Country Topic Models

2020-03-13T14:19:53-05:00

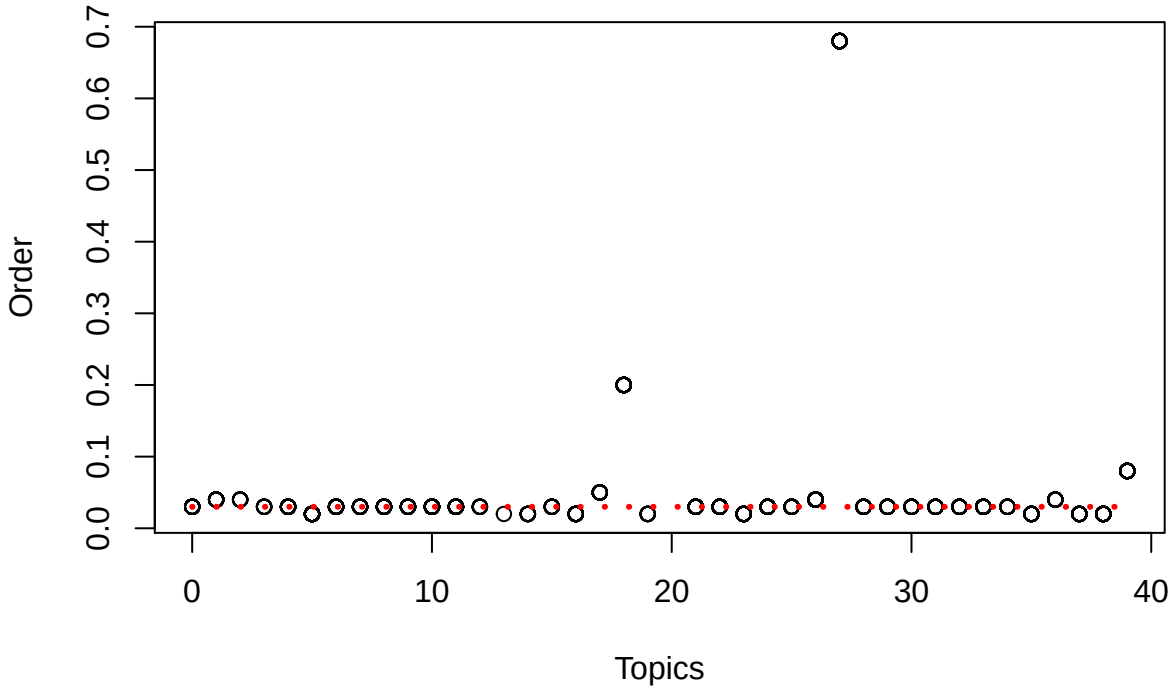
Here, we have a range of topics denoted in factors from 0 to 40.

```
## [1] dead red wrong open uh-huh energy drainin vert sex life dan fanned bleeding weight gamestop banks
## [2] watch calls mall underneath bloody ballin broke streets hittin big ball sky business royce dojo
## [3] slow snakes rat jump paper balmains quarterback relate found head
## [4] put pluto make bando trenches wine hard don party bout pray front messy
## [5] ptsd smell news mama affected shorter ryan fans forgot conceal shawl things andreas tank hanging
## [6] hot whip let's hah milli mouth bop good shits legit-ly hop finish cross fly woo cop swervin blun
## 40 Levels:                          ...
```

