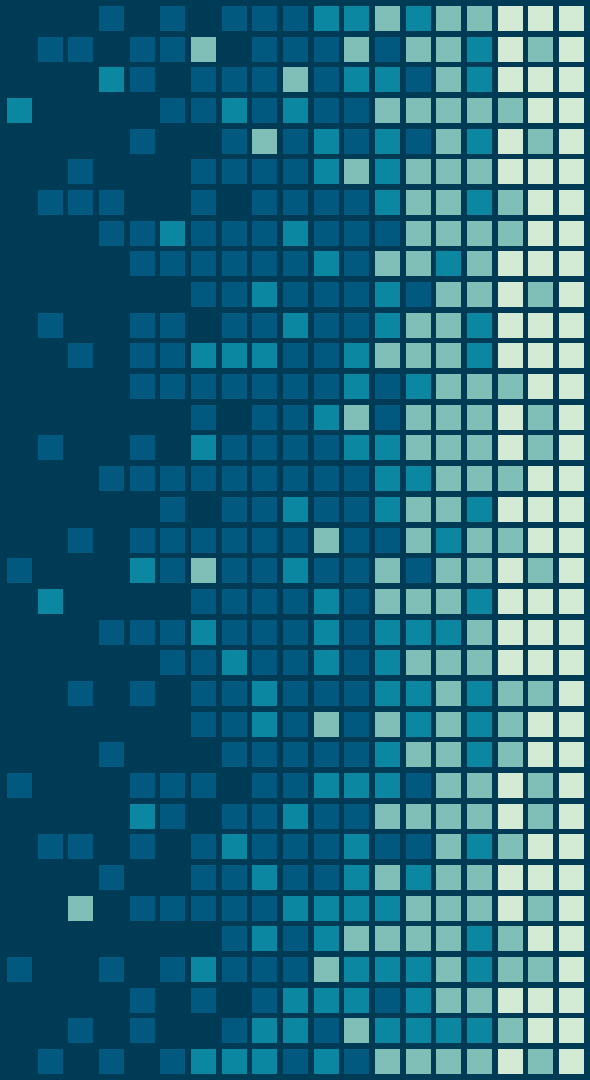
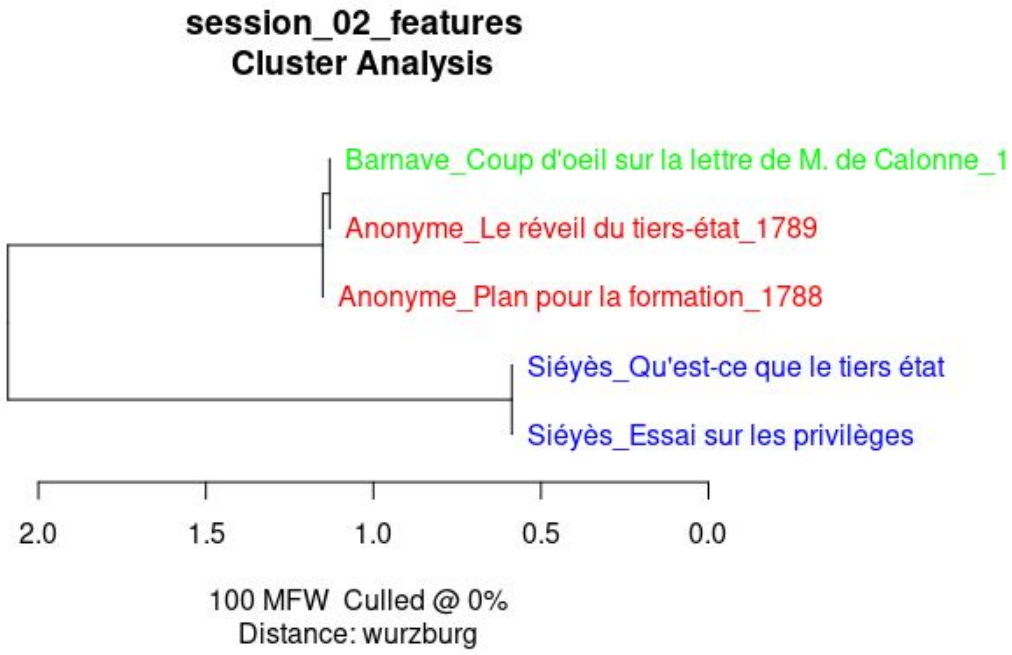
Stylometry with R

