LDA Topicmodel

Riccardo Ruta

05/2022

Contents

PART I - CREATE THE DTM

1) Convert the Document Feature Matrix (Dfm) in a Topic Model (Dtm)

```
dtm <- quanteda::convert(DFM_trimmed, to = "topicmodels")
```

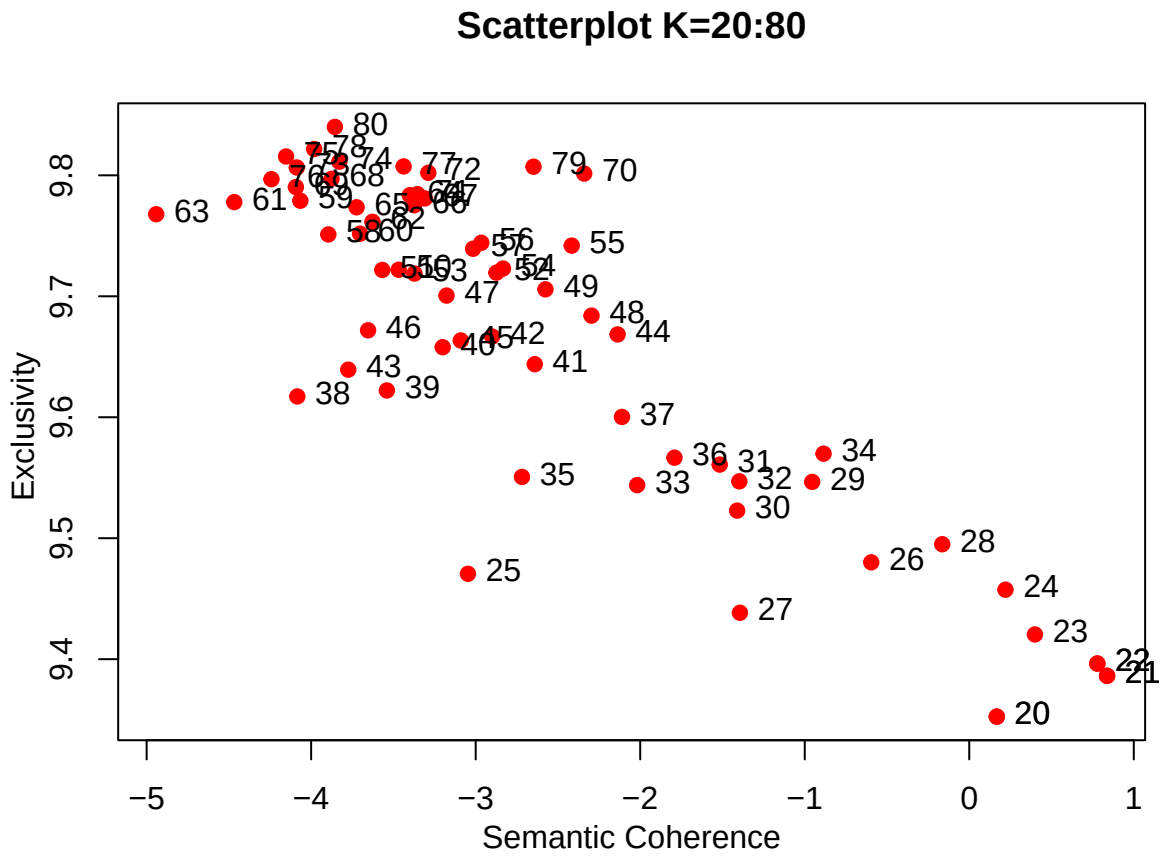
PART II - FIND THE BEST NUMBER OF TOPICS K

1) First try K = 20:80

```
## Finding the best K
top1 <- c(20:80)

## let's create an empty data frame
results1 <- data.frame(first=vector(), second=vector(), third=vector())


system.time(
  for (i  in top1)
  {
    set.seed(123)
    lda1 <- LDA(dtm, method= "Gibbs", k = (i),  control=list(verbose=50L, iter=100))
    topic <- (i)
    coherence <- mean(topic_coherence(lda1, dtm))
    exclusivity <- mean(topic_exclusivity(lda1))
    results1 <- rbind(results1 , cbind(topic, coherence, exclusivity ))
  }
)
# save(results1,file="data/results1.Rda")
```

topic	coherence	exclusivity
20	0.1674464	9.352493
21	0.8376643	9.386225
22	0.7783789	9.396400
23	0.3991629	9.420433
24	0.2202299	9.457425
25	-3.0470840	9.470567

	topic	coherence	exclusivity
59	78	-3.9815896	9.821740
60	79	-2.6484726	9.807149
61	80	-3.8557027	9.840021
62	20	0.1674464	9.352493
63	21	0.8376643	9.386225
64	22	0.7783789	9.396400

