LDA Topicmodel

Riccardo Ruta riccardo.ruta@studenti.unimi.it

06/05/2022

Contents

PART I - CREATE THE DTM

1) Import and clean dataset

```
# import the data
tw <-  read_csv("data/large_files/politicians_all_final_tweets.csv",show_col_types = FALSE)

# Adjust datetime (Run code in this order!)
Sys.setlocale("LC_TIME", "C")
tw$date <- as.Date(strptime(tw$creato_il,"%a %b %d %H:%M:%S %z %Y", tz = "CET"))
tw$date <- na.replace(tw$date, as.Date(tw$creato_il))

# Create week variable
tw <- tw %>% mutate(week = cut.Date(date, breaks = "1 week", labels = FALSE))

# Create month variable
tw <- tw %>% mutate(month = cut.Date(date, breaks = "1 month", labels = FALSE))

# Remove missing from tweets column (using remove_na tidyverse)
tw <- tw %>% drop_na(tweet_testo)

# Remove space from genere variable
a <- unique(tw$genere)
tw$genere <- gsub(a[3],"male",tw$genere)

# Select variables for the analysis
dataset <- tw %>% select(nome, tweet_testo, genere, party_id,chamber,status, date, week, month )
```

2) Create the corpus

```
corp <- corpus(dataset, text = "tweet_testo")
```

3) Create the Dfm removing stopwords and trimming

```
my_word <- as.list(read_csv("data/it_stopwords_new_list.csv", show_col_types = FALSE))

my_list <- c(" ","c'è","+"," ",my_word$stopwords, stopwords('italian'))

rev_dfm <- dfm(tokens(corp, remove_symbols = TRUE, remove_url = TRUE),
               remove = my_list, tolower = TRUE,
               remove_punct = TRUE, remove_numbers=TRUE)

rev_dfm <-   dfm_trim(rev_dfm, min_termfreq = 0.95, termfreq_type = "quantile",
                      max_docfreq = 0.1, docfreq_type = "prop")

rev_dfm <- dfm_group(rev_dfm, groups= month)
```

```
topfeatures(rev_dfm, 20)
```

```
##     governo      grazie      lavoro       paese <U+0001F1EE><U+0001F1F9>  presidente
##       25991       20760       18274       16444       15196       14215
##      grande    italiani      italia    l'italia         via    politica
##       13606       11993       11955       11728       11495        9930
##   cittadini       forza    ministro     insieme       mondo       legge
##        9331        8474        8411        8139        7664        7383
##       donne solidarietà
##        7103        6685
```

```
kable(topfeatures(rev_dfm, 20),col.names = "topfeature")
```

	topfeature
governo	25991
grazie	20760
lavoro	18274
paese	16444
<U+0001F1EE><U+0001F1F9>	15196
presidente	14215
grande	13606
italiani	11993
italia	11955
l'italia	11728
via	11495
politica	9930
cittadini	9331
forza	8474
ministro	8411
insieme	8139
mondo	7664
legge	7383
donne	7103
solidarietà	6685

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

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

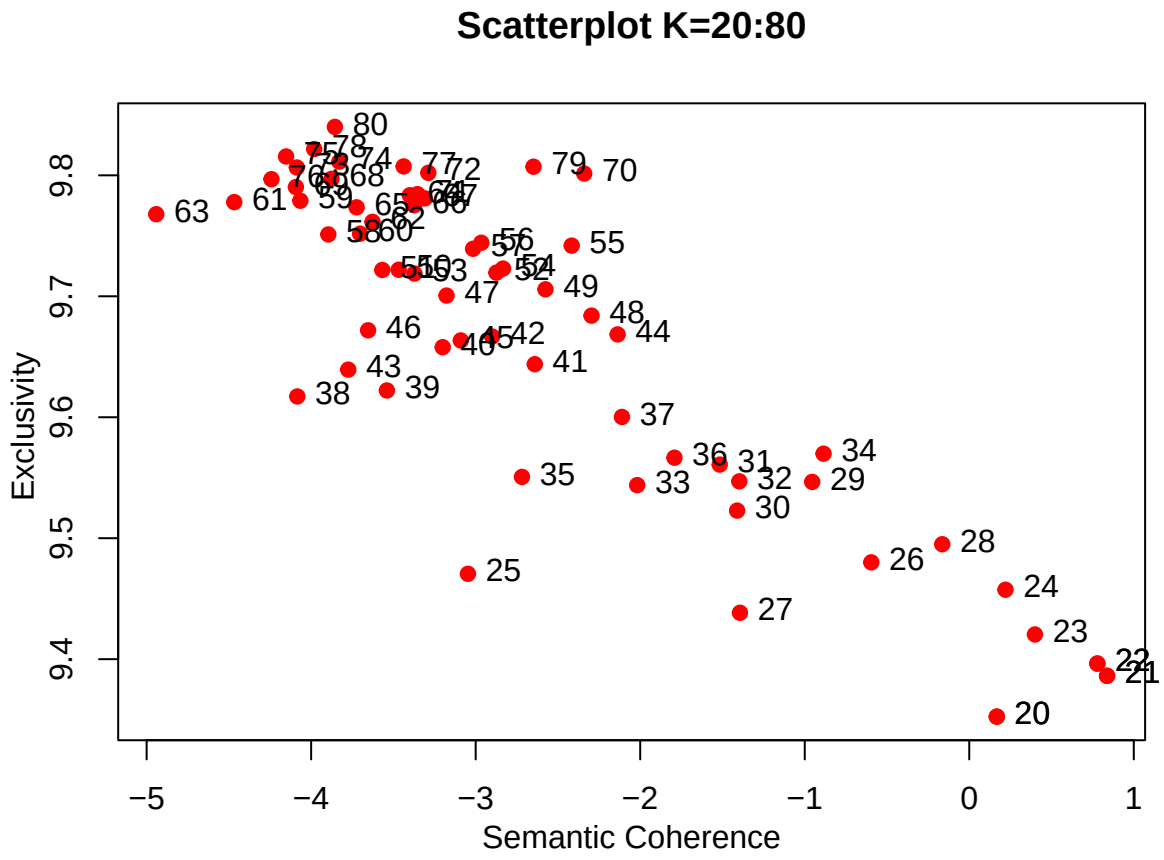
PART II - FIND THE BEST NUMBER OF TOPICS K