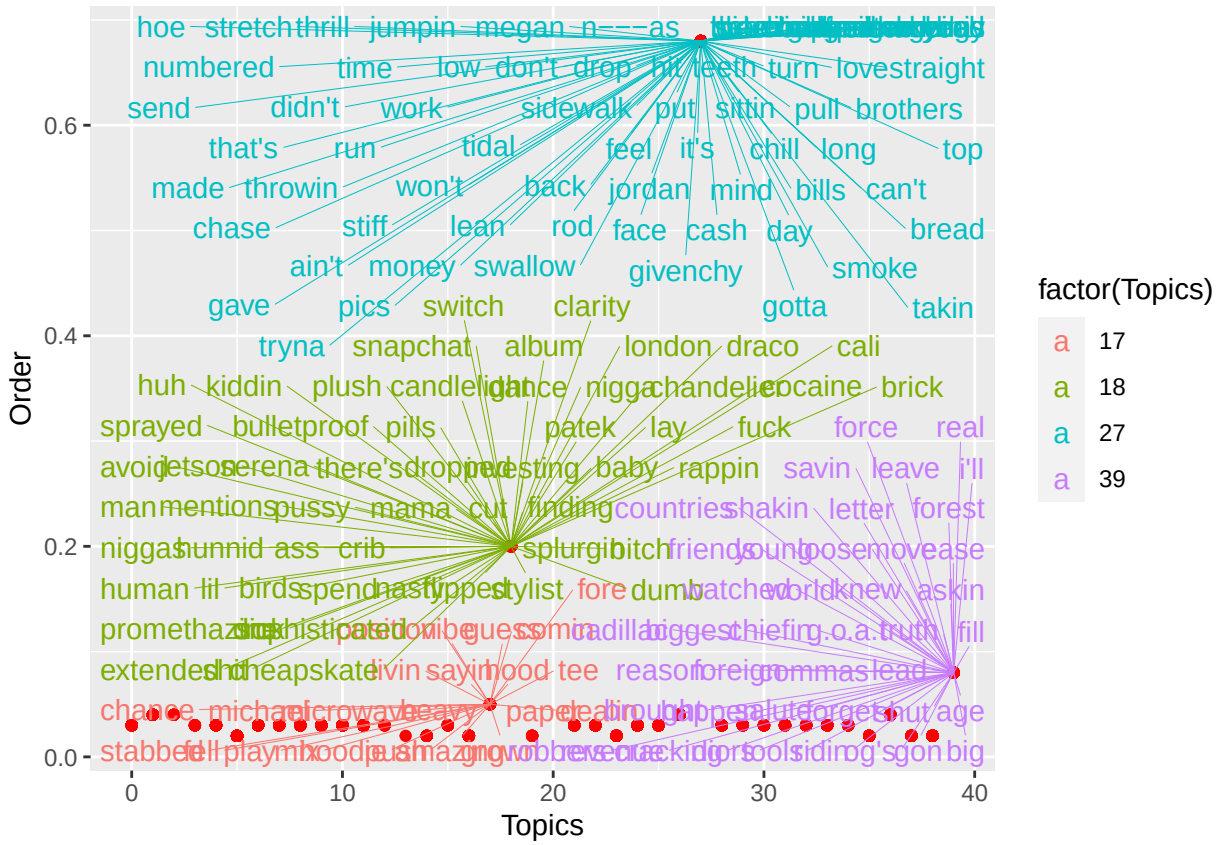
```
## [1] card follow flippin pluto law mama's month bruh hair decimals deals soda women tesla
## [2] goin spendin full haters hold telly celly murder told pay proud clear credit dripper december be
## [3] back store commercial swear put school chevy's corner aries flights attention savage band shows
## [4] usual call audemar'd spain virgil dollars nuts taxes workin lil caught cat alien worth bird work
## [5] yeah ain't boat choke note broke buy closet coats moncler walk heater bro rodeo buyin slick stor
## [6] big gon told young tryna forest move leave i'll chiefin age foreign ridin lead fill countries bi
## 40 Levels:                          ...
```

I graph them below. Along the y access is shown the percentage of probability that the set of words corresponding with one of the topics 0 through 40. Along the x axis is show the topic number.

```
## Loading required package: ggplot2
```