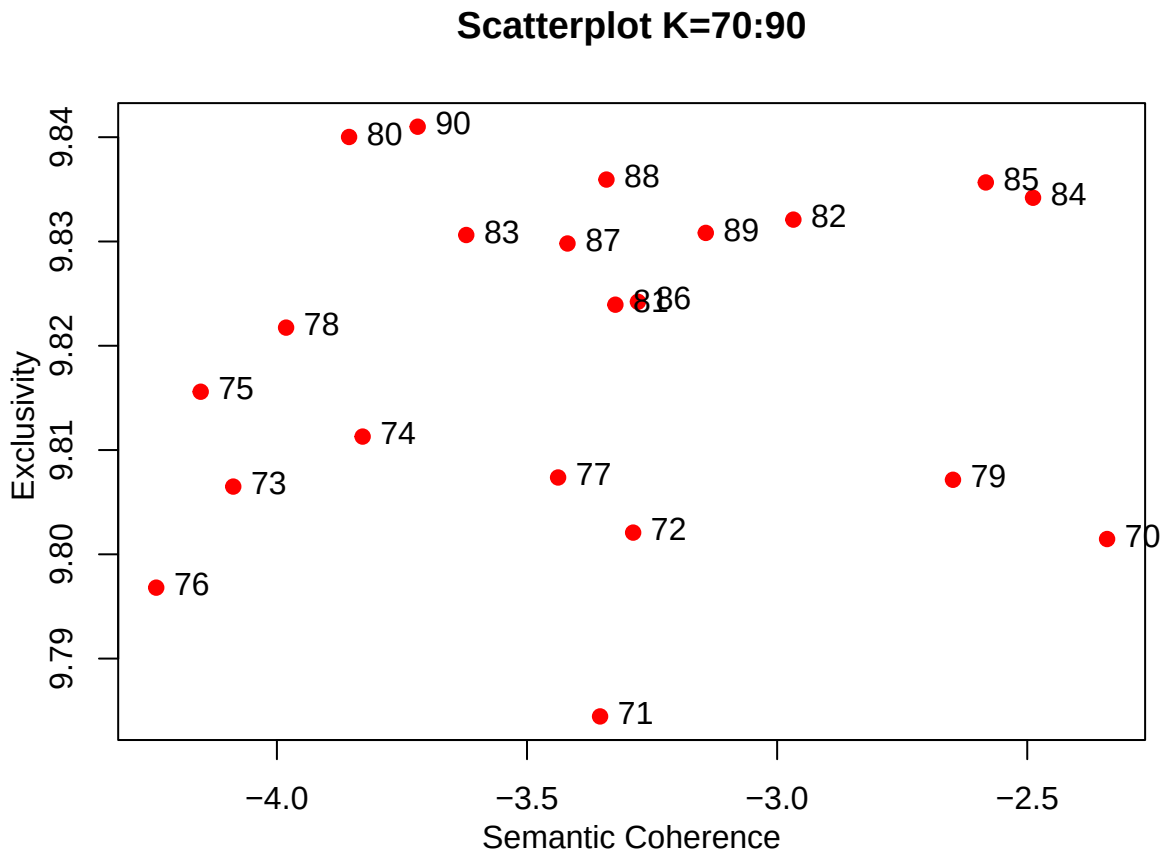


From this first try the best k seems to be 34

2) Second try K = 70:90

```
top2 <- c(70:90)
top2

results2 <- data.frame(first=vector(), second=vector(), third=vector())
results2
```

```
system.time(
  for (i  in top2)
  {
    set.seed(123)
    lda2 <- LDA(dtm, method= "Gibbs", k = (i),  control=list(verbose=50L, iter=100))
    topic <- (i)
    coherence <- mean(topic_coherence(lda2, dtm))
    exclusivity <- mean(topic_exclusivity(lda2))
    results2 <- rbind(results2 , cbind(topic, coherence, exclusivity ))
  }
)

# save(results2,file="data/results2.Rda")
```