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



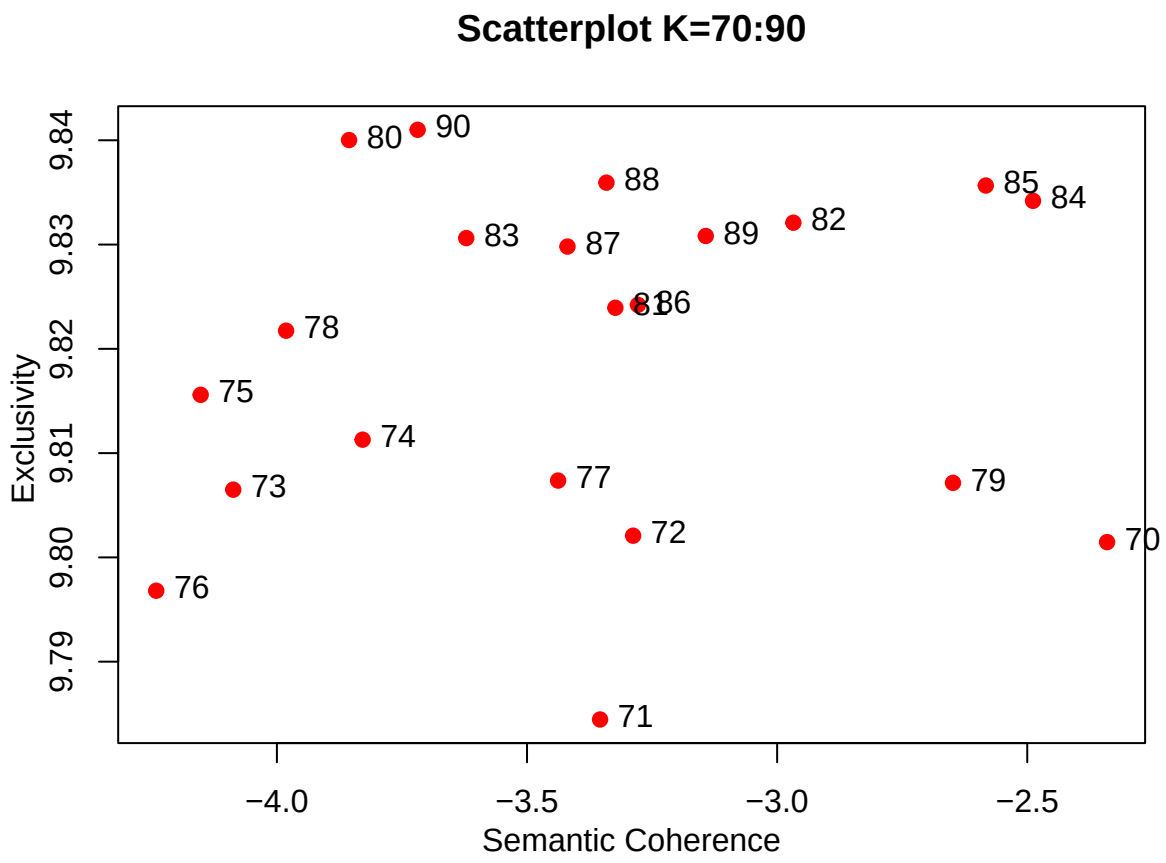
From this first try it's seems that the correct number of K is ……

