

LDA Topicmodel

Riccardo Ruta

05/2022

Contents

PART I - CREATE THE DTM	1
PART II - FIND THE BEST NUMBER OF TOPICS K	2
PART III - ANALISYS OF THE TOPICS	4

PART I - CREATE THE DTM

1) Remove all the account's mentions

```
DFM_trimmed@Dimnames$features <- gsub("^@", "", DFM_trimmed@Dimnames$features)
```

2) 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) Search the best number of Topics comparing coherence and exclusivity values

```
# 10 : 50 iter 1000
top1 <- c(10:50)

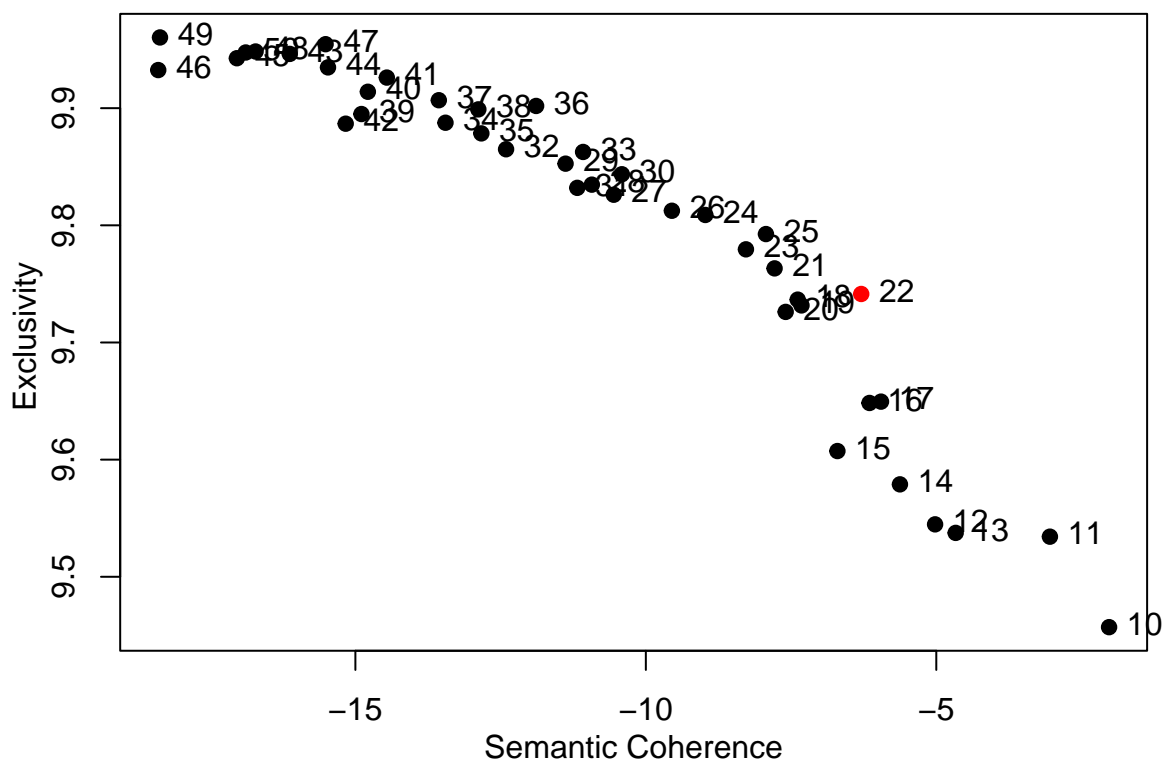
## let's create an empty data frame
risultati <- data.frame(first=vector(), second=vector(), third=vector())

system.time(
  for (i in top1)
  {
    set.seed(123)
    lda_test <- LDA(dtm, method= "Gibbs", k = (i), control=list(verbose=50L, iter=1000))
    topic <- (i)
    coherence_test <- mean(topic_coherence(lda_test, dtm))
    exclusivity_test <- mean(topic_exclusivity(lda_test))
    risultati <- rbind(risultati , cbind(topic, coherence_test, exclusivity_test ))
  }
)
# save(risultati, file="data/results_K_10-50.Rda")
```

K = 10:50

Plot the values of coherence and exclusivity in order to find the best K

Scatterplot k=10:50



Interpreting those data i can state that K= 22 has the best values of coherence and exclusivity.