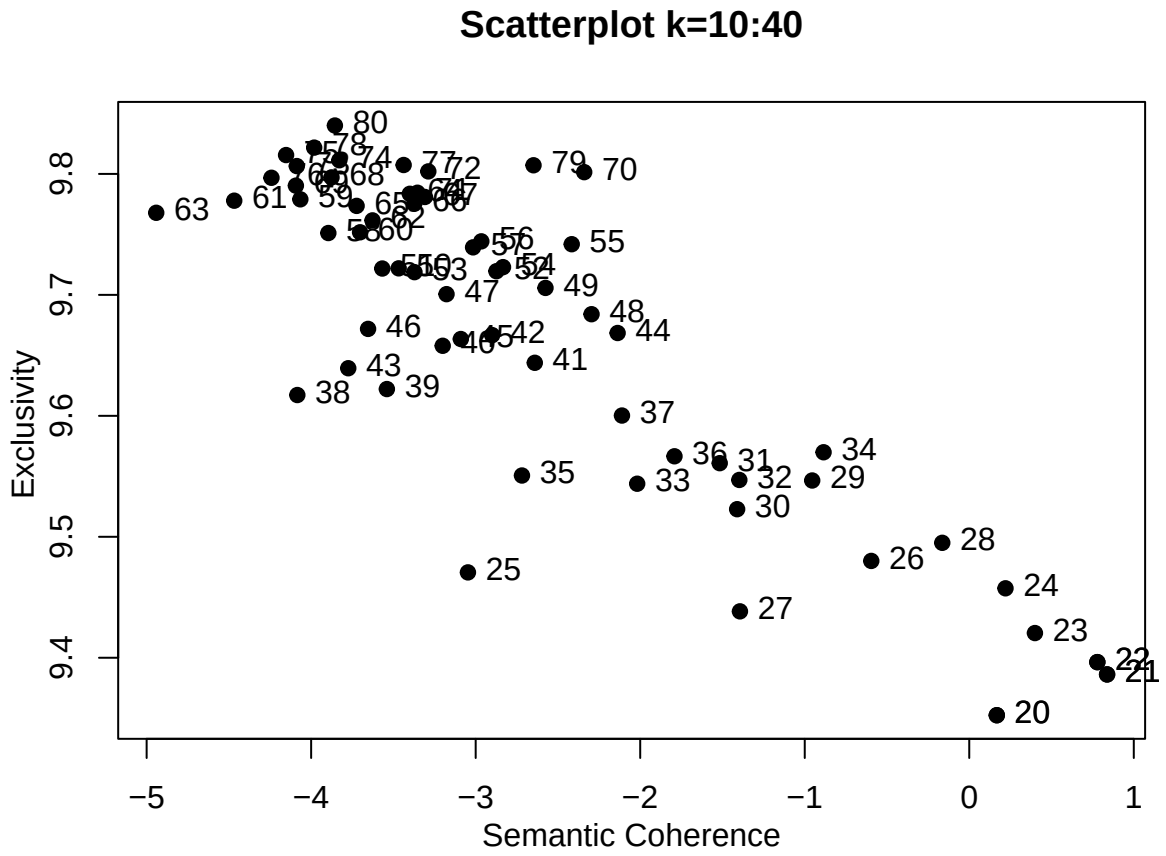


In this case 84 seems better, but the plot is very dispersive

3) Third try K = 10:40 with iteration = 1000

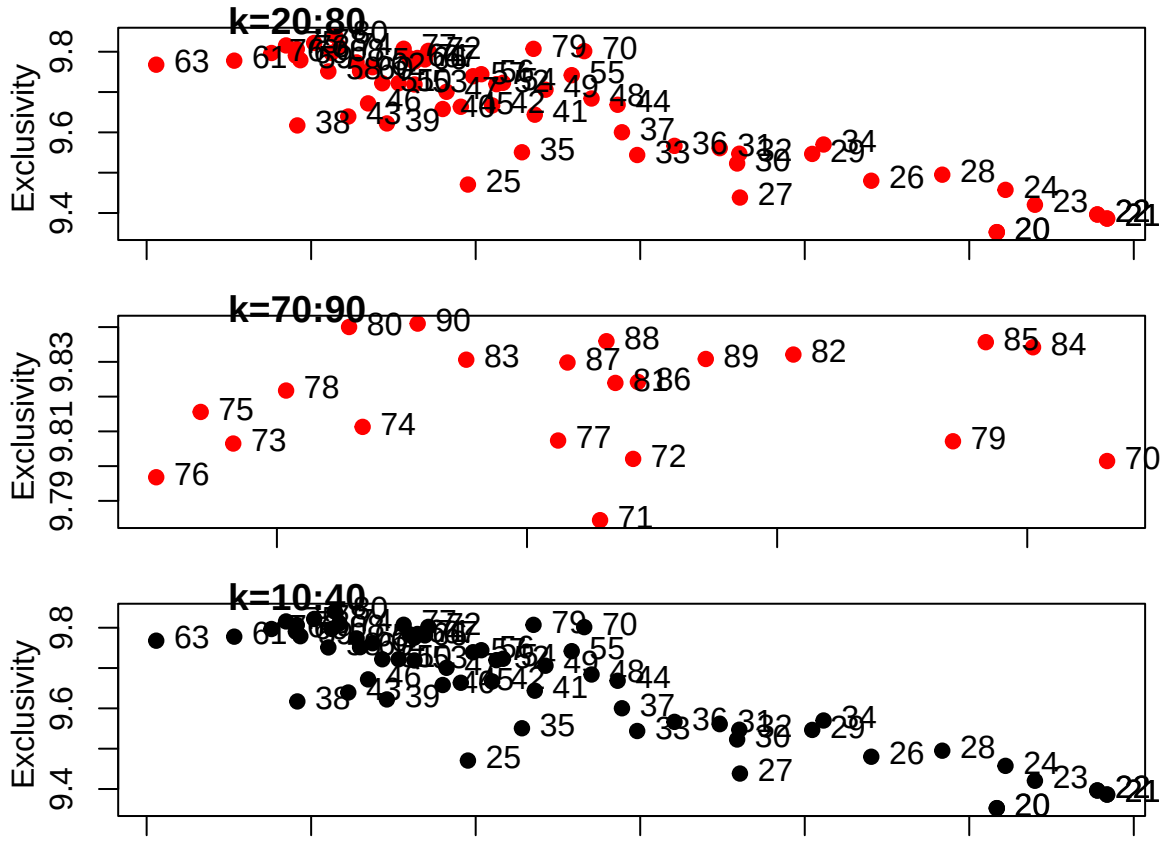
```
## Finding the best K
top_k <- c(10:40)
## let's create an empty data frame
results1 <- data.frame(first=vector(), second=vector(), third=vector())
system.time(
```

```
  for (i  in top_k)
  {
    set.seed(123)
    lda1 <- LDA(dtm, method= "Gibbs", k = (i),  control=list(verbose=50L, iter=1000))
    topic <- (i)
    coherence <- mean(topic_coherence(lda1, dtm))
    exclusivity <- mean(topic_exclusivity(lda1))
    results1 <- rbind(results1 , cbind(topic, coherence, exclusivity ))
  }
)
 #save(results1,file="data/results_k_10-40.Rda")
```



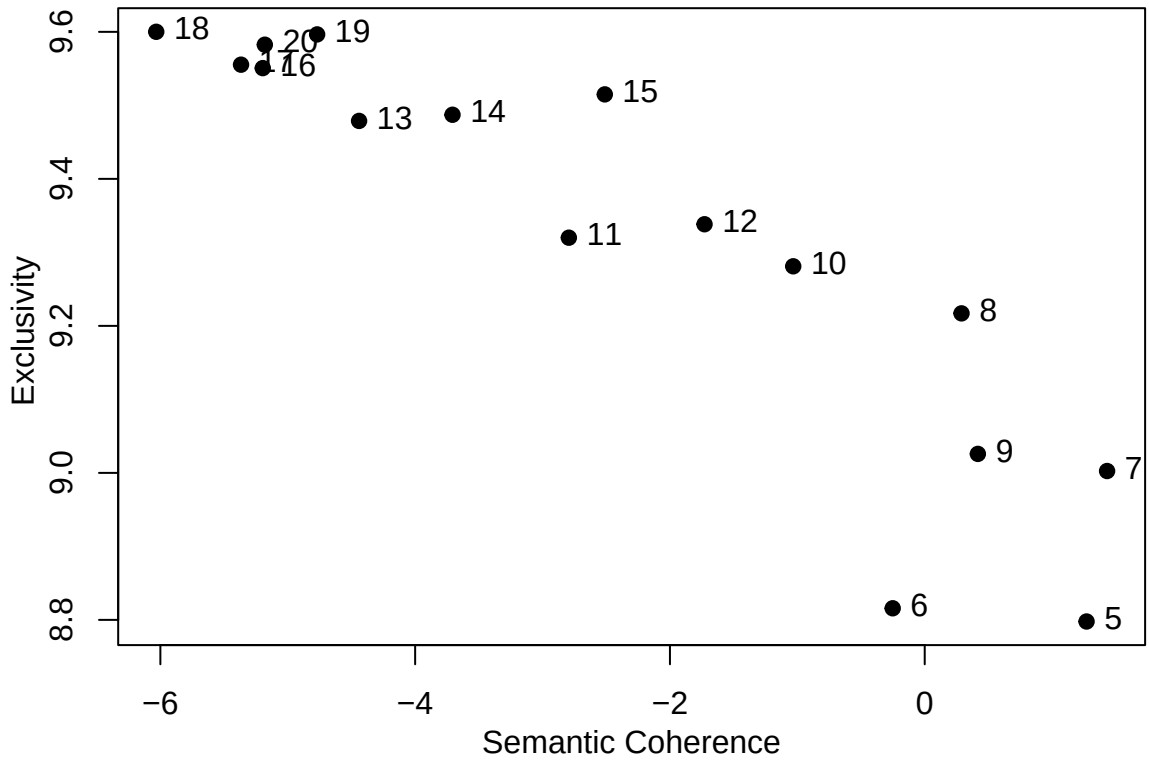
In this iteration the best choice can be 28 but the general level of coherence has fallen very low