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")
```

topic	coherence	exclusivity
70	-2.340422	9.801467
71	-3.353960	9.784454
72	-3.287822	9.802082
73	-4.087261	9.806496
74	-3.828622	9.811291
75	-4.152354	9.815591
76	-4.241225	9.796804
77	-3.437873	9.807368
78	-3.981590	9.821740
79	-2.648473	9.807149
80	-3.855703	9.840021
81	-3.323339	9.823939
82	-2.967526	9.832096
83	-3.621427	9.830621
84	-2.488446	9.834195
85	-2.582846	9.835661
86	-3.278275	9.824218
87	-3.419067	9.829809
88	-3.341212	9.835939
89	-3.142610	9.830824
90	-3.718580	9.841006

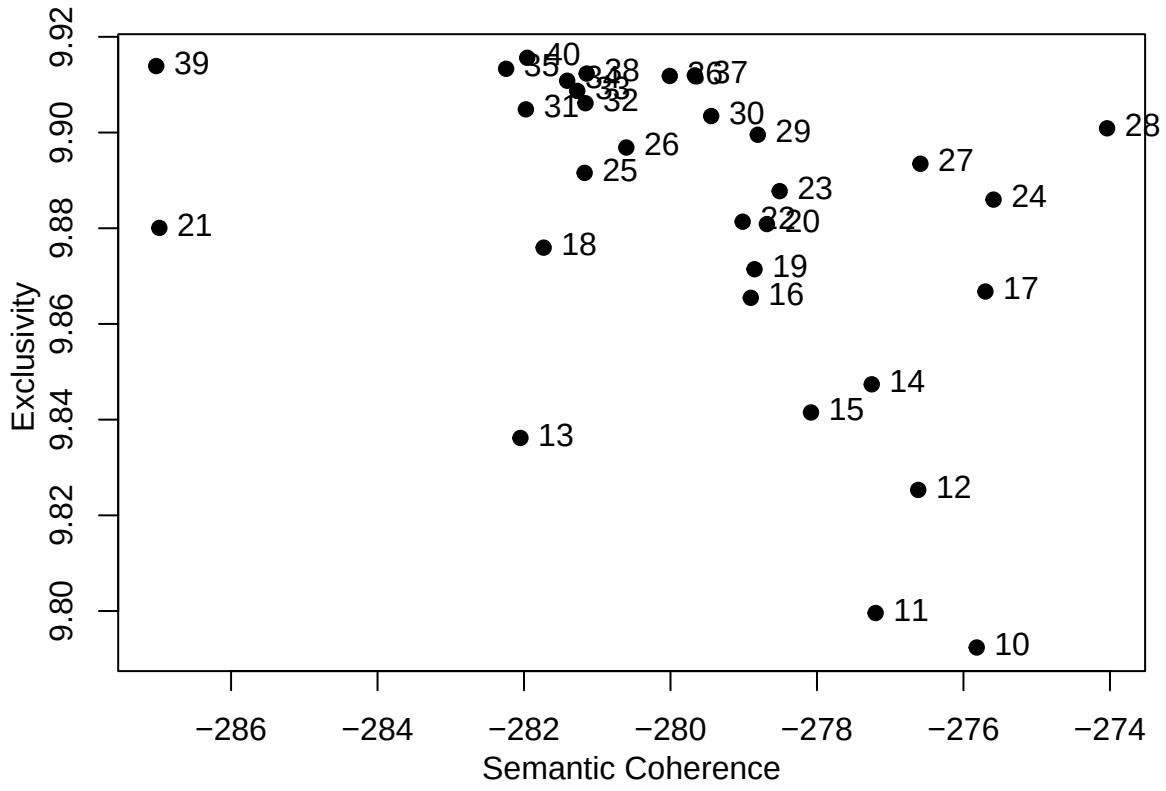


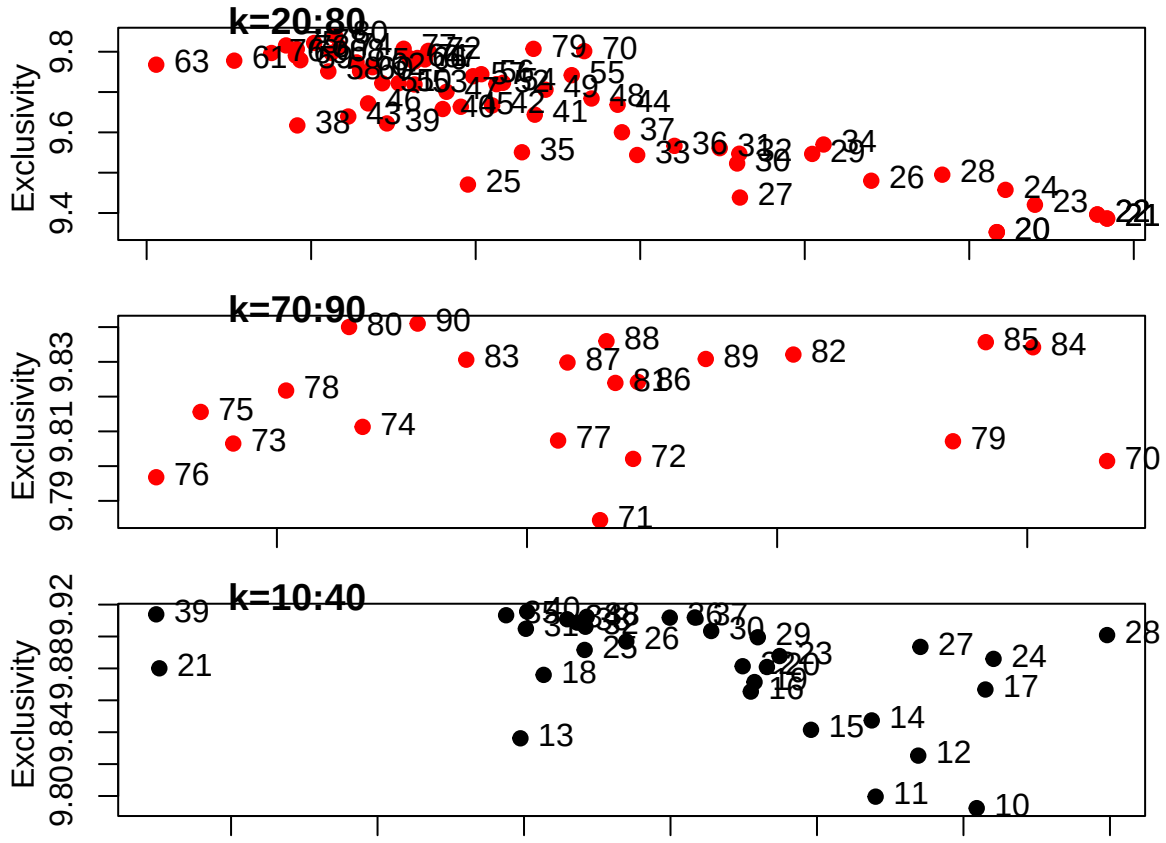
In this case … seems better than …

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")
```

topic	coherence	exclusivity
10	-275.8193	9.792377
11	-277.1995	9.799585
12	-276.6176	9.825317
13	-282.0491	9.836170
14	-277.2525	9.847386
15	-278.0818	9.841498
16	-278.9036	9.865459
17	-275.7004	9.866774
18	-281.7319	9.875940
19	-278.8529	9.871439
20	-278.6834	9.880872
21	-286.9793	9.880067
22	-279.0147	9.881363
23	-278.5099	9.887750
24	-275.5894	9.885981
25	-281.1725	9.891572
26	-280.6034	9.896865
27	-276.5881	9.893478
28	-274.0393	9.900883
29	-278.8082	9.899531
30	-279.4444	9.903447
31	-281.9744	9.904841
32	-281.1614	9.906131
33	-281.2734	9.908701
34	-281.4107	9.910822
35	-282.2445	9.913325
36	-280.0098	9.911821
37	-279.6674	9.911934
38	-281.1443	9.912321
39	-287.0217	9.913895
40	-281.9565	9.915620

