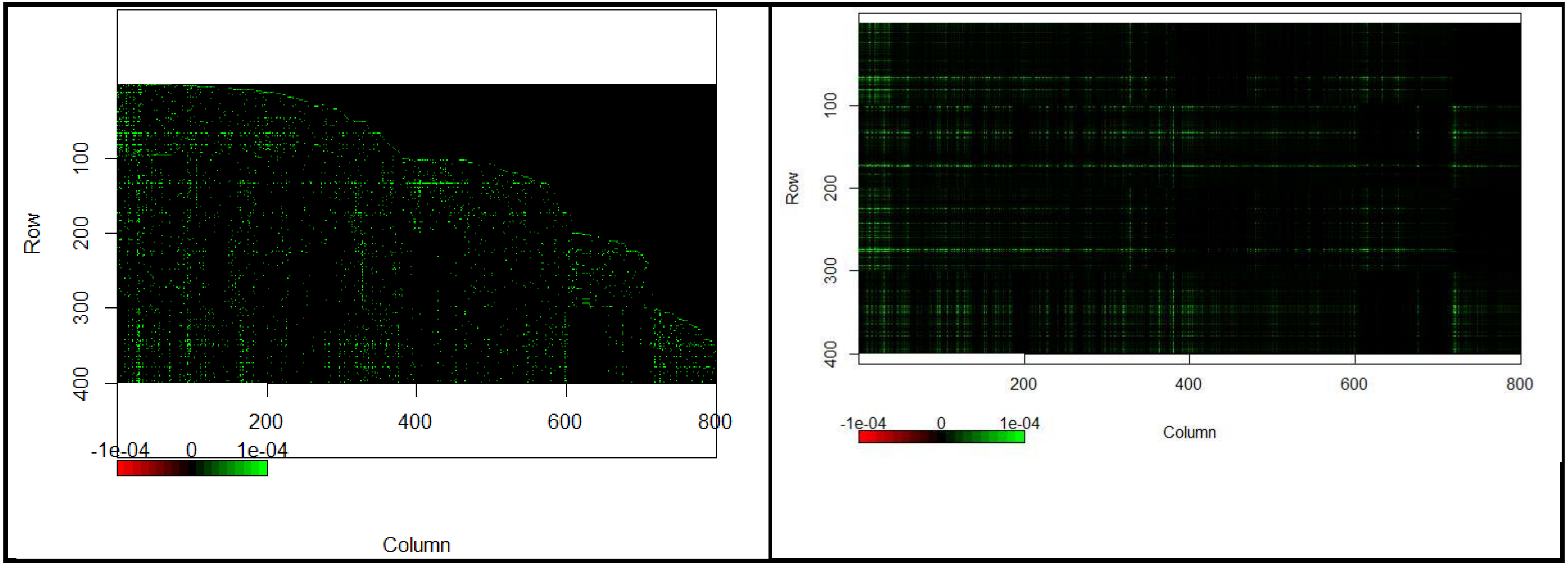
The overlap terms between two rows is not clear, at least for the top10 terms. So I will say that the performance using r =2 under this method is quite good.

<u>For reconstruction of the matrix:</u>

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

I compare both of them as below.

Original matrix	Reconstructed matrix (Using KL divergence with r =2)



It shows that using r = 2, has a blurry result comparing to the original matrix. Because the main point for any kind of method is 1) to have an efficient way to distinguish the topics between documents, and also 2) to reconstruct the matrix as close as it possibly can get. So, I will say that for the 1^{st} purpose, using r =2 is still work, but for 2^{ed} purpose, it's not good to use r =2.

ii)