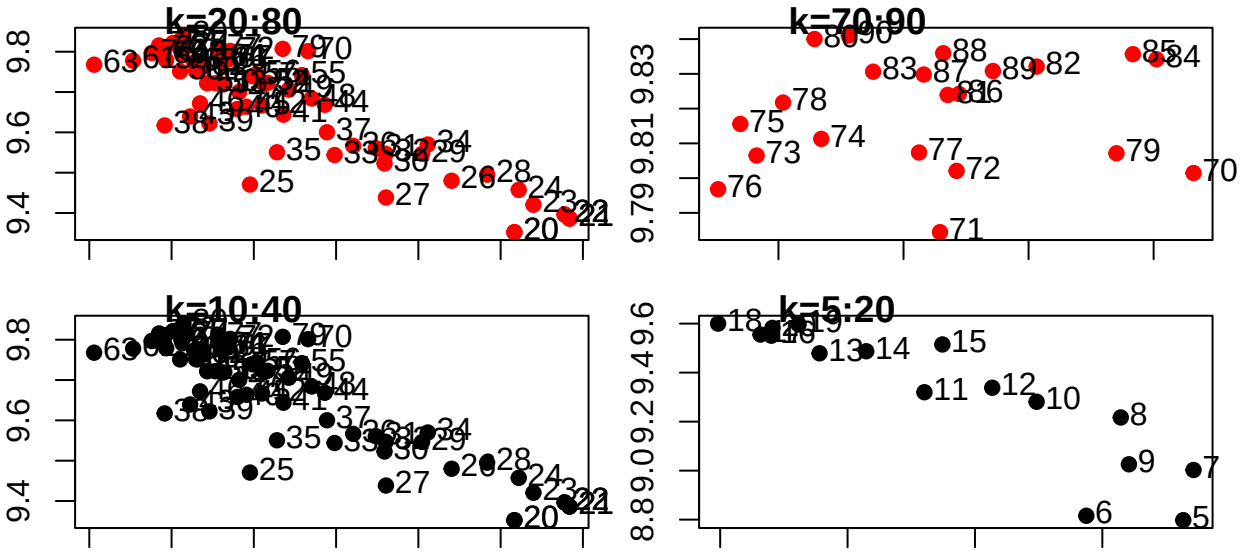
```
grid <- plot_grid(plot1, plot2, plot3,
                  labels = list("k=20:80", "k=70:90", "k=10:40"), label_x = .15, nrow = 3)

# ggsave("figs/plot_grid.png", plot = grid)
grid
```