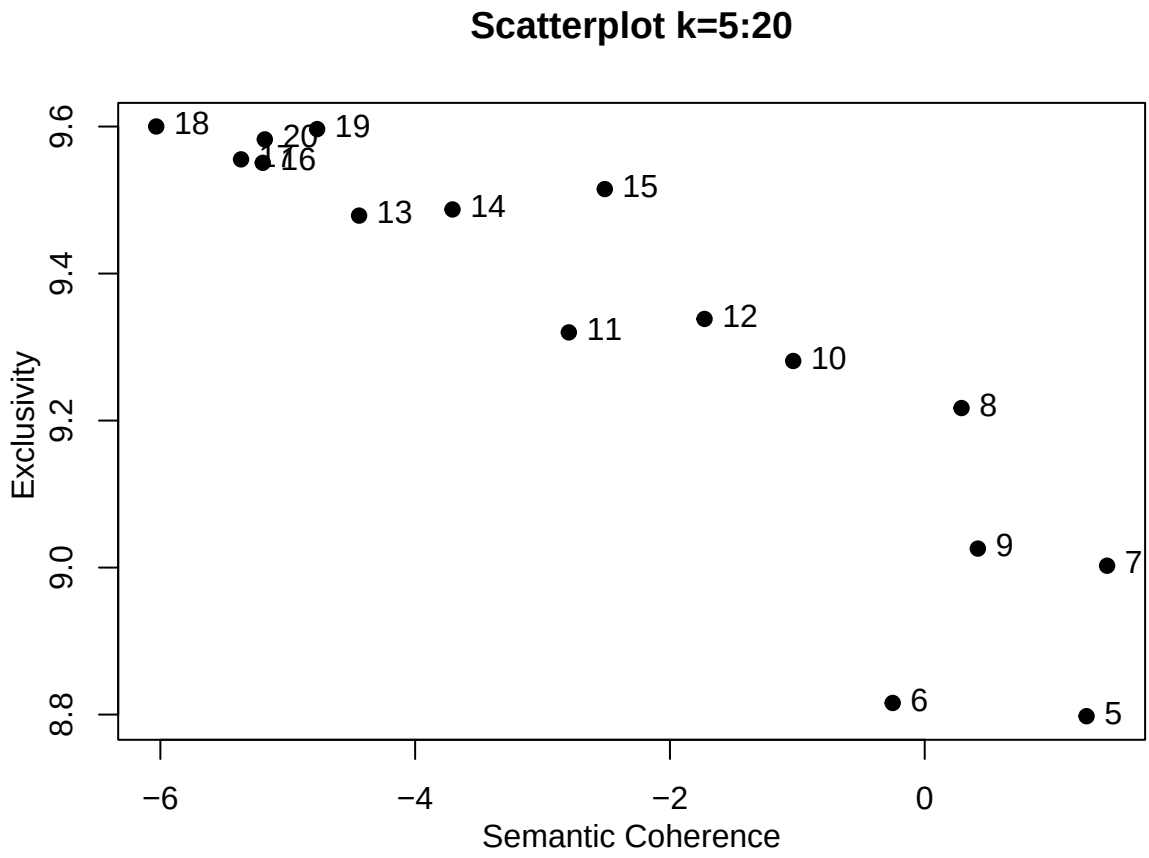

Scatterplot k=10:40
Exclusivity
Semantic Coherence
9.92
9.90
9.88
9.86
9.84
9.82
9.80
−286
−284
−282
−280
−278
−276
−274
39
40
35
28
37
31
32
30
29
28
26
27
25
23
24
21
22
20
18
19
16
17
14
15
13
12
11
10

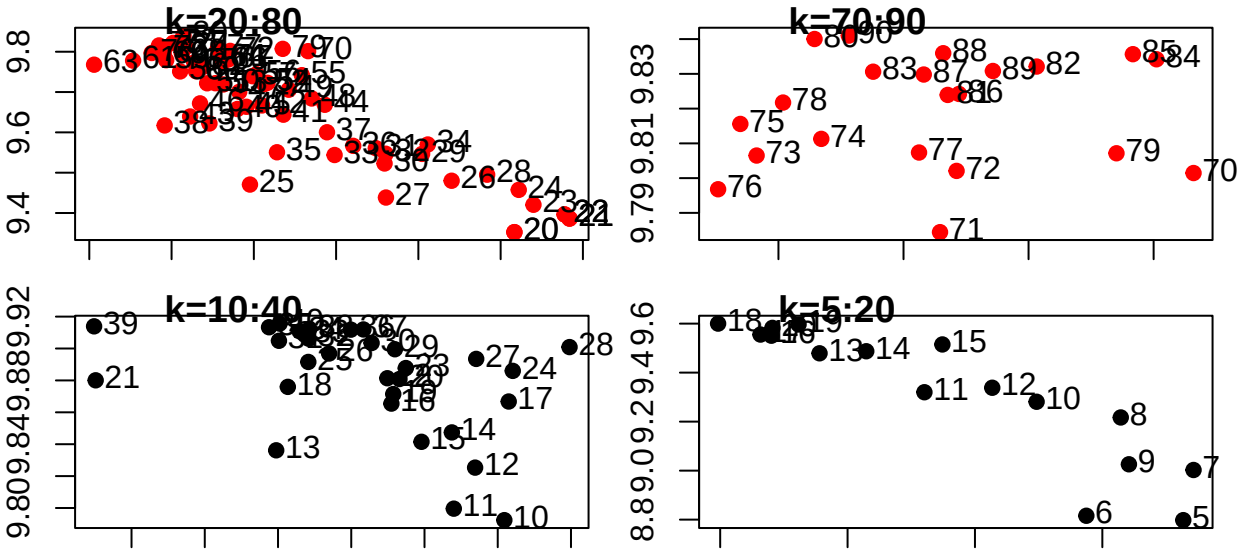


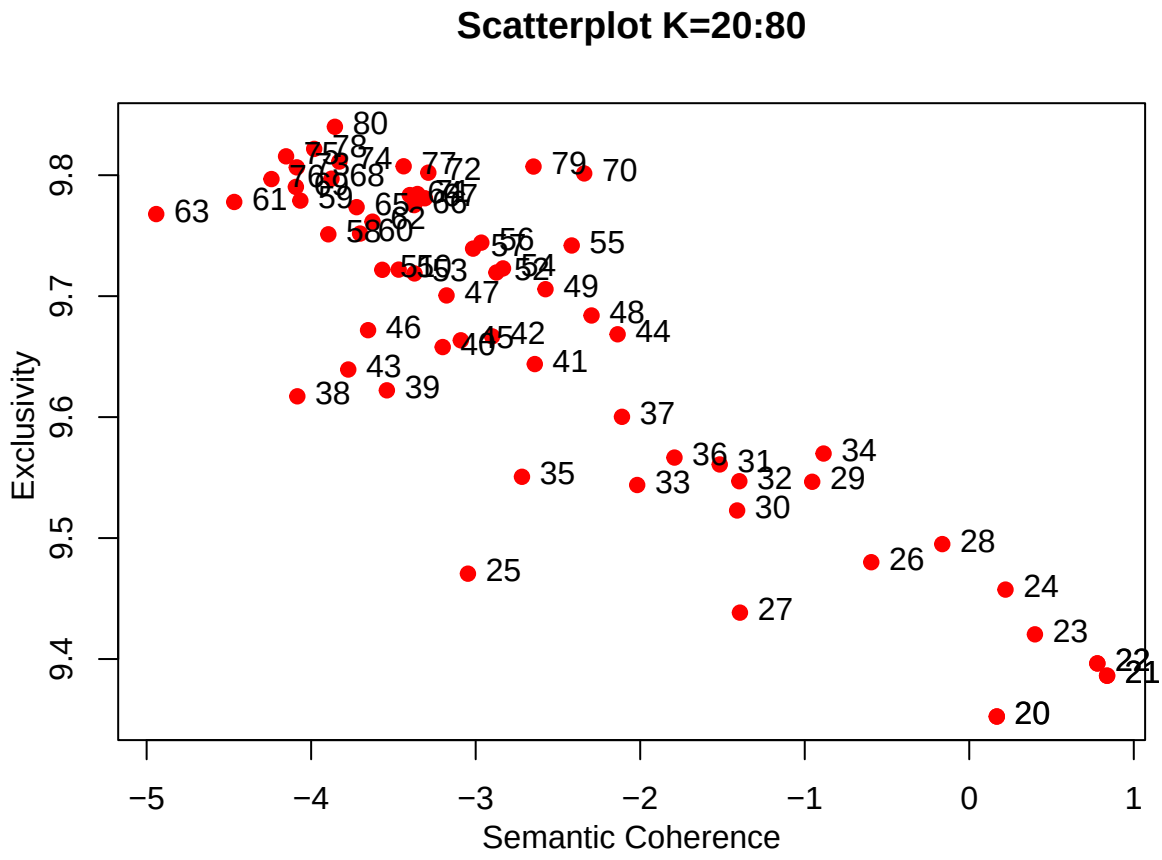
4) Fourth try k = 5:20 iteration n = 2000