Run the process with r =8.

```
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")
     })
```

```
       space       launch        orbit      mission         nasa       shuttl         venu      station       soviet       option
3.482521e-05 2.267867e-05 1.836416e-05 1.802477e-05 1.686169e-05 1.553478e-05 1.434391e-05 1.061032e-05 1.046705e-05 1.045919e-05
------------------------------------------------------------------------------------------------------------------------------------
      church    christian        peopl        group     religion         bibl       cathol        thing     question        faith
3.065920e-05 2.118346e-05 1.897872e-05 1.219438e-05 1.074609e-05 9.480580e-06 8.876667e-06 8.826587e-06 8.702235e-06 8.678640e-06
------------------------------------------------------------------------------------------------------------------------------------
      govern    administr      clinton          law      protect      privaci         drug        devic        legal         data
2.151724e-05 1.413485e-05 1.377205e-05 1.195599e-05 1.177867e-05 1.022457e-05 1.009080e-05 9.889989e-06 9.687329e-06 9.492221e-06
------------------------------------------------------------------------------------------------------------------------------------
      system       scienc       comput           de         data    scientist     scientif         part      protein      program
2.319916e-05 1.551924e-05 1.236823e-05 1.153659e-05 9.509902e-06 8.990667e-06 8.552133e-06 8.008130e-06 7.698934e-06 7.102140e-06
------------------------------------------------------------------------------------------------------------------------------------
         god     homosexu    christian         paul         jesu        peopl          sin         word          law        faith
5.462475e-05 2.076040e-05 1.961800e-05 1.929040e-05 1.707111e-05 1.427537e-05 1.377295e-05 1.137034e-05 1.018813e-05 9.718540e-06
------------------------------------------------------------------------------------------------------------------------------------
         kei      encrypt        secur         chip      clipper       system        peopl       escrow       public          law
4.814733e-05 3.637384e-05 2.640802e-05 1.992934e-05 1.980732e-05 1.693102e-05 1.311337e-05 1.197076e-05 1.128866e-05 1.027552e-05
------------------------------------------------------------------------------------------------------------------------------------
       henri         test        thing          net      toronto          hst        water          pat         high         tast
1.226398e-05 1.161085e-05 1.108025e-05 1.007863e-05 9.119414e-06 8.175726e-06 7.981435e-06 7.924872e-06 7.776659e-06 7.298430e-06
------------------------------------------------------------------------------------------------------------------------------------
      diseas       doctor        medic      candida       effect      patient        studi         food         diet       health
1.478210e-05 1.442238e-05 1.389079e-05 1.334113e-05 1.283070e-05 1.280005e-05 1.225277e-05 1.081659e-05 1.045623e-05 1.019070e-05
------------------------------------------------------------------------------------------------------------------------------------
```

Adjust the presentation as below.

ROW 1	space	launch	orbit	mission	nasa	shuttl	venu	station	soviet	option
	0.0000348	0.0000227	0.0000184	0.0000180	0.0000169	0.0000155	0.0000143	0.0000106	0.0000105	0.0000105
ROW 2	church	christian	peopl	group	religion	bibl	cathol	thing	question	faith
	0.0000307	0.0000212	0.0000190	0.0000122	0.0000107	0.0000095	0.0000089	0.0000088	0.0000087	0.0000087
ROW 3	govern	administr	clinton	law	protect	privaci	drug	devic	legal	data
	0.0000215	0.0000141	0.0000138	0.0000120	0.0000118	0.0000102	0.0000101	0.0000099	0.0000097	0.0000095
ROW 4	system	scienc	comput	de	data	scientist	scientif	part	protein	program
	0.0000232	0.0000155	0.0000124	0.0000115	0.0000095	0.0000090	0.0000086	0.0000080	0.0000077	0.0000071
ROW 5	god	homosexu	christian	paul	jesu	peopl	sin	word	law	faith
	0.0000546	0.0000208	0.0000196	0.0000193	0.0000171	0.0000143	0.0000138	0.0000114	0.0000102	0.0000097
ROW 6	kei	encrypt	secur	chip	clipper	system	peopl	escrow	public	law
	0.0000481	0.0000364	0.0000264	0.0000199	0.0000198	0.0000169	0.0000131	0.0000120	0.0000113	0.0000103
ROW 7	henri	test	thing	net	toronto	hst	water	pat	high	tast
	0.0000123	0.0000116	0.0000111	0.0000101	0.0000091	0.0000082	0.0000080	0.0000079	0.0000078	0.0000073
ROW 8	diseas	doctor	medic	candida	effect	patient	studi	food	diet	health
	0.0000148	0.0000144	0.0000139	0.0000133	0.0000128	0.0000128	0.0000123	0.0000108	0.0000105	0.0000102