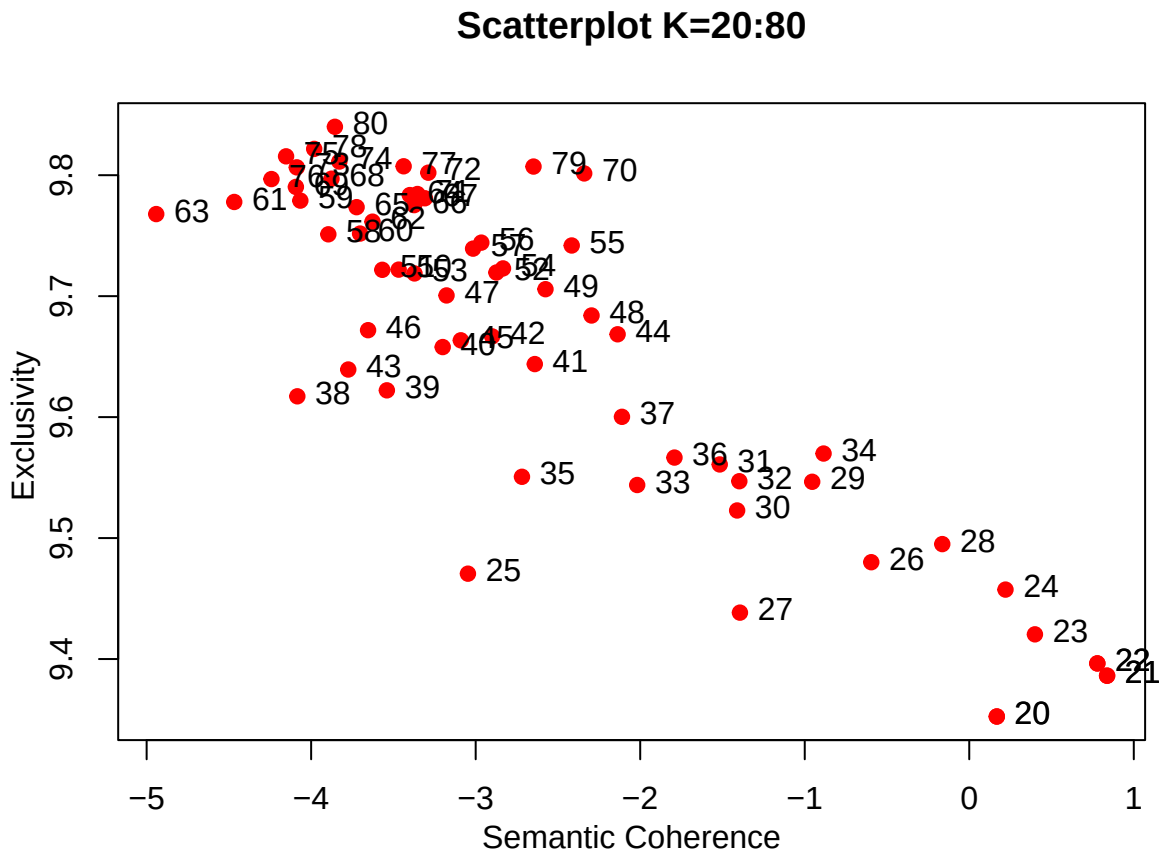


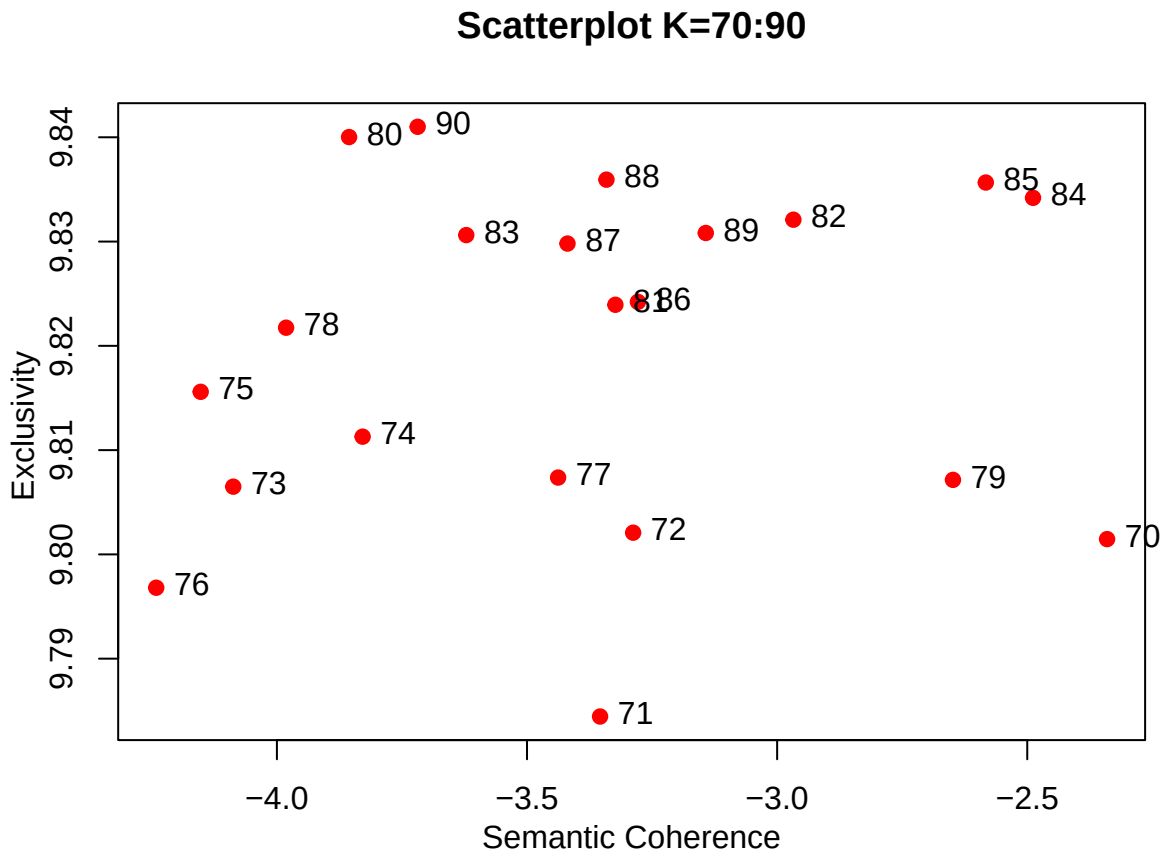
4) Fourth try k = 5:20 iteration n = 2000

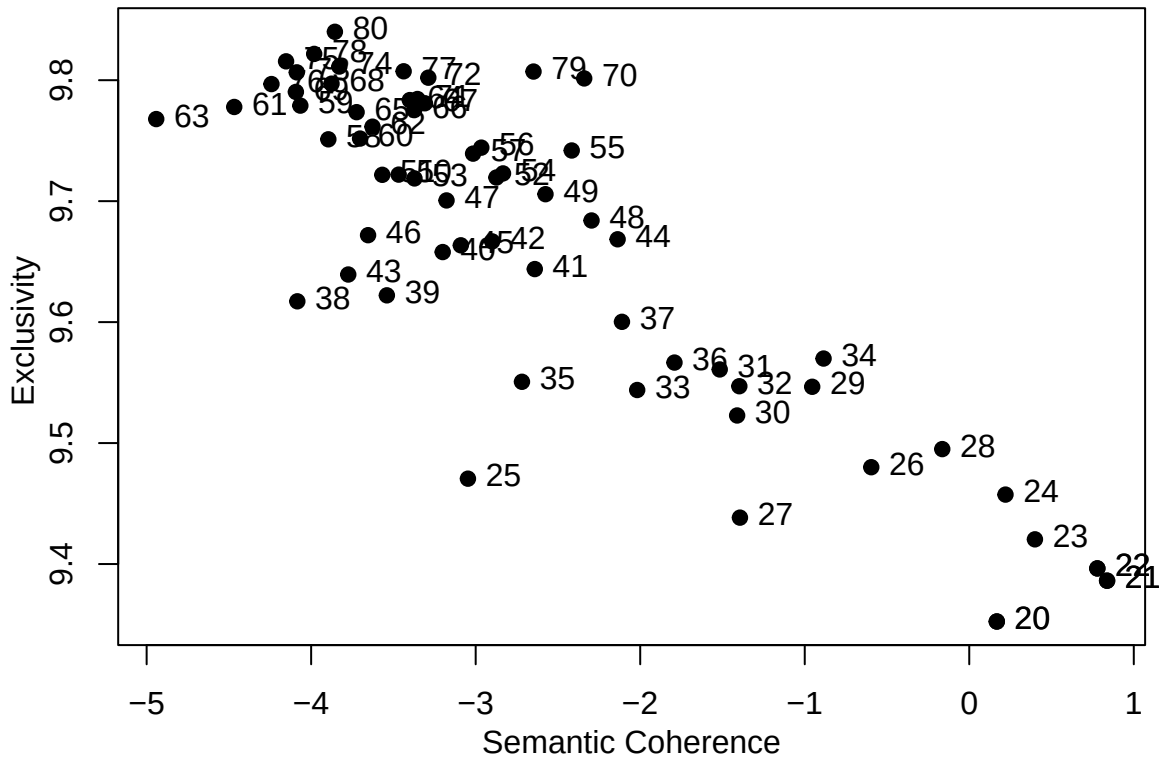
```
## Finding the best K
top1234 <- c(5:20)
top1234

## let's create an empty data frame
results1 <- data.frame(first=vector(), second=vector(), third=vector())
results1

system.time(
  for (i  in top1234)
  {
    set.seed(123)
    lda1 <- LDA(dtm, method= "Gibbs", k = (i),  control=list(verbose=50L, iter=2000))
    topic <- (i)
    coherence <- mean(topic_coherence(lda1, dtm))
    exclusivity <- mean(topic_exclusivity(lda1))
    results1 <- rbind(results1 , cbind(topic, coherence, exclusivity ))
  }
)
# save(results1,file="data/k_5-20.Rda")
```