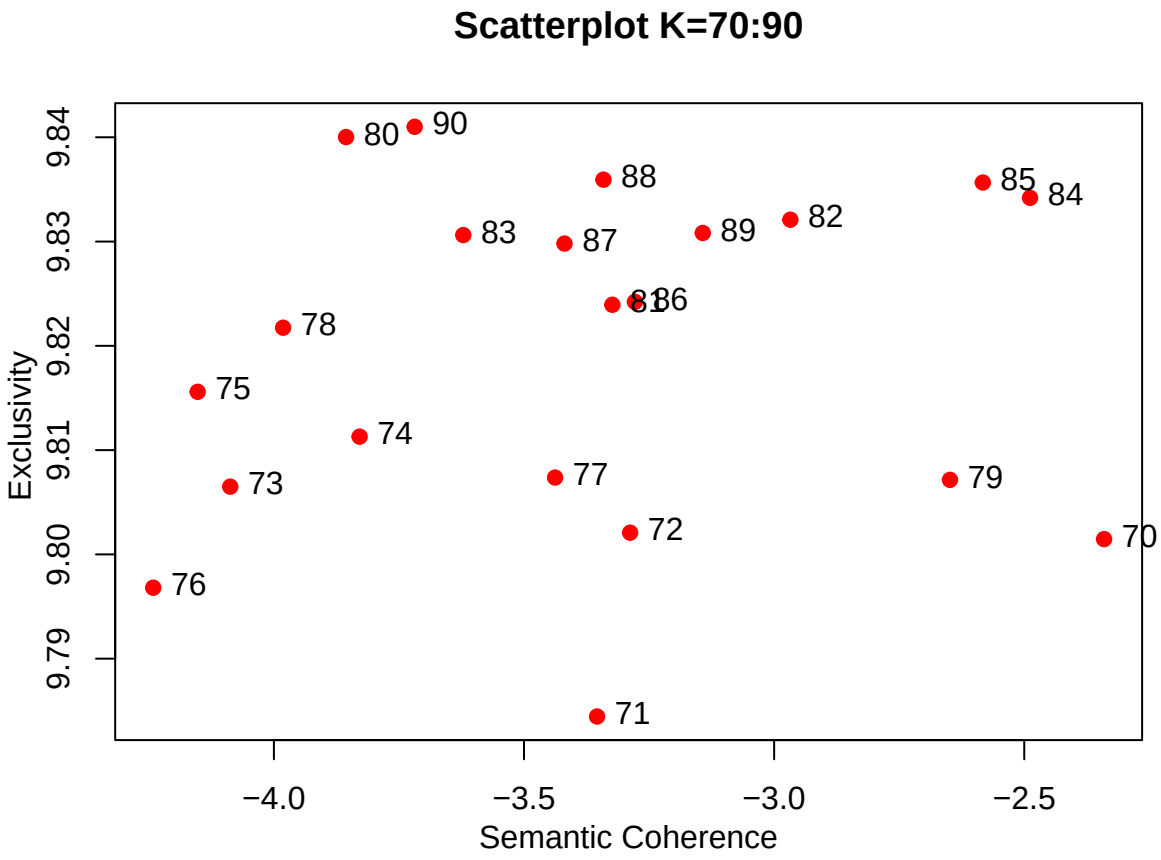
```
## Finding the best K
top1234 <- c(5:20)
top1234

## 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")
```