If we still hold only 4 topics as “sci.crypt”, “sci.med”, “sci.space”, and “soc.religion.christian”, then it is easier to see which row belongs to which topic. With the indication of those words with different colored, it is easy to see that for ROW 1 and ROW 4, it’s about “sci.space”; for ROW 2 and ROW 5, it’s about “soc.religion.christian”; for ROW 3 and ROW 6, it’s about “sci.crypt”; for ROW 7 and ROW 8, it’s about “sci.med”. And the most important part is that the OVERLAPE term should be as less as it can be, then we can say that it’s an efficient way to split.

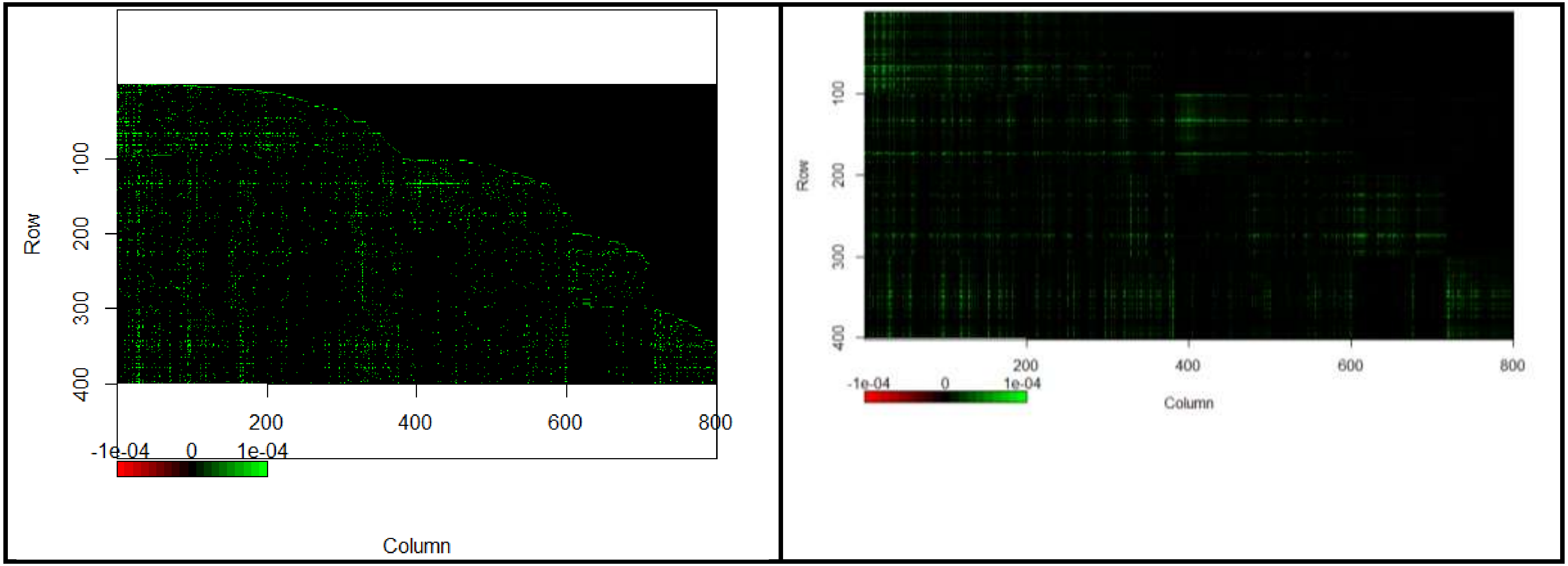
And for all the terms above(total 80 terms), only 8 terms are overlapped, which are "data", "christian", "peopl", "law", "faith", "system", and "thing". So we can see this as a 8 topics division for all 400 documents; and that it’s efficient to use r =8.