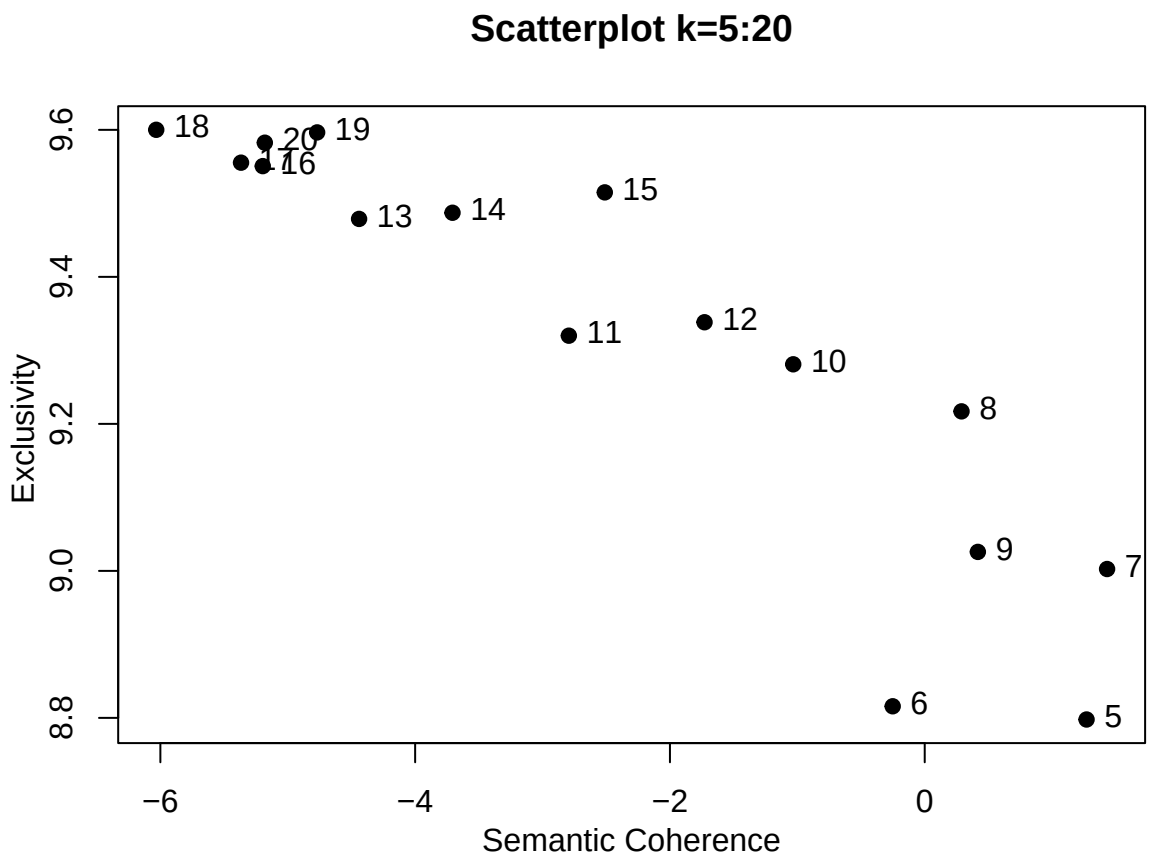

Scatterplot k=5:20
Exclusivity
Semantic Coherence
9.6
9.4
9.2
9.0
8.8
−6
−4
−2
0
18
19
20
17
16
13
14
15
11
12
10
8
9
7
6
5


k=20:80
k=70:90
k=10:40
k=5:20


Scatterplot K=20:80
Exclusivity
Semantic Coherence
9.4
9.5
9.6
9.7
9.8
−5
−4
−3
−2
−1
0
1


Scatterplot K=70:90
Exclusivity
9.84
9.83
9.82
9.81
9.80
9.79
80
90
88
85
84
83
87
89
82
86
78
75
74
73
77
79
72
70
76
71
−4.0
−3.5
−3.0
−2.5
Semantic Coherence


Scatterplot k=10:40
Exclusivity
Semantic Coherence
9.4
9.5
9.6
9.7
9.8
−5
−4
−3
−2
−1
0
1
80
79
70
63
61
55
49
48
47
46
44
43
41
39
38
37
36
35
34
33
32
31
30
29
28
27
26
25
24
23
22
21
20



After all these tests, I believe that 10 is the best choice because it achieves good levels of coherence and exclusivity and is consistent with the choice of looking for substantive policy themes

PART III - ANALISYS OF THE TOPICS

First try with k = 30

```
system.time(lda <- LDA(dtm, method= "Gibbs", k = 30, control = list(seed = 123)))
# save(lda, file = "data/lda_k_30.Rda")
```