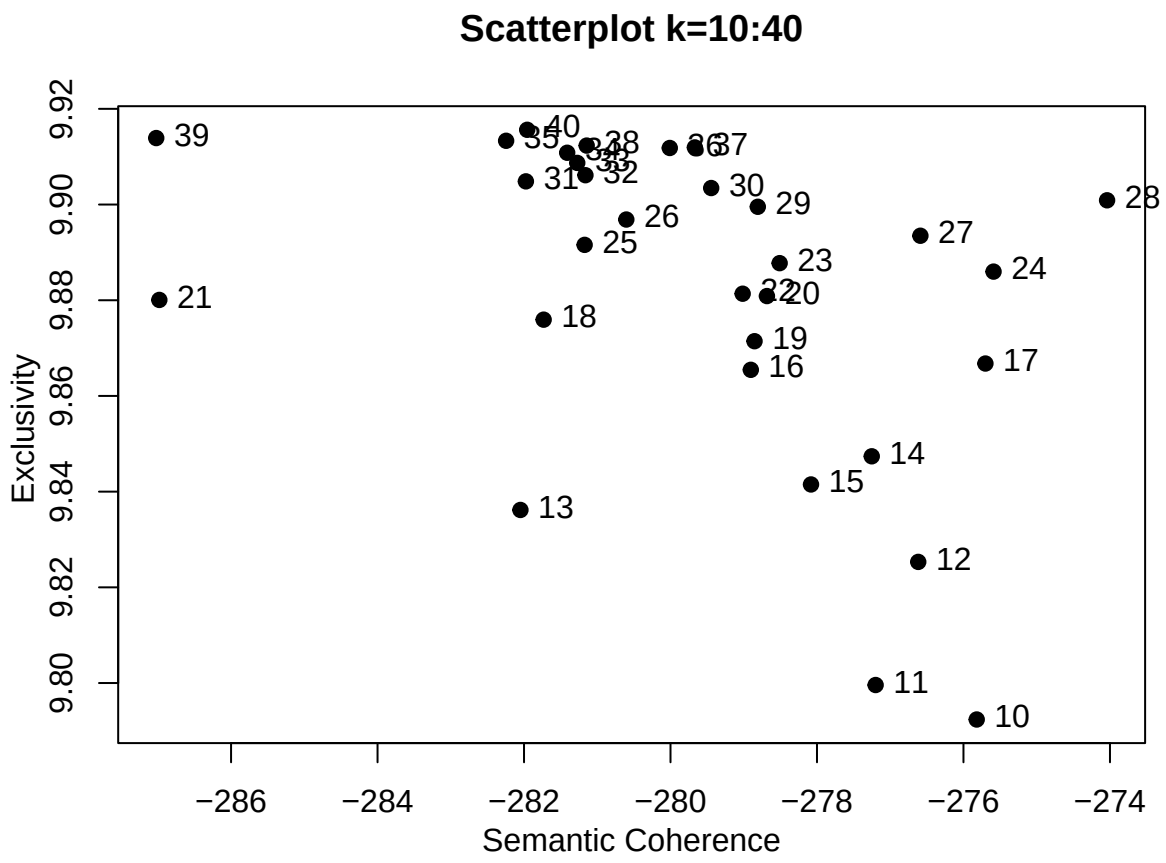
topic	coherence	exclusivity
5	1.2703475	8.797794
6	-0.2513313	8.815796
7	1.4315439	9.002518
8	0.2889243	9.217023
9	0.4171350	9.025830
10	-1.0321615	9.281054
11	-2.7934947	9.319961
12	-1.7284863	9.338250
13	-4.4400630	9.478833
14	-3.7067269	9.487235
15	-2.5113135	9.514924
16	-5.1962084	9.550672
17	-5.3668034	9.555340
18	-6.0324648	9.600086
19	-4.7693042	9.596540
20	-5.1799566	9.582518



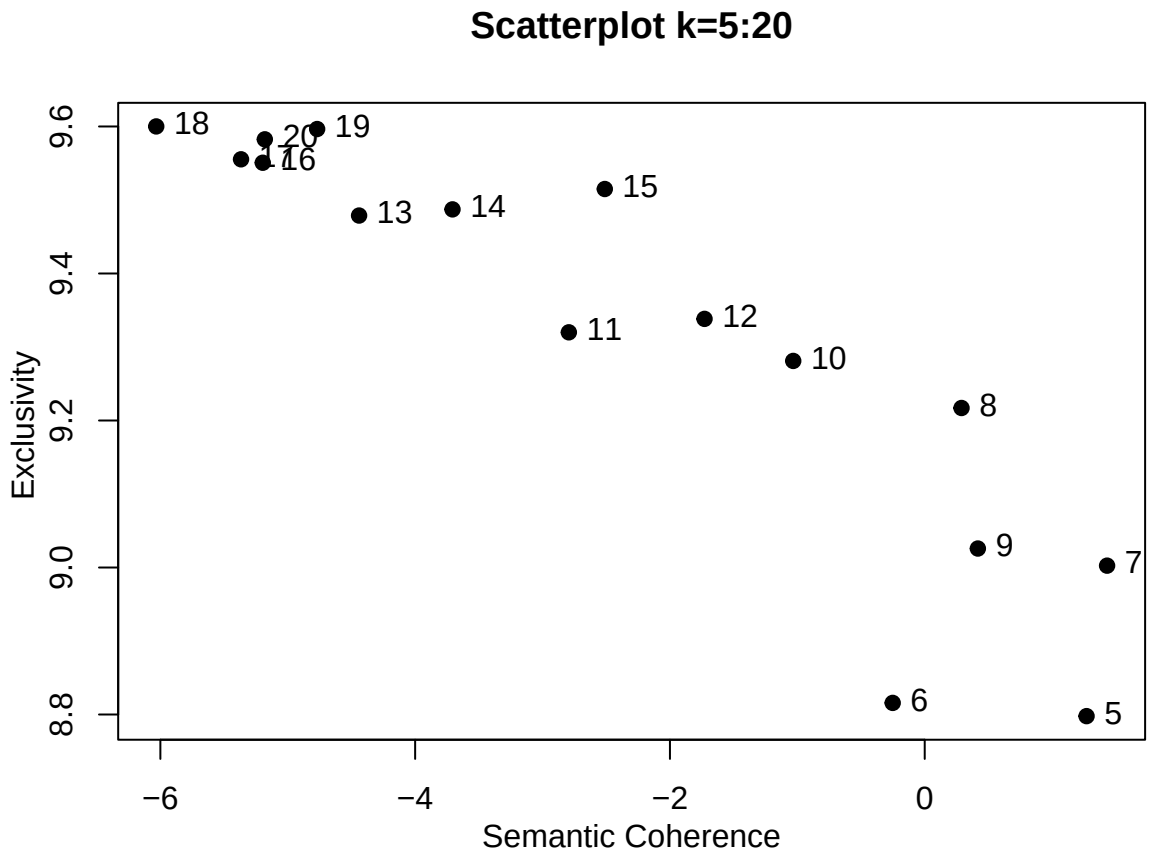

k=20:80
k=70:90
k=10:40
k=5:20


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
9.92
9.90
9.88
9.86
9.84
9.82
9.80
−286
−284
−282
−280
−278
−276
−274
39
40
35
28
37
31
32
30
29
28
26
27
25
23
24
21
22
20
18
19
16
17
14
15
13
12
11
10