Part 3. Words behind trees

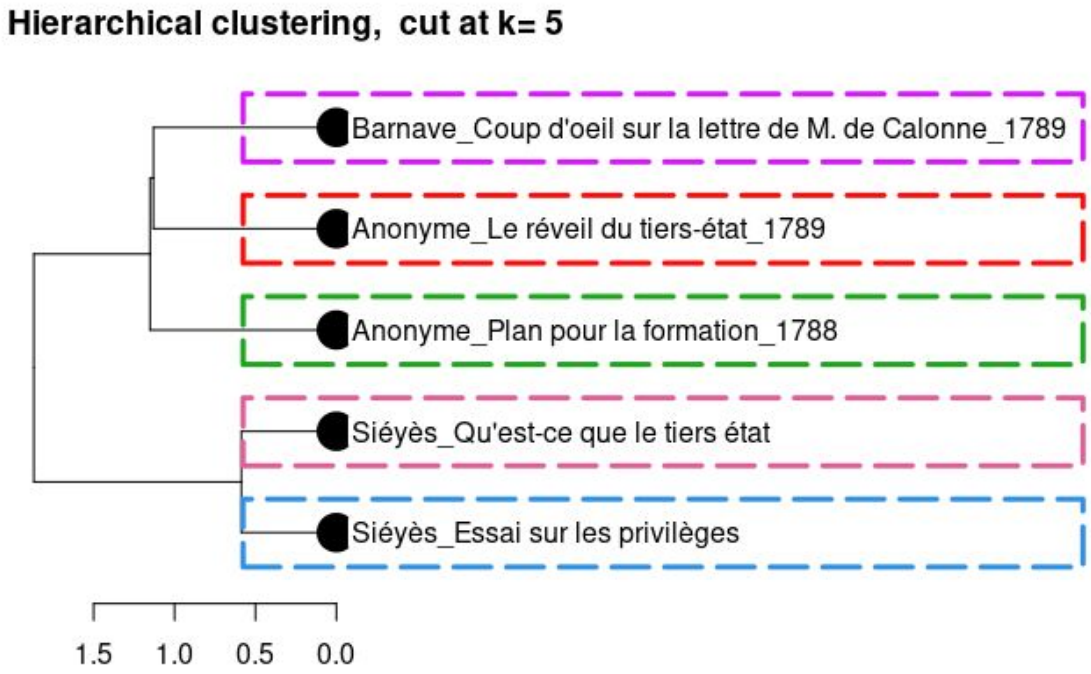
Joanna Byszuk and
Artjoms Šeļa



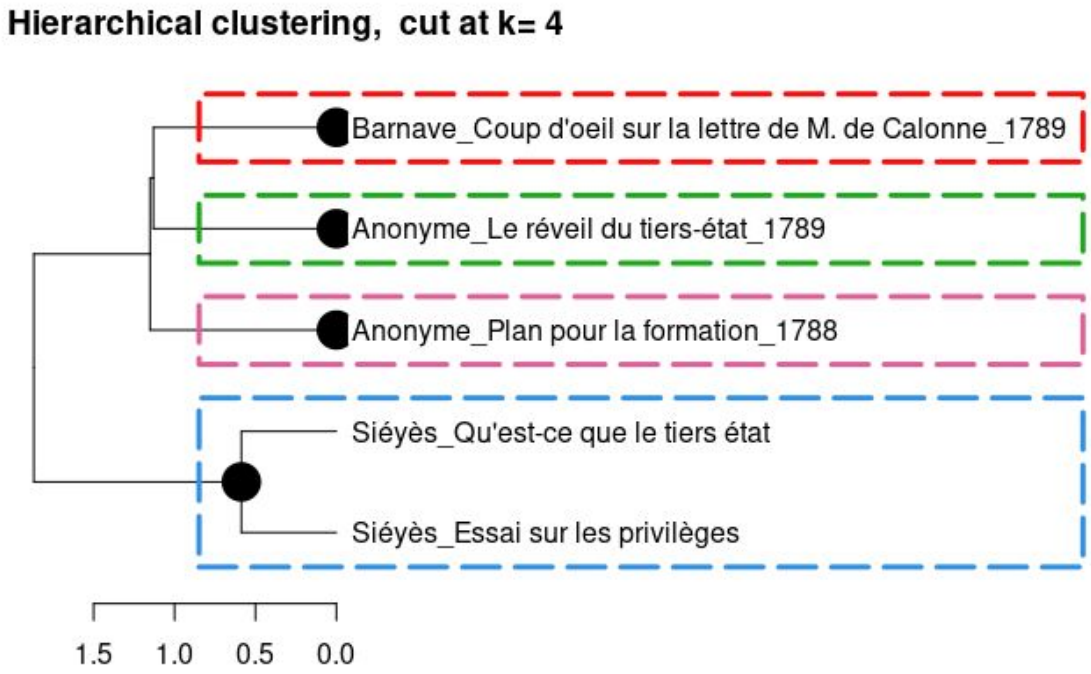