For reconstruction of the matrix:

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

I compare both of them as below.

Original matrix	Reconstructed matrix (Using KL divergence with r =8)



As expected, the two graph are much more similar since we use a higher rank (as r =8).

e)

Use Guassian NMF:

```
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")
     })
         venu        soviet         space         probe       mission         earth        launch         orbit        explor          year 
7.865402e-05 5.158317e-05 4.977810e-05 4.581903e-05 3.905513e-05 2.834690e-05 2.521950e-05 2.401224e-05 2.229390e-05 1.932836e-05 
----------------------------------------------------------------------------------------------------------------------------------
          god         peopl      homosexu     christian          paul           sin          jesu           law        church         faith 
7.438044e-05 3.565252e-05 3.503802e-05 3.482337e-05 3.248718e-05 2.739292e-05 2.195861e-05 1.864565e-05 1.774603e-05 1.699184e-05 
----------------------------------------------------------------------------------------------------------------------------------
```

```
      cancer       diseas         diet         drug       health        studi       effect        medic      patient           dr 
4.408339e-05 4.154017e-05 4.108598e-05 4.027694e-05 3.550131e-05 3.267301e-05 2.976185e-05 2.747793e-05 2.632505e-05 2.437573e-05 
-----------------------------------------------------------------------------------------------------------------------------------
     encrypt          kei          law        secur        devic      protect         chip       govern       system      clipper 
6.228736e-05 4.885666e-05 3.089337e-05 3.021957e-05 2.855234e-05 2.763946e-05 2.682029e-05 2.652668e-05 2.176663e-05 2.148359e-05 
-----------------------------------------------------------------------------------------------------------------------------------
```
Adjust the presentation as below.

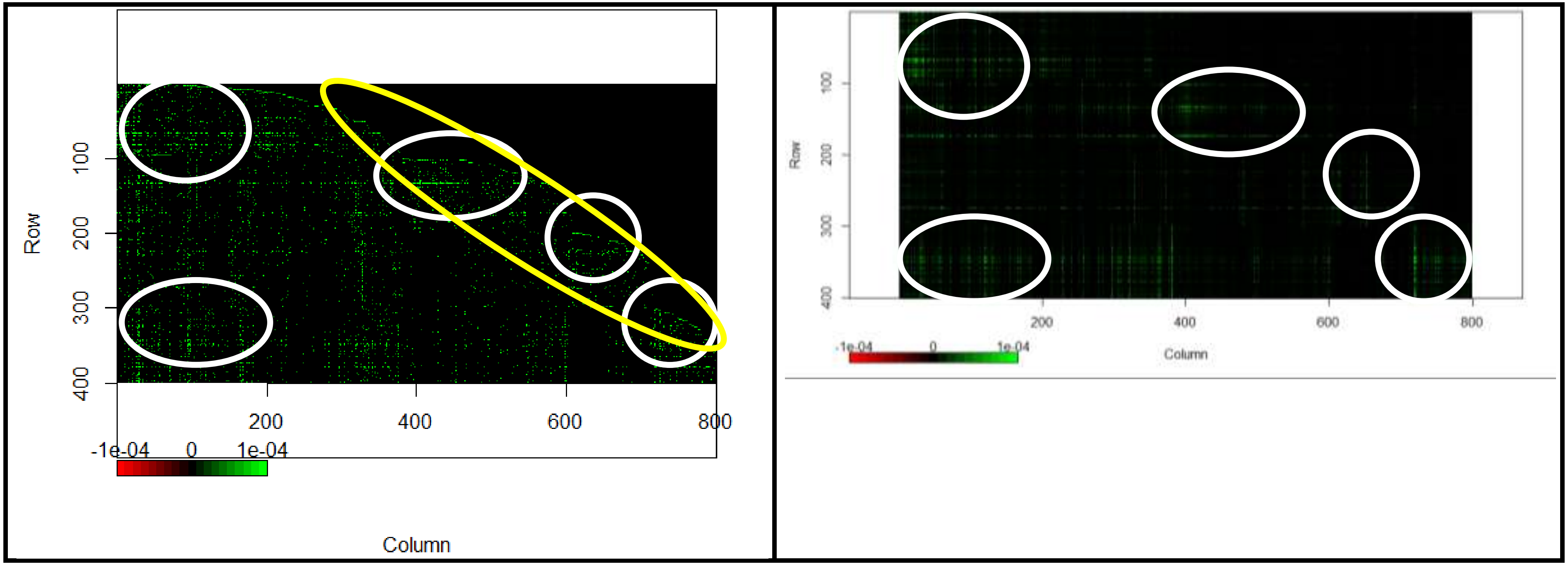
ROW 1	venu	soviet	space	probe	mission	earth	launch	orbit	explor	year
	0.000079	0.000052	0.000050	0.000046	0.000039	0.000028	0.000025	0.000024	0.000022	0.000019
ROW 2	god	peopl	homosexu	christian	paul	sin	jesu	law	church	faith
	0.000074	0.000036	0.000035	0.000035	0.000032	0.000027	0.000022	0.000019	0.000018	0.000017
ROW 3	cancer	diseas	diet	drug	health	studi	effect	medic	patient	dr
	0.000044	0.000042	0.000041	0.000040	0.000036	0.000033	0.000030	0.000027	0.000026	0.000024
ROW 4	encrypt	kei	law	secur	devic	protect	chip	govern	system	clipper
	0.000062	0.000049	0.000031	0.000030	0.000029	0.000028	0.000027	0.000027	0.000022	0.000021