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



After this try i can state that 30(81) is the best choice

PART III - ANALISYS OF THE TOPICS

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

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

	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

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
anni	grazie	#coronavirus	imprese	forza	solidarietà	italia	paese	via	governo
grande	lavoro	#covid19	milioni	lega	libertà	ministro	politica	diretta	italiani
storia	buon	covid	euro	roma	diritti	l'italia	presidente	domani	legge
mondo	donne	scuola	governo	@matteosalvinimi	rispetto	presidente	bene	de	conte
vittime	nazionale	sicurezza	lavoratori	@forza_italia	parole	the	futuro	intervista	parlamento
famiglia	importante	salute	decreto	#lega	democrazia	commissione	momento	intervento	@fratelliditalia
coraggio	giornata	pass	sostegno	salvini	giustizia	italiani	dobbiamo	sera	pd
presto	grande	dati	piano	#salvini	@stampasgarbi	consiglio	cittadini		draghi
italiano	auguri	emergenza	famiglie	regione	diritto	to	dare	parlare	voto
pensiero	mondo	green	risorse	sindaco	violenza	crisi	serve	appena	vuole

	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
	anni	grazie	#coronavirus	imprese	forza	solidarietà	italia	paese	via	governo
	grande	lavoro	#covid19	milioni	lega	libertà	ministro	politica	diretta	italiani
	storia	buon	covid	euro	roma	diritti	l'italia	presidente	domani	legge
	mondo	donne	scuola	governo	@matteosalvinimi	rispetto	presidente	bene	de	conte
	vittime	nazionale	sicurezza	lavoratori	@forza_italia	parole	the	futuro	intervista	parlamento
	famiglia	importante	salute	decreto	#lega	democrazia	commissione	momento	intervento	@fratelliditalia
	coraggio	giornata	pass	sostegno	salvini	giustizia	italiani	dobbiamo	sera	pd
	presto	grande	dati	piano	#salvini	@stampasgarbi	consiglio	cittadini		draghi
	italiano	auguri	emergenza	famiglie	regione	diritto	to	dare	parlare	voto
	pensiero	mondo	green	risorse	sindaco	violenza	crisi	serve	appena	vuole

Here i extract the most important terms from the model

Using 30 topics I imagined that..... COMMENT HERE

This method turns out to be cumbersome and not very informative because for many "topics" it is difficult for me to find a label.

So, I also decided to repeat the search using a much lower K.

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

Titles:

PART IV - RE-ANALISYS OF THE TOPICS as GENRE

I produced this pdf by filling out an R-Markdown, here i paste my complete script:

```
################################
## EXERCISE 4 --> TOPIC MODEL ##
################################

library(quanteda.textmodels)
library(quanteda)
library(topicmodels)
library(topicdoc)
library(cowplot)
library(ggplot2)
library(ggplotify)
library(knitr)
library(tibble)
library(kableExtra)

getwd()
setwd("C:/Users/Riccardo/Documents/UNIVERSITA'/MAGISTRALE/Big Data Analysis/LAB4/assignment_4")
getwd()

## Open the file with movie reviews inside the quanteda.textmodels package
data("data_corpus_moviereviews",
     package ="quanteda.textmodels")

## Create the corpus
corp <- tail(data_corpus_moviereviews,500) +
  head(data_corpus_moviereviews, 500)

## Extract the actual texts of the movie reviews and adding them as docvars in the corpus
docvars(corp, "texts") <- texts(corp)

## Create the Dfm removing stopwords
rev_dfm <- dfm(corp, remove =c(stopwords("english"), ("+"), ("<"), (">"),
                               ("rt"), ("00*"),("="),
                               ("`"), ("d"), ("r"), ("t"), ("m"), ("l"),
                               ("b"), ("g"), ("$"), ("j"), ("o"), ("u"),
                               ("e")), tolower = TRUE, stem = FALSE,
               remove_punct = TRUE, remove_numbers=TRUE)

## Timming the Dfm
rev_dfm <-   dfm_trim(rev_dfm, min_termfreq = 0.95, termfreq_type = "quantile",
                      max_docfreq = 0.1, docfreq_type = "prop")

kable(topfeatures(rev_dfm, 20))
```

```
## Keeping only documents with number of tokens >0
rev_dfm[ntoken(rev_dfm) == 0,]
rev_dfm <- rev_dfm[ntoken(rev_dfm) > 0,]

str(rev_dfm@docvars)

## Convert the Document Feature Matrix (Dfm) in a Topic Model (Dtm)
dtm <- convert(rev_dfm, to = "topicmodels")

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

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

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")
# load("data/results1.Rda")

kable(head(results1))
kable(tail(results1))
str(results1)


plot1 <- as.ggplot(~plot(results1$coherence, results1$exclusivity, main="Scatterplot K=20:80",
     xlab="Semantic Coherence", ylab="Exclusivity ", pch=19, col=ifelse(results1$coherence<=-155.8,"bla
       text(results1$coherence, results1$exclusivity,
            labels=results1$topic, cex= 1,  pos=4))

# ggsave("figs/plot1.png", plot = plot1)
plot1

## From this first try it's seems that the correct number of K is close to 71

#############################################################################################
## Now i repeat the procedure with 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")
# load("data/results2.Rda")

kable(results2)
str(results2)


plot2 <- as.ggplot(~plot(results2$coherence, results2$exclusivity, main="Scatterplot K=70:90",
                         xlab="Semantic Coherence", ylab="Exclusivity ", pch=19,
                         col=ifelse(results2$coherence > -155.3,"red","black")) +
                    text(results2$coherence, results2$exclusivity, labels=results2$topic, cex= 1,  pos=

# ggsave("figs/plot2.png", plot = plot2)
plot2
## 81 seems better than 71

##############################################################################################
## now i repeat the cycle with the correct number of iteration (2000) for 75:85

top3 <- c(75:95)
top3

results3 <- data.frame(first=vector(), second=vector(), third=vector())
results3

system.time(
  for (i  in top3)
  {
    set.seed(123)
    lda3 <- LDA(dtm, method= "Gibbs", k = (i), control=list(verbose=50L,
                                                            iter=2000))
    topic <- (i)
    coherence <- mean(topic_coherence(lda3, dtm))
    exclusivity <- mean(topic_exclusivity(lda3))
    results3 <- rbind(results3 , cbind(topic, coherence, exclusivity ))
  }
)
```

```
# save(results3,file="data/results3.Rda")
# load("data/results3.Rda")

kable(results3)
str(results3)

## k=81
plot3 <- as.ggplot(~plot(results3$coherence, results3$exclusivity, main="Scatterplot k=75:95",
     xlab="Semantic Coherence", ylab="Exclusivity ", pch=19, col=ifelse(results3$coherence<=-153.9 | res

       text(results3$coherence, results3$exclusivity,
            labels=results3$topic, cex= 1,  pos=4))
plot3
#ggsave("figs/plot3.png", plot = plot3)

grid <- plot_grid(plot1, plot2, plot3, labels = list("k=20:80", "k=70:90",
                                                     "k=75:95"),
                  label_x = .15, nrow = 3)

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

## After this try i can state that 81 is the best choice

#############################################################################################################
load("data/lda.Rda")
# system.time(lda <- LDA(dtm, method= "Gibbs", k = 81, control = list(seed = 123)))
# save(lda, file = "data/lda.Rda")

## Here i extract the most important terms from the model

terms <- get_terms(lda, 10)

dt1 <- terms[,1:10]
dt2 <- terms[,11:20]
dt3 <- terms[,21:30]
dt4 <- terms[,31:40]
dt5 <- terms[,41:50]
dt6 <- terms[,51:60]
dt7 <- terms[,61:70]
dt8 <- terms[,71:81]


knitr::kable(dt1, col.names = c("Top terms 01","Top terms 02","Top terms 03",
                          "Top terms 04","Top terms 05","Top terms 06","Top terms 07","Top terms 08","To
  kable_styling(latex_options = "scale_down")

knitr::kable(dt2, col.names = c("Top terms 11","Top terms 12","Top terms 13",
                          "Top terms 14","Top terms 15","Top terms 16","Top terms 17","Top terms 18","To
  kable_styling(latex_options = "scale_down")

knitr::kable(dt3, col.names = c("Top terms 21","Top terms 22","Top terms 23",
                          "Top terms 24","Top terms 25","Top terms 26","Top terms 27","Top terms 28","To
```

```
  kable_styling(latex_options = "scale_down")

knitr::kable(dt4, col.names = c("Top terms 31","Top terms 32","Top terms 33",
                          "Top terms 34","Top terms 35","Top terms 36","Top terms 37","Top terms 38","To
  kable_styling(latex_options = "scale_down")

knitr::kable(dt5, col.names = c("Top terms 41","Top terms 42","Top terms 43",
                          "Top terms 44","Top terms 45","Top terms 46","Top terms 47","Top terms 48","To
  kable_styling(latex_options = "scale_down")

knitr::kable(dt6, col.names = c("Top terms 51","Top terms 52","Top terms 53",
                          "Top terms 54","Top terms 55","Top terms 56","Top terms 57","Top terms 58","To
  kable_styling(latex_options = "scale_down")

knitr::kable(dt7, col.names = c("Top terms 61","Top terms 62","Top terms 63",
                          "Top terms 64","Top terms 65","Top terms 66","Top terms 67","Top terms 68","To
  kable_styling(latex_options = "scale_down")

knitr::kable(dt8, col.names = c("Top terms 71","Top terms 72","Top terms 73",
                          "Top terms 74","Top terms 75","Top terms 76","Top terms 77","Top terms 78","To
  kable_styling(latex_options = "scale_down")


titles <- c("shakespeare", "armageddon", "3", "the usual sospects", "seven", "shrek", "7", "8", "9", "10
            "11", "austin powers","griffin", "14", "15", "16", "the batman", "18", "19", "20",
            "godzilla", "22","the grat lebowski", "24", "25", "26", "27", "28", "29", "wonderful christm
            "patch adams", "32","tommy lee jones", "34", "35", "36", "37", "38", "39", "40",
            "mission impossible", "42","ted", "teen", "45", "eddie murphy", "47", "48", "49", "50",
            "the truman show", "52","song", "special effect", "55", "56", "disney cartoon", "john travol
            "61", "62","bill campbell", "64", "Saving private Ryan", "pulp fiction", "67", "aliens", "69
            "rocky", "72","73", "74", "75", "76", "77", "78", "79", "80", "81"
            )

table_titles <- rbind (titles, terms)

t1 <- table_titles[,1:10]
t2 <- table_titles[,11:20]
t3 <- table_titles[,21:30]
t4 <- table_titles[,31:40]
t5 <- table_titles[,41:50]
t6 <- table_titles[,51:60]
t7 <- table_titles[,61:70]
t8 <- table_titles[,71:81]

kable(t1)%>%
  kable_styling(latex_options = "scale_down")

kable(t1)%>%
  kable_styling(latex_options = "scale_down")

kable(t1)%>%
  kable_styling(latex_options = "scale_down")
```

```
kable(t1)%>%
  kable_styling(latex_options = "scale_down")

kable(t1)%>%
  kable_styling(latex_options = "scale_down")

kable(t1)%>%
  kable_styling(latex_options = "scale_down")

kable(t1)%>%
  kable_styling(latex_options = "scale_down")

kable(t1)%>%
  kable_styling(latex_options = "scale_down")

##############################################################################
## Repeat the procedure with K = 3:15 looking for film genre

genere <- c(3:15)

results.genere <- data.frame(first=vector(), second=vector(), third=vector())

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

# save(results.genere,file="data/results_genere.Rda")
# load("data/results_genere.Rda")

plot.genere <- as.ggplot(~plot(results.genere$coherence,
                               results.genere$exclusivity, main="Scatterplot K=3:9",xlab="Semantic Cohe
                         col=ifelse(results.genere$coherence > -155.3,"red","black")) + text(results.ge
                                                                                             labels=resu


# ggsave("figs/plot_genere.png", plot = plot.genere)

plot.genere


compar <- plot_grid(plot3, plot.genere, label_x = .15, ncol = 2 )

compar
# ggsave("figs/plot_compar.png", plot = compar)
```

```
#############################################################################################
## Looking for topics = genre 6 seems to be the best choice

load("data/lda_genere.Rda")
# system.time(lda.genere <- LDA(dtm, method= "Gibbs", k = 6, control = list(seed = 123)))
# save(lda.genere, file = "data/lda_genere.Rda")

## Extract the 15 most important terms from the model

terms2 <- get_terms(lda.genere, 15)

A <- kable(terms2, col.names = c("Top terms genre1", "Top terms genre2", "Top terms genre3", "Top terms
A

B <-kable(terms2, col.names = c("Action", "Thriller", "Comedy", "Fantasy", "Family", "Romance"))
B
```