Cluster analysis: grouping suggestion



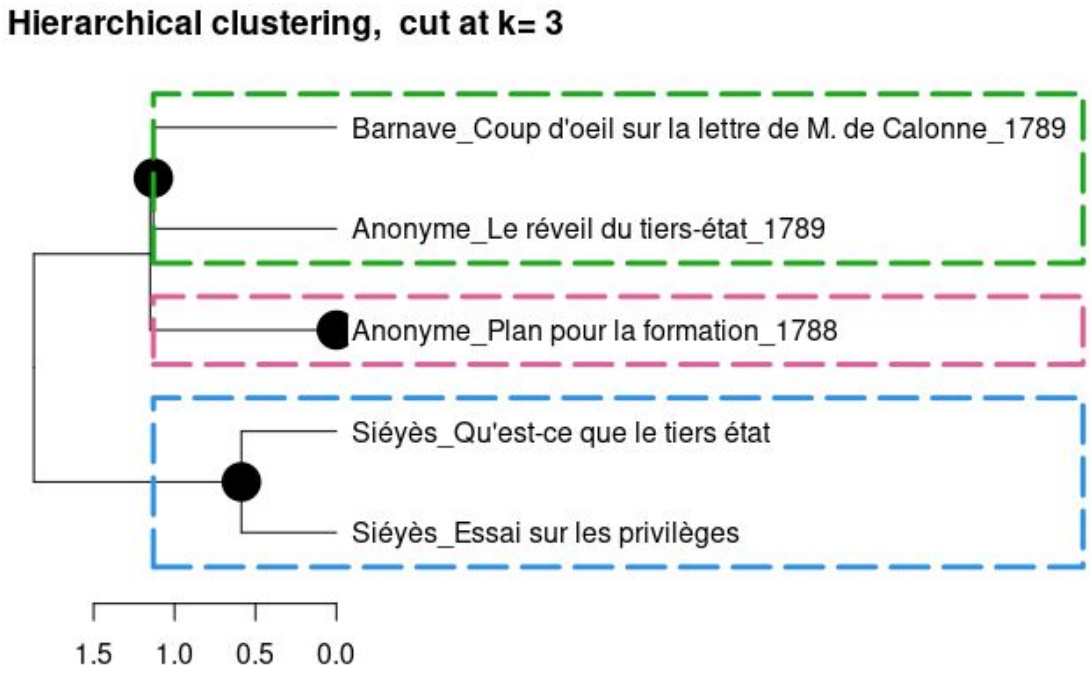
Cutting trees by groups!