Here i extract the most important terms from the model

Top terms 01	Top terms 02	Top terms 03	Top terms 04	Top terms 05	Top terms 06	Top terms 07	Top terms 08	Top terms 09	Top terms 10
#afghanistan	president	forza	governo		#iostoconsalvini	#draghi	maggio	the	guerra
#tokyo2020	#quirinal	grand	italiani	donn	governo	draghi	governo	@fratellilditalia	#ucraina
talebani	repubblica	buon	+	via	luglio	governo	#decretorilancio	of	pace
	#presidentedellarepubblica	pd	lavoro	violenza	#recoveryfund	#governodraghi	#fase2	to	ucraina
agosto	#quirinale2022	l'italia	#covid19	giornata	president	lavoro	lavoro	and	putin
pass	draghi	politica	crisi		#cont	paes	ministro	violenza	donn
	gennaio		impres	minacc	#salvini	president	#bonafed	donn	marzo
@stampasgarbi	quirinal	c'è	diretta		legg	@fratellilditalia	#silviaromano	novembr	
grazi	grand	casa	momento	mondo	cont	buon	#recoveryfund	covid	russia
afghanistan	#mattarella	@pdnetwork	l'italia	pensiero	l'italia	l'italia	ripartir	#morradimett	ucraino

Top terms 11	Top terms 12	Top terms 13	Top terms 14	Top terms 15	Top terms 16	Top terms 17	Top terms 18	Top terms 19	Top terms 20
#dpcm	pass	maggio	pass	giugno	giugno	governo	governo	#coronavirus	lavoro
governo	sindaco	april	draghi	#2giugno	#primalitalia	ministro	cont	#mes	paes
#iostoconsalvini	green	vaccinal	natal	scuola	roma	paes	#crisidigoverno	#covid19	italia
ottobr	#greenpass	coprifuoco	green	minacc	bocca	c'è	crisi	#cont	donn
#cont	settembr	@stampasgarbi	vaccinati	+	luglio	cittadini	#cont	april	giornata
#mes	candidato	@fratelliditalia	dicembr	governo	piazza	president	paes	#forzalombardia	commission
covid	città	#nocoprifuoco		#cont	@stampasgarbi	bene	maggioranza	mes	l'italia
@fratelliditalia	draghi	#pnrr	@fratelliditalia	paes	forza		president	liquidità	italiani
de	piazza	pandemia	covid	@luigidimaio		grazi	#governo	ripartir	#lega
jole	roma	draghi	@fattoquotidiano	cont	draghi	parlamento	@fratelliditalia	#fase2	insiem

Top terms 21	Top terms 22	Top terms 23	Top terms 24	Top terms 25	Top terms 26	Top terms 27	Top terms 28	Top terms 29	Top terms 30
settembr	governo	#sanremo2022	#coronavirus	grazi	natal	vaccini	governo	pass	#referendumgiustizia
#iovotono	#iostoconsalvini	febbraio	grazi	anni	cont	donn	agosto	green	giustizia
elettoral	#oggivotolega	draghi	misur	lavoro	governo	buon	anni	ottobr	pass
#processateanchem	#salvini	green	l'emergenza	grand	bilancio	@pdnetwork	vittim	#ddlzan	luglio
#referendum	salvini	pass	momento	diritti	@fratelliditalia	marzo	settembr	roma	#greenpass
parlamentari	#borgonzonipresident	#ucraina	emergenza	via	dicembr	@fratelliditalia	#iovotono	sindaco	gazebo
voto	#prescrizion	parlamento	coronavirus	legg	#mes	auguri	@fratelliditalia	legg	#ddlzan
@fratelliditalia	@fratelliditaiia	@fratelliditalia	casa	president	#natal	draghi	covid		riforma
scuola	#emiliaromagna	guerra	decreto	giovani	italiani	vaccinal	bonus	+	firm
referendum	#m5s	#greenpass	#iorestoacasa	città	mes	vaccino	scuola	@forza_italia	draghi

COMMENTHERE

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
titles	1	2	3	4	5	6	7	8	9	10
	#afghanistan	president	forza	governo		#iostoconsalvini	#draghi	maggio	the	guerra
	#tokyo2020	#quirinal	grand	italiani	donn	governo	draghi	governo	@fratelliditalia	#ucraina
	talebani	repubblica	buon	+	via	luglio	governo	#decretorilancio	of	pace
		#presidentedellarepubblica	pd	lavoro	violenza	#recoveryfund	#governodraghi	#fase2	to	ucraina
	agosto	#quirinale2022	l'italia	#covid19	giornata	president	lavoro	lavoro	and	putin
	pass	draghi	politica	crisi		#cont	paes	ministro	violenza	donn
		gennaio		impres	minacc	#salvini	president	#bonafed	donn	marzo
	@stampasgarbi	quirinal	c'è	diretta		legg	@fratelliditalia	#silviaromano	novembr	
	grazi	grand	casa	momento	mondo	cont	buon	#recoveryfund	covid	russia
	afghanistan	#mattarella	@pdnetwork	l'italia	pensiero	l'italia	l'italia	ripartir	#morradimett	ucraino

	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
titles	11	12	13	14	15	16	17	18	19	20
	#dpcm	pass	maggio	pass	giugno	giugno	governo	governo	#coronavirus	lavoro
	governo	sindaco	april	draghi	#2giugno	#primalitalia	ministro	cont	#mes	paes
	#iostoconsalvini	green	vaccinal	natal	scuola	roma	paes	#crisidigoverno	#covid19	italia
	ottobr	#greenpass	coprifuoco	green	minacc	bocca	c'è	crisi	#cont	donn
	#cont	settembr	@stampasgarbi	vaccinati	+	luglio	cittadini	#cont	april	giornata
	#mes	candidato	@fratelliditalia	dicembr	governo	piazza	president	paes	#forzalombardia	commission
	covid	città	#nocoprifuoco		#cont	@stampasgarbi	bene	maggioranza	mes	l'italia
	@fratelliditalia	draghi	#pnrr	@fratelliditalia	paes	forza		president	liquidità	italiani
	de	piazza	pandemia	covid	@luigidimaio		grazi	#governo	ripartir	#lega
	jole	roma	draghi	@fattoquotidiano	cont	draghi	parlamento	@fratelliditalia	#fase2	insiem