PART III - ANALISYS OF THE TOPICS

Repeat the analysis selecting $k = 22$

```
system.time(lda_22 <- LDA(dtm, method= "Gibbs", k = 22, control = list(seed = 123)))
# save(lda, file = "data/lda_k_22.Rda")
```

The most important terms from the model, for each topic

Top terms 01	Top terms 02	Top terms 03	Top terms 04	Top terms 05	Top terms 06
vaccini	guerra	#draghi	grazie	#referendumgiustizia	#tokyo2020
aprile	#ucraina	draghi	mondo	luglio	#afghanistan
vaccinale	ucraina	governo	insieme	giustizia	talebani
pandemia	pace	#governodraghi	presidente	#ddlzan	agosto
@fratelliditalia	putin	lavoro	lavoro	gazebo	afghanistan
buon	marzo	paese	solidarietà	#primalitalia	medaglia
draghi	russia	@fratelliditalia	donne	giugno	oro
#draghi		presidente	paese	riforma	@fattoquotidiano
covid	popolo	buon	grande	firme	kabul
@pdnetwork	ucraino	vaccini	nazionale	#euro2020	medaglie

Top terms 7	Top terms 8	Top terms 9	Top terms 10	Top terms 11	Top terms 12
#coronavirus	maggio	sindaco	#iostoconsalvini	#iostoconsalvini	conte
misure	#ddlzan	ottobre	#oggivotolega	luglio	governo
momento	coprifuoco	pass	salvini	paese	crisi
coronavirus	@stampasgarbi	roma	#borgonzonipresidente	#salvini	#crisidigoverno
emergenza	vaccinale	@forza_italia	governo	giugno	#conte
grazie	#fratelliditalia	green	#salvini	#conte	maggioranza
l'emergenza	lavoro	città	#prescrizione	rilancio	paese
#covid19	#meloni	candidato	@fratelliditalia	conte	presidente
casa	#coprifuoco	#roma	#m5s	#2giugno	gennaio
medici	sinistra	elettorale	#emiliaromagna	#recoveryfund	fiducia

Top terms 13	Top terms 14	Top terms 15	Top terms 16	Top terms 17	Top terms 18
#dpcm	#fase2	presidente	draghi	the	natale
#iostoconsalvini	maggio	#quirinale	pass	@fratelliditalia	dicembre
ottobre	#mes	repubblica	green	to	bilancio
#mes	lavoro	#presidentedellarepubblica	covid	of	anno
jole	ripartire	#mattarella	via	and	#atreju21
covid	aprile	#quirinale2022	#greenpass	bilancio	@fattoquotidiano
contagi	imprese	quirinale	@fratelliditalia	covid	#natale
@fratelliditalia	liquidità	gennaio	vaccinati	#mes	euro
#conte	#forzalombardia	mattarella	@stampasgarbi	natale	buon
de	#recoveryfund	david	pandemia	#covid	auguri

Top terms 19	Top terms 20	Top terms 21	Top terms 22
governo	settembre	novembre	grazie
italiani	#iovotono	de	lavoro
#covid19	elettorale	violenza	anni
paese	@fratelliditalia	pass	politica
#governo	#referendum	@fattoquotidiano	governo
imprese	scuola	donne	grande
conte	parlamentari	reddito	bene
l'italia	#iostoconsalvini	et	l'italia
crisi	referendum	@theskeptical__	forza
decreto	#processateancheme	renzi	via

```
titles_22.1 <- c("1", "2", "3", "4", "5", "6")
titles_22.2 <- c("7", "8", "9", "10", "11", "12")
titles_22.3 <- c("13", "14", "15", "16", "17", "18")
titles_22.4 <- c("19", "20", "21", "22")
```

```
table_titles_22.1 <- rbind (titles_22.1, terms)
```

```
## Warning in rbind(titles_22.1, terms): number of columns of result is not a
## multiple of vector length (arg 1)
```

```
table_titles_22.2 <- rbind (titles_22.2, terms)
```

```
## Warning in rbind(titles_22.2, terms): number of columns of result is not a
## multiple of vector length (arg 1)
```

```
table_titles_22.3 <- rbind (titles_22.3, terms)
```

```
## Warning in rbind(titles_22.3, terms): number of columns of result is not a
## multiple of vector length (arg 1)
```

```
table_titles_22.4 <- rbind (titles_22.4, terms)
```

```
## Warning in rbind(titles_22.4, terms): number of columns of result is not a
## multiple of vector length (arg 1)
```

```
t22.1 <- table_titles_22.1[,1:6]
t22.2 <- table_titles_22.2[,7:12]
t22.3 <- table_titles_22.3[,13:18]
t22.4 <- table_titles_22.4[,19:22]
```

```
kable(t22.1)%>%
  kable_styling(latex_options = c("HOLD_position", "scale_down"))
```