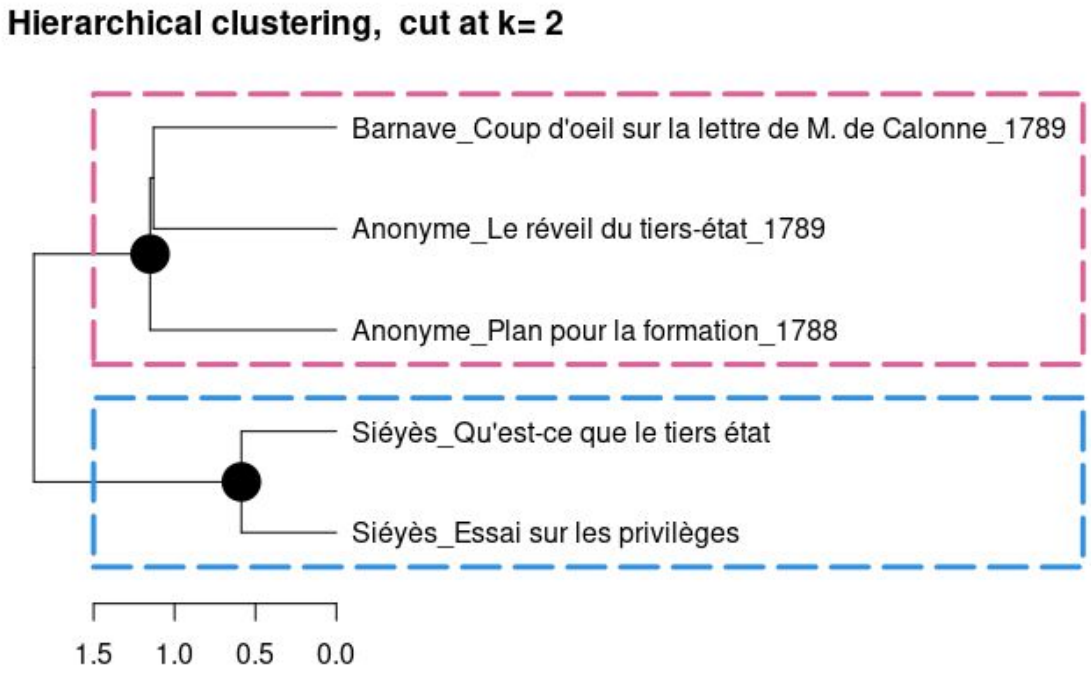
Cutting trees by groups!



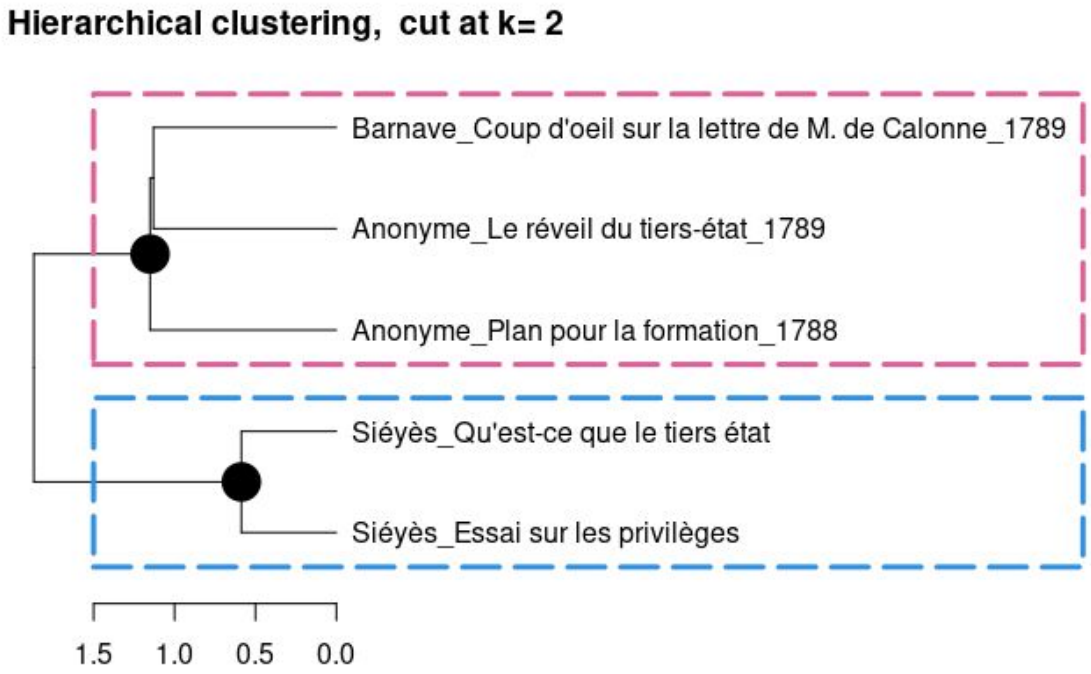
Cutting trees by groups!



Cutting trees by groups!