	Topic 21	Topic 22	Topic 23	Topic 24	Topic 25	Topic 26	Topic 27	Topic 28	Topic 29	Topic 30
titles	21	22	23	24	25	26	27	28	29	30
	settembr	governo	#sanremo2022	#coronavirus	grazi	natal	vaccini	governo	pass	#referendumgiustizia
	#iovotono	#iostoconsalvini	febbraio	grazi	anni	cont	donn	agosto	green	giustizia
	elettoral	#oggivotolega	draghi	misur	lavoro	governo	buon	anni	ottobr	pass
	#processateanchem	#salvini	green	l'emergenza	grand	bilancio	@pdnetwork	vittim	#ddlzan	luglio
	#referendum	salvini	pass	momento	diritti	@fratelliditalia	marzo	settembr	roma	#greenpass
	parlamentari	#borgonzonipresident	#ucraina	emergenza	via	dicembr	@fratelliditalia	#iovotono	sindaco	gazebo
	voto	#prescrizion	parlamento	coronavirus	legg	#mes	auguri	@fratelliditalia	legg	#ddlzan
	@fratelliditalia	@fratelliditaiia	@fratelliditalia	casa	president	#natal	draghi	covid		riforma
	scuola	#emiliaromagna	guerra	decreto	giovani	italiani	vaccinal	bonus	+	firm
	referendum	#m5s	#greenpass	#iorestoacasa	città	mes	vaccino	scuola	@forza_italia	draghi

COMMENTHERE

Repeat the search using a much lower K.

K = 10

```
system.time(lda_k_10 <- LDA(dtm, method= "Gibbs", k = 10, control = list(seed = 123)))
# save(lda_k_10, file = "data/lda_k_30.Rda")
```

Top terms 01	Top terms 02	Top terms 03	Top terms 04	Top terms 05	Top terms 06	Top terms 07	Top terms 08	Top terms 09	Top terms 10
guerra	grazie	#coronavirus	governo	draghi	governo	green	giustizia	#iostoconsalvini	governo
presidente	italia	#covid19	paese	#draghi	conte	via	grazie	salvini	italiani
#ucraina	legge	governo	presidente	lavoro	@fratellilditalia	draghi	#referendumgiustizia	governo	#covid19
draghi	donne	momento	lavoro	@fratellilditalia	the	@fattoquotidiano	#tokyo2020	elettorale	l'italia
ucraina	grande	misure	politica	vaccini	#conte	#greenpass	#greenpass	#salvini	@fratellilditalia
#quirinale	insieme	emergenza	italiani	pandemia	italiani	vaccinati	luglio	settembre	imprese
@fratellilditalia	via	grazie	l'italia	governo	#covid19	@fratellilditalia	@stampasgarbi	#iovotono	lavoro
putin	commissione	imprese	grande	vaccinale	#governo	sindaco	gazebo	#lega	#recoveryfund
marzo	libertà	decreto	grazie	#covid19	covid	ottobre	riforma	@matteosalvinimi	conte
russia	mondo	l'emergenza	parlamento	paese	pandemia	lavoro	green	@pdnetwork	#conte

COMMENTHERE

Titles:

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
titles	1	2	CORONAVIRUS	SOSTEGNI ECONOMICI	LEGA	DIRITTI	GOVERNO	INFORMAZIONE GIORNALISTICA	9	10
	guerra	grazie	#coronavirus	governo	draghi	governo	green	giustizia	#iostoconsalvini	governo
	presidente	italia	#covid19	paese	#draghi	conte	via	grazie	salvini	italiani
	#ucraina	legge	governo	presidente	lavoro	@fratellilditalia	draghi	#referendumgiustizia	governo	#covid19
	draghi	donne	momento	lavoro	@fratellilditalia	the	@fattoquotidiano	#tokyo2020	elettorale	l'italia
	ucraina	grande	misure	politica	vaccini	#conte	#greenpass	#greenpass	#salvini	@fratellilditalia
	#quirinale	insieme	emergenza	italiani	pandemia	italiani	vaccinati	luglio	settembre	imprese
	@fratellilditalia	via	grazie	l'italia	governo	#covid19	@fratellilditalia	@stampasgarbi	#iovotono	lavoro
	putin	commissione	imprese	grande	vaccinale	#governo	sindaco	gazebo	#lega	#recoveryfund
	marzo	libertà	decreto	grazie	#covid19	covid	ottobre	riforma	@matteosalvinimi	conte
	russia	mondo	l'emergenza	parlamento	paese	pandemia	lavoro	green	@pdnetwork	#conte

COMMENTHERE

Repeat the search using a much lower K.

K = 5

```
system.time(lda_k_5 <- LDA(dtm, method= "Gibbs", k = 5, control = list(seed = 123)))
# save(lda_k_5, file = "data/lda_k_5.Rda")
```

Top terms 01	Top terms 02	Top terms 03	Top terms 04	Top terms 05
governo	grazie	#coronavirus	grazie	grazie
presidente	lavoro	governo	pass	governo
paese	anni	#covid19	anni	anni
italiani	draghi	grazie	via	salvini
lavoro	#draghi	decreto	green	lavoro
grazie	grande	imprese	lavoro	#iostoconsalvini
anni	vaccini	emergenza	draghi	forza
politica	forza	l'italia	roma	grande
grande	vaccinale	misure	#greenpass	cittadini
via	lega	momento	grande	#lega

COMMENTHERE

Titles:

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
titles	1	2	3	4	5
	governo	grazie	#coronavirus	grazie	grazie
	presidente	lavoro	governo	pass	governo
	paese	anni	#covid19	anni	anni
	italiani	draghi	grazie	via	salvini
	lavoro	#draghi	decreto	green	lavoro
	grazie	grande	imprese	lavoro	#iostoconsalvini
	anni	vaccini	emergenza	draghi	forza
	politica	forza	l'italia	roma	grande
	grande	vaccinale	misure	#greenpass	cittadini
	via	lega	momento	grande	#lega

COMMENTHERE