From the result above, it's easy to see that for ROW 1, it's about "sci.space"; for ROW 2, it's about "soc.religion.christian"; for ROW 3, it's about "sci.med"; for ROW 4, it's about "sci.crypt". So, for this part, I think it is also efficient to use this method.

For reconstruction of the matrix:

```
Phat <- lr.gnmf$L %*% lr.gnmf$R
showcol(pmin(Phat,0.0001))
```
I compare both of them as below.

Original matrix	Reconstructed matrix (Using Gaussian NMF with r =4)



It can also show the dense part, but still have the blurry boundary problem.

Which NMF variant produces better results?

ANS:

Compare the two method I've used before:

i) Using Guassian NMF:

space	venu	soviet	space	probe	mission	earth	launch	orbit	explor	year
	0.000079	0.000052	0.000050	0.000046	0.000039	0.000028	0.000025	0.000024	0.000022	0.000019
religion	god	peopl	homosexu	christian	paul	sin	jesu	law	church	faith
	0.000074	0.000036	0.000035	0.000035	0.000032	0.000027	0.000022	0.000019	0.000018	0.000017

med	cancer	diseas	diet	drug	health	studi	effect	medic	patient	dr
	0.000044	0.000042	0.000041	0.000040	0.000036	0.000033	0.000030	0.000027	0.000026	0.000024
crypt	encrypt	kei	law	secur	devic	protect	chip	govern	system	clipper
	0.000062	0.000049	0.000031	0.000030	0.000029	0.000028	0.000027	0.000027	0.000022	0.000021

ii) Using KL divergence:

space	space	launch	orbit	mission	nasa	shuttl	venu	system	year	earth
	0.000047	0.000028	0.000023	0.000022	0.000021	0.000020	0.000018	0.000015	0.000014	0.000014
religion	god	christian	peopl	church	homosexu	paul	jesu	faith	thing	question
	0.000062	0.000043	0.000035	0.000034	0.000023	0.000020	0.000020	0.000020	0.000017	0.000016
med	studi	diseas	effect	doctor	medic	candida	patient	peopl	drug	food
	0.000019	0.000019	0.000018	0.000018	0.000018	0.000017	0.000016	0.000016	0.000015	0.000014
crypt	kei	encrypt	system	secur	govern	chip	clipper	law	peopl	de
	0.000058	0.000048	0.000029	0.000029	0.000028	0.000026	0.000023	0.000021	0.000021	0.000016