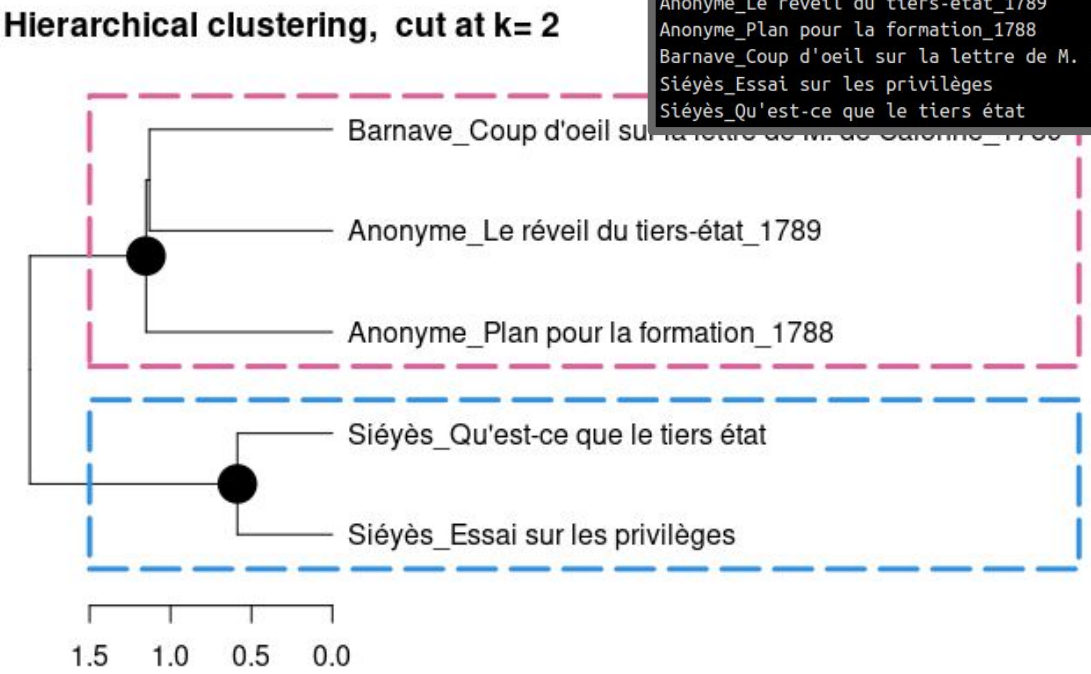


Think of new ‘pink’ and ‘blue’ classes



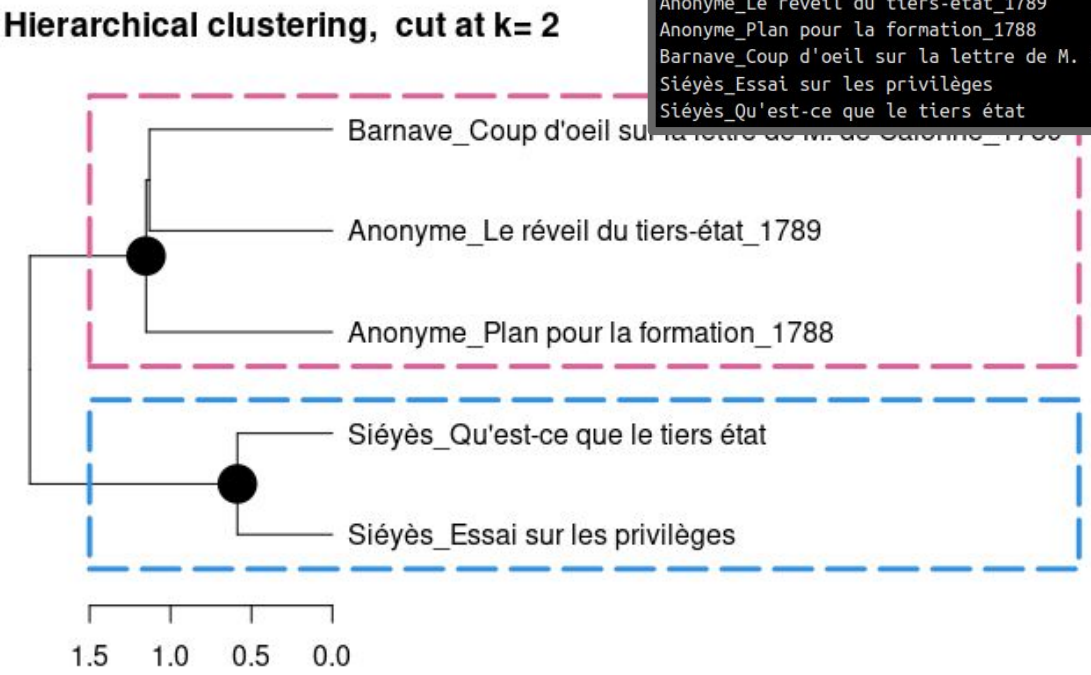
Classes driven by word frequencies!

