

amazon

dz1397

12/2/2020

```
library(quanteda)

## Package version: 2.1.2
## Parallel computing: 2 of 12 threads used.
## See https://quanteda.io for tutorials and examples.
##
## Attaching package: 'quanteda'
## The following object is masked from 'package:utils':
##
##      View

library(tidyr)
library(tidytext)
library(wordcloud)

## Loading required package: RColorBrewer

library(tm)

## Loading required package: NLP
##
## Attaching package: 'NLP'
## The following objects are masked from 'package:quanteda':
##
##      meta, meta<-
##
## Attaching package: 'tm'
## The following objects are masked from 'package:quanteda':
##
##      as.DocumentTermMatrix, stopwords
```

PART 1: Exploratory Data Analysis & Text Analytics

```
data1 <- read.csv("USvideos.csv")
```

1. How many complaints have been generated?

```
dim(data1)
```

```
## [1] 40949    16
```

```
colnames(data1)
```

```
## [1] "video_id"          "trending_date"      "title"
## [4] "channel_title"     "category_id"        "publish_time"
## [7] "tags"              "views"              "likes"
## [10] "dislikes"          "comment_count"      "thumbnail_link"
## [13] "comments_disabled" "ratings_disabled"    "video_error_or_removed"
## [16] "description"
```

```
process_list<- list("description","tags","title")
precorpus<- data1
```

```
dim(precorpus) # dim of file
```

```
## [1] 40949    16
```

```
for (i in process_list){
  precorpus[[i]] <- gsub("'", "", precorpus[[i]]) # remove apostrophes
precorpus[[i]] <- gsub("[:punct:]", " ", precorpus[[i]]) # replace punctuation with space
precorpus[[i]] <- gsub("[:cntrl:]", " ", precorpus[[i]]) # replace control characters with space
precorpus[[i]] <- gsub("^[:space:]+", "", precorpus[[i]]) # remove whitespace at beginning of documents
precorpus[[i]] <- gsub("[:space:]+$", "", precorpus[[i]]) # remove whitespace at end of documents
precorpus[[i]] <- gsub("[^a-zA-Z -]", " ", precorpus[[i]]) # allows only letters
precorpus[[i]] <- tolower(precorpus[[i]])
}
```

```
newscorpus_des<- corpus(precorpus$description,
                        docnames=precorpus$X,
                        docvar=data.frame(pos=precorpus$video_id,
                                           date=precorpus$trending_date,
                                           loc = precorpus$views))
```

```
newscorpus_tag<- corpus(precorpus$tags,
                        docnames=precorpus$X,
                        docvar=data.frame(pos=precorpus$video_id,
                                           date=precorpus$trending_date,
                                           loc = precorpus$views))
```

```
newscorpus_title<- corpus(precorpus$title,
                           docnames=precorpus$X,
                           docvar=data.frame(pos=precorpus$video_id,
                                              date=precorpus$trending_date,
                                              loc = precorpus$views))
```

```
newscorpus<-newscorpus_des
```

```
#explore the corpus
```

```
#summary(newscorpus) #summary of corpus
```

```
# create a custom dictionary
```

```
swlist = c("www", "http","https", "n", "com")
```

```
dfm.stem<- dfm(newscorpus,
               remove = c(swlist,stopwords("english")),
               verbose=F,
               stem=TRUE)
```

```
topfeatures(dfm.stem, n=50)
```

```
##      ly      bit      youtub  twitter  instagram    video    facebook
##      54462    51884    48900    41557    38833    37651    35484
##      nhttps      s      nhttp    watch    nfollow      us      goo
##      28966    27429    24997    20410    17901    17773    16810
##      gl      c      v      music    channel    new      youtu
##      16785    16483    16266    16017    15983    14927    13432
##      get      t  subscrib  smarturl    show    amzn      use
##      13375    13074    13065    12533    12017    11152    10967
##      ntwitter ninstagram    make    like    list  nsubscrib nfacebook
##      10923    10705    10661    10439    10380    10143    9782
##      can    product    news    one    now    time    live
##      9426    9392    9247    8660    8611    8507    8355
##      love      go    world    tumblr      m    nwatch    link
##      7959    7802    7777    7613    7237    7082    7035
##      nthe
##      6950
```

```
dfm.simple<- dfm(newscorpus,
  remove = c(swlist, stopwords("english")),
  verbose=F,
  stem=F)
```

```
set.seed(142)  #keeps cloud' shape fixed
freq<-topfeatures(dfm.simple, n=500)
```

```
wordcloud(names(freq),
  freq, max.words=200,
  scale=c(2, .7),
  colors=brewer.pal(8, "Dark2"))
```



```
newscorpus<-newscorpus_tag
```

```
swlist = c("www", "http", "https", "n", "com")
dfm.stem<- dfm(newscorpus,
               remove = c(swlist, stopwords("english")),
               verbose=F,
               stem=TRUE)
topfeatures(dfm.stem, n=50)
```

##	video	makeup	funni	music	new	show	movi	news
##	10938	7958	7323	7307	7108	6412	6202	5889
##	trailer	youtub	game	anim	food	comedi	live	best
##	5590	5261	5194	5155	4977	4768	4317	4073
##	challeng	cat	offici	tutori	review	black	star	make
##	3905	3786	3526	3485	3413	3325	3314	3288
##	interview	vlog	celebr	tv	world	vs	life	recip
##	3226	3162	3104	2942	2911	2906	2893	2877
##	late	song	war	react	beauti	s	cook	night
##	2819	2818	2741	2727	2720	2680	2647	2643
##	iphon	test	pop	kid	season	diy	entertain	jame
##	2618	2595	2591	2585	2576	2557	2512	2466
##	record	famili						
##	2455	2351						

```
set.seed(142)    #keeps cloud' shape fixed
freq<-topfeatures(dfm.stem, n=500)
```

```
wordcloud(names(freq),
          freq, max.words=200,
          scale=c(3, .1),
          colors=brewer.pal(8, "Dark2"))
```



```
newscorpus<-newscorpus_title
#explore the corpus
#summary(newscorpus) #summary of corpus
# create a custom dictionary

swlist = c("www", "http","https", "n", "com")

dfm.stem<- dfm(newscorpus,
               remove = c(swlist,stopwords("english")),
               verbose=F,
               stem=TRUE)
topfeatures(dfm.stem, n=50)
```

##	offici	video	trailer	ft	s	make	vs	new
##	4039	2952	2258	1315	1192	1072	1046	1008
##	live	music	makeup	audio	day	hd	first	star
##	983	941	876	872	817	796	762	754
##	get	game	tri	challeng	show	time	war	full
##	749	749	700	675	666	665	641	621
##	love	movi	lyric	black	world	test	one	season
##	615	609	607	593	580	547	546	520

```
set.seed(142)    #keeps cloud' shape fixed
freq<-topfeatures(dfm.stem, n=500)

wordcloud(names(freq),
           freq, max.words=200,
           scale=c(3, .7),
           colors=brewer.pal(8, "Dark2"))
```