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
titles_22.1	1	2	3	4	5	6
	vaccini	guerra	#draghi	grazie	#referendumgiustizia	#tokyo2020
	aprile	#ucraina	draghi	mondo	luglio	#afghanistan
	vaccinale	ucraina	governo	insieme	giustizia	talebani
	pandemia	pace	#governodraghi	presidente	#ddlzan	agosto
	@fratelliditalia	putin	lavoro	lavoro	gazebo	afghanistan
	buon	marzo	paese	solidarietà	#primalitalia	medaglia
	draghi	russia	@fratelliditalia	donne	giugno	oro
	#draghi		presidente	paese	riforma	@fattoquotidiano
	covid	popolo	buon	grande	firme	kabul
	@pdnetwork	ucraino	vaccini	nazionale	#euro2020	medaglie

```
kable(t22.2)%>%
  kable_styling(latex_options = c("HOLD_position", "scale_down"))
```

	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
titles_22.2	7	8	9	10	11	12
	#coronavirus	maggio	sindaco	#iostoconsalvini	#iostoconsalvini	conte
	misure	#ddlzan	ottobre	#oggivotolega	luglio	governo
	momento	coprifuoco	pass	salvini	paese	crisi
	coronavirus	@stampasgarbi	roma	#borgonzonipresidente	#salvini	#crisidigoverno
	emergenza	vaccinale	@forza_italia	governo	giugno	#conte
	grazie	#fratelliditalia	green	#salvini	#conte	maggioranza
	l'emergenza	lavoro	città	#prescrizione	rilancio	paese
	#covid19	#meloni	candidato	@fratelliditalia	conte	presidente
	casa	#coprifuoco	#roma	#m5s	#2giugno	gennaio
	medici	sinistra	elettorale	#emiliaromagna	#recoveryfund	fiducia

```
kable(t22.3)%>%
  kable_styling(latex_options = c("HOLD_position", "scale_down"))
```

	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18
titles_22.3	13	14	15	16	17	18
	#dpcm	#fase2	presidente	draghi	the	natale
	#iostoconsalvini	maggio	#quirinale	pass	@fratelliditalia	dicembre
	ottobre	#mes	repubblica	green	to	bilancio
	#mes	lavoro	#presidentedellarepubblica	covid	of	anno
	jole	ripartire	#mattarella	via	and	#atreju21
	covid	aprile	#quirinale2022	#greenpass	bilancio	@fattoquotidiano
	contagi	imprese	quirinale	@fratelliditalia	covid	#natale
	@fratelliditalia	liquidità	gennaio	vaccinati	#mes	euro
	#conte	#forzalombardia	mattarella	@stampasgarbi	natale	buon
	de	#recoveryfund	david	pandemia	#covid	auguri

```
kable(t22.4)%>%
  kable_styling(latex_options = c("HOLD_position", "scale_down"))
```

	Topic 19	Topic 20	Topic 21	Topic 22
titles_22.4	21	22	19	20
	governo	settembre	novembre	grazie
	italiani	#iovotono	de	lavoro
	#covid19	elettorale	violenza	anni
	paese	@fratelliditalia	pass	politica
	#governo	#referendum	@fattoquotidiano	governo
	imprese	scuola	donne	grande
	conte	parlamentari	reddito	bene
	l'italia	#iostoconsalvini	et	l'italia
	crisi	referendum	@theskeptical_	forza
	decreto	#processateancheme	renzi	via