	de	la	les	l	à	le
Anonyme_Le réveil du tiers-état_1789	4.9919255	3.1490453	2.9068111	2.0993635	1.8333808	1.9663722
Anonyme_Plan pour la formation_1788	4.3847242	3.1117397	4.6676096	1.6030174	1.3672796	1.7444602
Barnave_Coup d'oeil sur la lettre de M. de Calonne_1789	4.2275472	3.0196766	2.8443405	2.2209234	1.1494253	2.1235145
Siéyès_Essai sur les privilèges	4.3275072	2.729227	2.4114403	2.0095336	2.3460136	1.8412936
Siéyès_Qu'est-ce que le tiers état	3.9608393	2.9340656	2.3968003	1.874459	2.0416082	1.8983375



Feature ~ cluster association

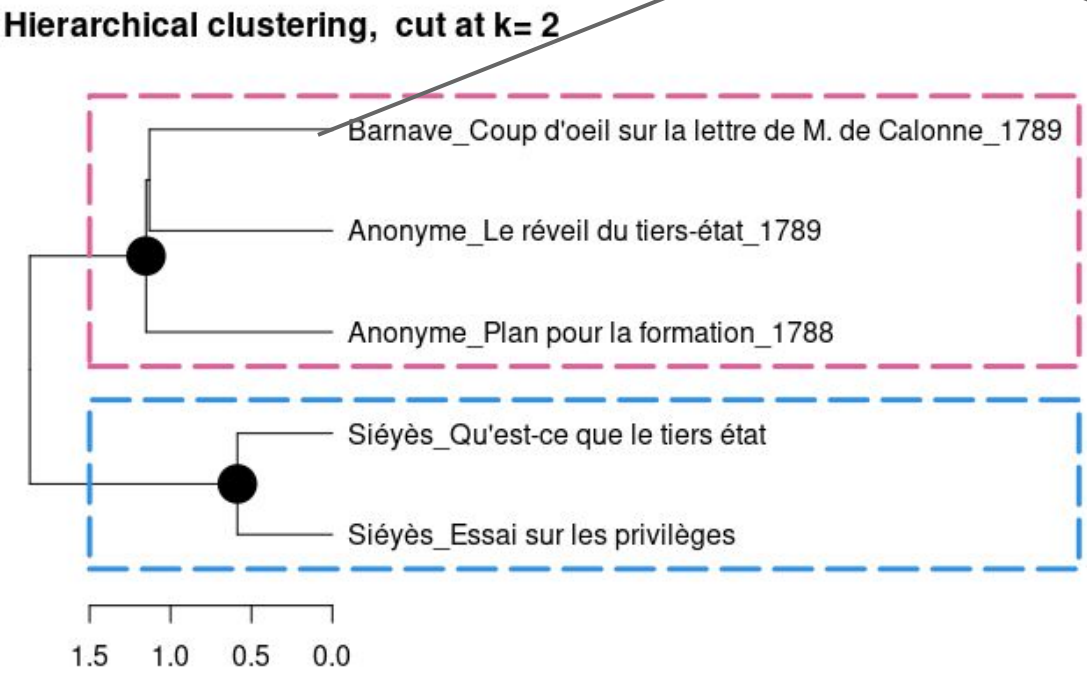
	de	la	les	l	à	le
Anonyme_Le réveil du tiers-état_1789	4.9919255	3.1490453	2.9068111	2.0993635	1.8333808	1.9663722
Anonyme_Plan pour la formation_1788	4.3847242	3.1117397	4.6676096	1.6030174	1.3672796	1.7444602
Barnave_Coup d'oeil sur la lettre de M. de Calonne_1789	4.2275472	3.0196766	2.8443405	2.2209234	1.1494253	2.1235145
Siéyès_Essai sur les privilèges	4.3275072	2.729227	2.4114403	2.0095336	2.3460136	1.8412936
Siéyès_Qu'est-ce que le tiers état	3.9608393	2.9340656	2.3968003	1.874459	2.0416082	1.8983375



You an use 'emerging' class information from the tree to define corpora and check **which features differ** across clusters.