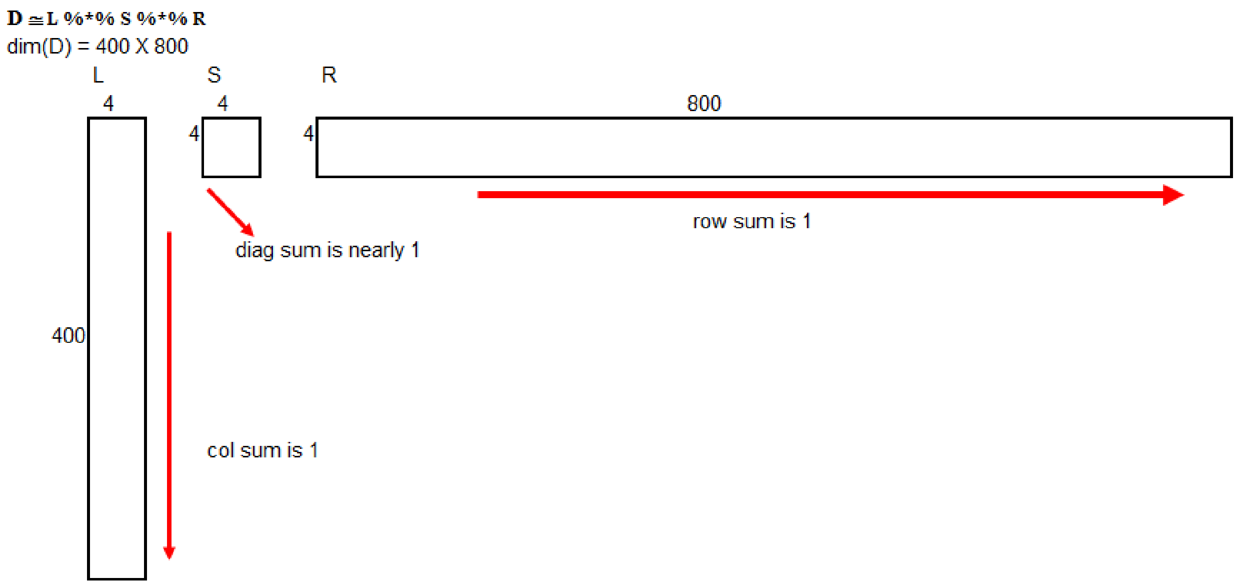
If we focus on the value for the top 10 terms under each topics, we can see that the values using Guassian NMF is much more higher than those of using KL divergence in general. And one of the purpose of these methodology is to have an efficient way or distinguishable values to split the documents into sub groups, and that depends on the calculated values as mentioned. So I will say that within the methodology I’ve tried, Guassian NMF is a better choice.