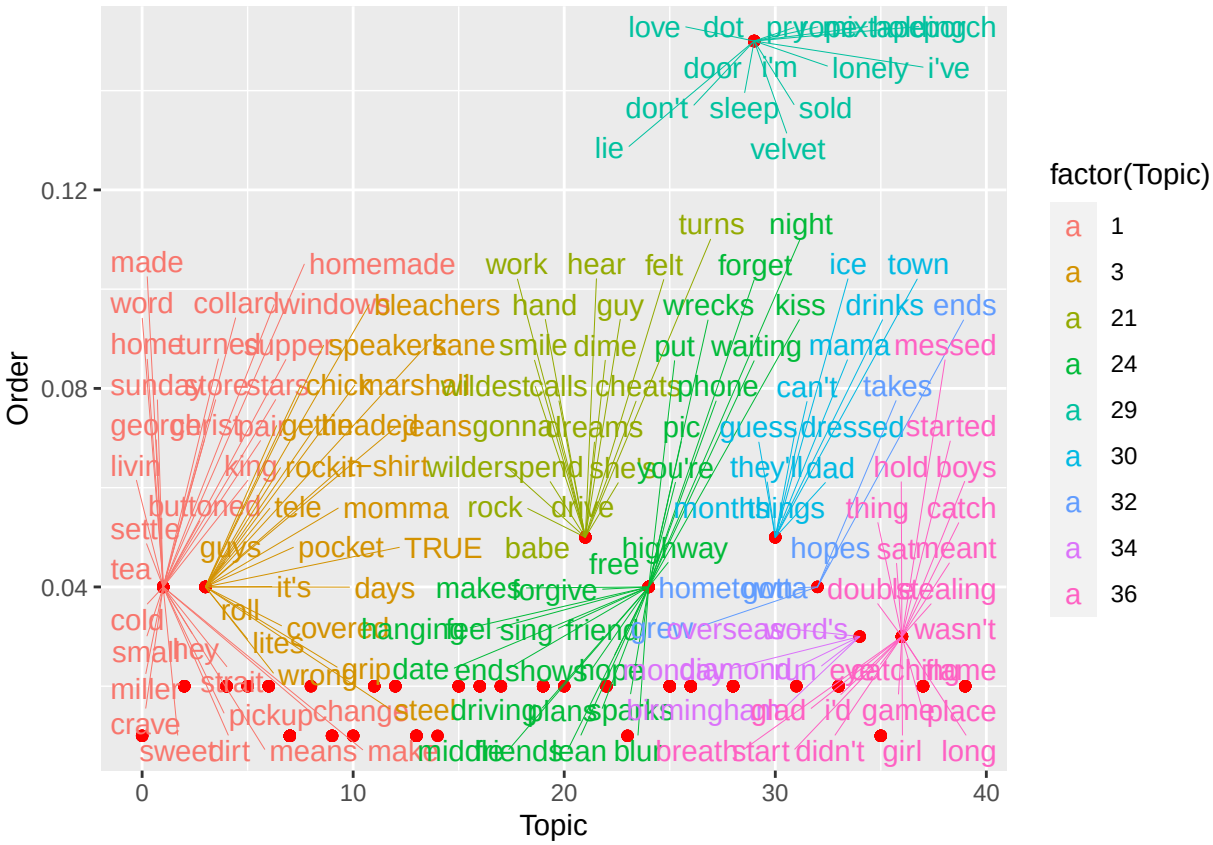
Next I will combine the two graphs by inserting the individual terms making up a particular topic.



Seems to get 68% of its words from topic number 27, 20% from topic number 18, 5% from topic 17 and 8% from topic 39. Anything less than 5% I chose not to show.

```
library(ggrepel)
library(gridExtra)
data(mtcars)
library(ggplot2)
setwd("~/mallet/")
keys1 <- read.csv("Country_keys.csv", header = TRUE, sep = ",")
Edges1 <- read.csv("Country_word_topics.csv", header = TRUE, sep = ",")
all1 <- merge(x = keys1, y = Edges1, by = "Topic")

ggplot(all1, aes(Topic, Order)) +
  geom_point(color = "red") +
  geom_text_repel(data = subset(all1, Order > 0.02),
 segment.size = 0.15,
segment.alpha = 1.0,
aes(
color = factor(Topic),
label = Words,
))
```



The thing that I find most interesting here, is that in the first graph there are no relationship words, love, friend, lonely, cheats, you, her, etcetera. However, in the second set, there are many relationship words.

In the next section, SA in the menu, I will show some preliminary sentiment analysis on these texts.

Notes