Think about it as **keyword** problem.

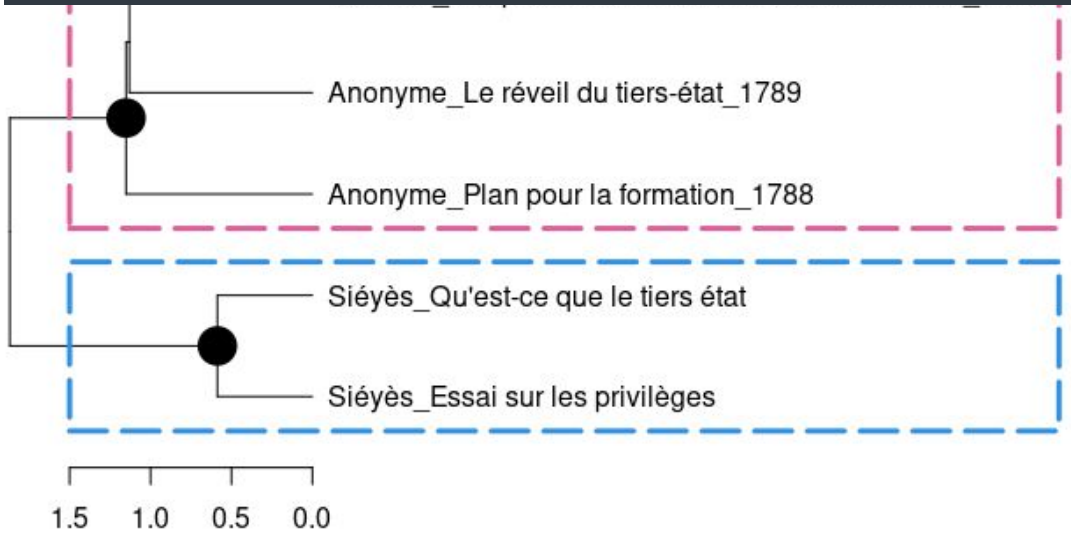
	de	la	les	l	à	le
Anonyme_Le réveil du tiers-état_1789	4.9919255	3.1490453	2.9068111	2.0993635	1.8333808	1.9663722
Anonyme_Plan pour la formation_1788	4.3847242	3.1117397	4.6676096	1.6030174	1.3672796	1.7444602
Barnave_Coup d'oeil sur la lettre de M. de Calonne_1789	4.2275472	3.0196766	2.8443405	2.2209234	1.1494253	2.1235145
Siéyès_Essai sur les privilèges	4.3275072	2.729227	2.4114403	2.0095336	2.3460136	1.8412936
Siéyès_Qu'est-ce que le tiers état	3.9608393	2.9340656	2.3968003	1.874459	2.0416082	1.8983375