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
```

I acquire some traits of matrix L’, S’, and R’.
I show the traits as below.



So to enforce a probability concept, I can put some explanation in to the product of these three matrixs.

Let Di refer to document(i); Wi refer to term(i);and Zi refer to topic(i).

I. Each element in Matrix L represents “for a selected topic Zi, the probability to pick up document Mi”, i.e. P(Di|Zi).
II. Each element in Matrix S represents “the probability to pick topic Zi”, i.e. P(Zi).
III. Each element in Matrix R represents “for a selected topic Zi, the probability to has term Wi in the topic set”, i.e. P(Wi|Zi).

※But there are some assumption has to be made for this probability explanation to be built.

b)

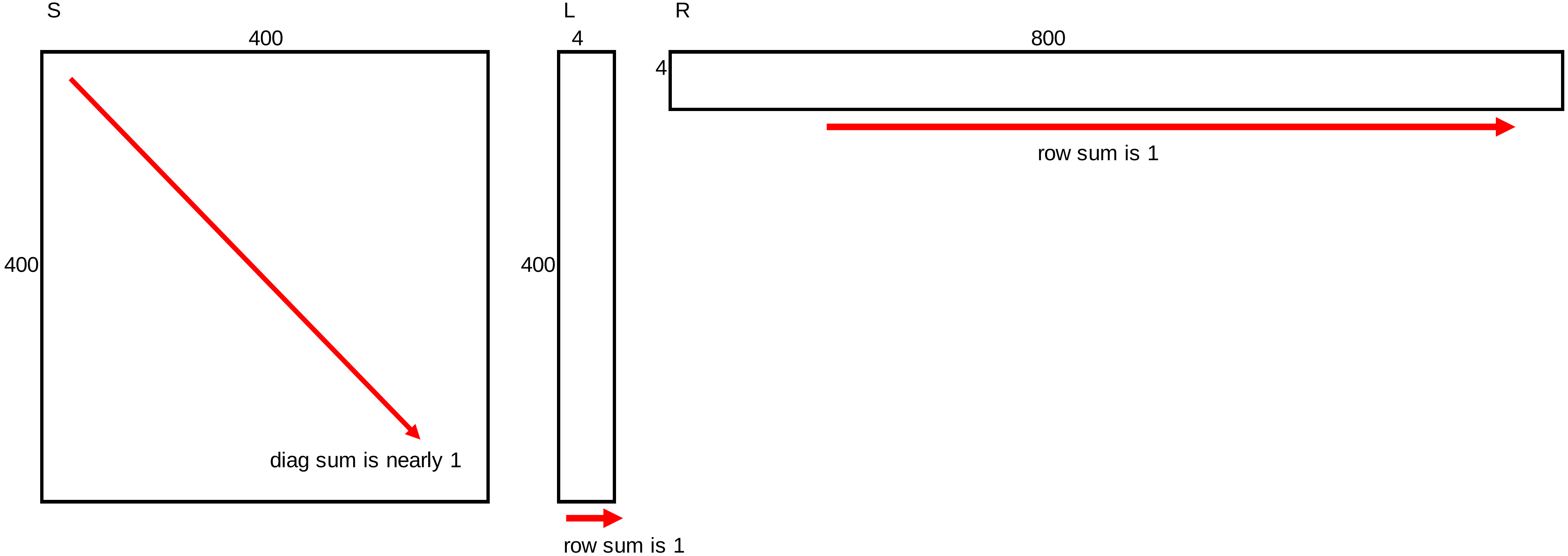
```
slr.gkl <- nmf.slr(lr.gkl)
summary(slr.gkl)
  Length Class     Mode
S 160000 ddiMatrix S4
L   1600 -none-    numeric
R   3200 -none-    numeric
```

```
> 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 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 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 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 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 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 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 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 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 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
```

I acquire some traits of matrix S’, L’, and R’.
I show the traits as below.

D ≅ S %*% L %*% R
dim(D) = 400 X 800



So to enforce a probability concept, I can put some explanation in to the product of these three matrixs.

Let Di refer to document(i); Wi refer to term(i);and Zi refer to topic(i).

I. Each element in Matrix S represents “the probability to pick document Di”, i.e. P(Zi).
II. Each element in Matrix L represents “for a selected document Di, the probability that it contains topic Zi”, i.e. P(Zi|Di).

III. Each element in Matrix R represents “for a selected topic Zi, the probability to has term Wi in the topic set”, i.e. P(Wi|Zi).

※But still, there are some assumption has to be made for this probability explanation to be built.

3 Clustering

Method	Accuracy
a) k-means	0.265
b) k-means on $U_4\Sigma_4$	0.4925
c) k-means on the $\tilde{L}$ matrix of the NMF	0.315
d) k-means on the L' matrix of factorization $L'\Sigma'R'$ obtained from the NMF	0.5025
e) k-means on the L'' matrix of factorization $\Sigma''L''R''$ obtained from the NMF.	0.725

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 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
[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 1 1 1 1 4 1 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 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 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 2 1 1 1 1 1 1 1 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 1 1 1 1 1
[257] 1 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 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 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 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 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 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
[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 4 2 2 2 2 2 2 2 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 2 2 2 2 2
[257] 2 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 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 4 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 3 4 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 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 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
```