Cluster 2	Cluster 1
est	fe
sont	
sans	
sa	
sur	
et	
ses	

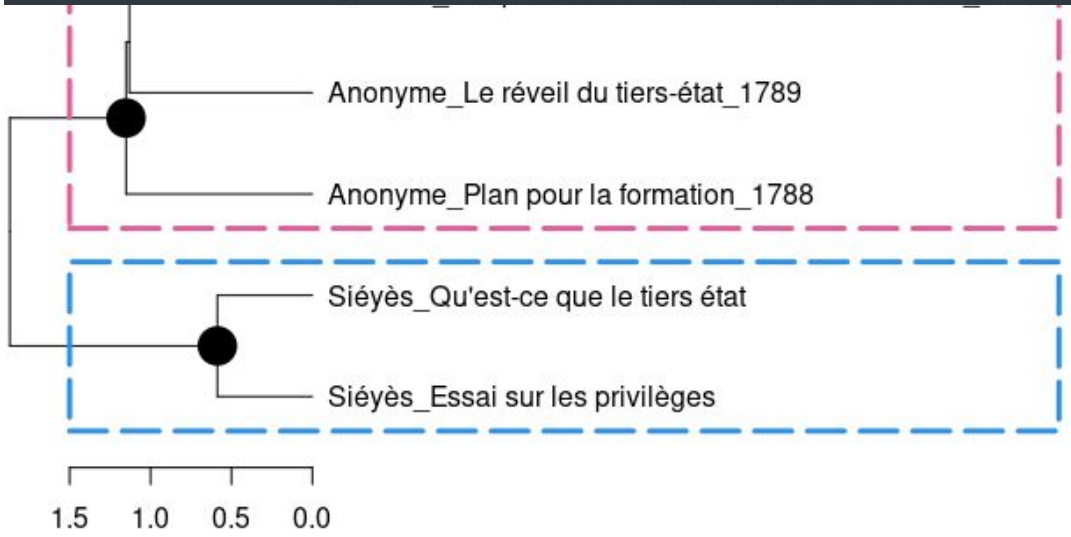
pute, lequel fe 1 Anemblée de TéMion ou ârrondiïiement. Ces Députés ne pourront être élus que parmi les propriétaires domiciliés ou parmi les forains qui auront des propriétés dans le lieu payant cinquante livres de charges réelles & pour être élu, il ne fera pas néceflaire d'être présent à l'Afl'emblée.. Les Etars indiquer 6ht les chefs-lieux des arrondiflemens ailleurs que dans les villes qui Ont des Députés particuliers $ &.pouf la première convocation, les Députes de ï'Éfëétion de Grenoble fe réuniront a Vitille ceux de l'Election de Vienne, à Bouroin,; ceux de l'Elefliôn de Romans à Beaurepaire; ceux de l'Ele&ion de Valence, Chabeuil ceux de l'Eleftiqn de,Gap, ^Charges ceux de l'Eleélion de, Monte- entr*eux les^ Députés qui devront repréfeii* -t|r du içMiûtiB; aux EtatliLé IMrocès-verbai fera envoyé au Secrétaire* le nom des

H



pute, lequel fe 1 Anemblée de TéMion ou ârrondiïiement. Ces Députés ne pourront être élus que parmi les propriétaires domiciliés ou parmi les forains qui auront des propriétés dans le lieu payant cinquante livres de charges réelles & pour être élu, il ne fera pas néceflaire d'être présent à l'Afl'emblée.. Les Etars indiquer 6ht les chefs-lieux des arrondiflemens ailleurs que dans les villes qui Ont des Députés particuliers $ &.pouf la première convocation, les Députes de ï'Éfëétion de Grenoble fe réuniront a Vitille ceux de l'Election de Vienne, à Bouroin,; ceux de l'Elefliôn de Romans à Beaurepaire; ceux de l'Ele&io élion de, Monte- e Lé IMrocès-

OCR: becomes **more** of a problem when text sources are different



An early alpha function from me

```r
library(stylo)
source("src/view_tree.R")

res1 <- stylo(corpus.dir="../session_01_sample
              gui=F,
              distance.measure="wurzburg")


view_tree(res1, ## saved results variable fro
          k=2, ## to how many groups you want
          p=0.05,  ## p-value threshold for
          color_leaves=F ## should it color
          )
```

```
CLUSTER 1
==============
TEXTS
Anonyme_Plan pour la formation_1788
Anonyme_Le réveil du tiers-état_1789
Barnave_Coup d'oeil sur la lettre de M. de Calonne_1789
==============
FEATURES associated (p<0.05)

fe




CLUSTER 2
==============
TEXTS
Siéyès_Essai sur les privilèges
Siéyès_Qu'est-ce que le tiers état
==============
FEATURES associated (p<0.05)

est sont sans sa sur et ses

```

An early alpha function from me

Some tips:

- Bad idea to use when a group includes one ‘leaf’
- Bad idea to use when ***k* > 5**
- For now, does not work with consensus trees
- Untested, might throw horrible errors 🧛👻🧟

Assembling a corpus based on ***k=2*** cut and doing