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 2 1 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 1 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 2 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 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 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"]
```

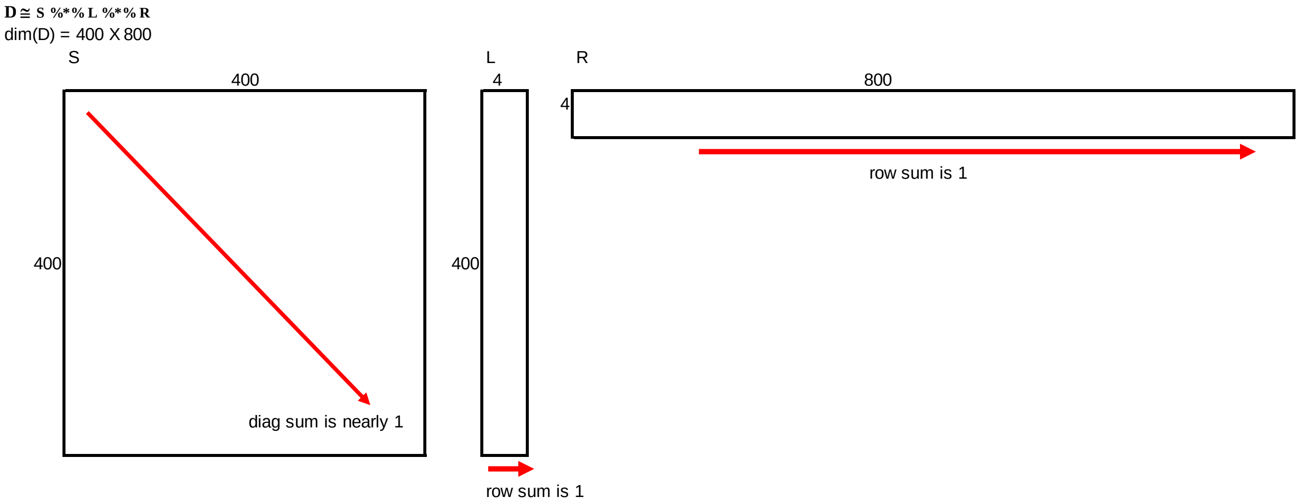
Compare the result:

```
> trail_a
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4
         1 100   0   1   0
         2   0   2   0   0
         3   0   1   0   0
```

```
         4   0  97  99 100

> trail_b
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3  4
         1 97  2  0 11
         2  0  1  0  0
         3  3 97 99 89
         4  0  0  1  0

> trail_c
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4
         1 100  97  99  78
         2   0   3   0   0
         3   0   0   1   0
         4   0   0   0  22

> trail_d
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4
         1  97   0   0   0
         2   0   3   0   0
         3   3   0   1   0
         4   0  97  99 100

> trail_e
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3  4
         1 95 96  1  1
         2  0  1  3  0
         3  2  2 96  1
         4  3  1  0 98
```

Observation 1:

For the accuracy, it is hugely better using method (e). I think it is because of the extraction of the document information from matrix L, which is presented below. In this way, the noisy in the matrix L can be incredibly reduced.



Observation 2:

topic 1=sci.crypt

topic 2=sci.med

topic 3=sci.space

topic 4=soc.religion.christian

For the results showed above, I see most off the error happens with Topic 2 and 3. I think it is because the articles (or documents) over these kind of topics use more general words than of the other two topics, causing the trouble during clustering.