Exploración y curado de chistes

28/02/2022

Contenidos

1 Primera aproximacion de NLP para el análisis de un conjunto de chistes con título

1.1 Carga de datos

```
data_raw=read_csv("data/tots.csv")
```

```
## Rows: 840 Columns: 3
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): titulo, texto
## dbl (1): id
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
str(data_raw)
```

```
## spec_tbl_df [840 x 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ id    : num [1:840] 1 2 3 4 5 6 7 8 9 10 ...
##  $ titulo: chr [1:840] "Meter la PATA (versión chiste)" "¿Me engañas?" "Quitar el autocorrector" "Si
##  $ texto : chr [1:840] "- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy ma
##  - attr(*, "spec")=
##   .. cols(
##   ..   id = col_double(),
##   ..   titulo = col_character(),
##   ..   texto = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
knitr::kable(head(data_raw,20))
```

id	titulo	texto
1	Meter la PATA (versión chiste)	- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es verdad, si te vi embarazada.#- Es adoptada#- Mejor me voy
2	¿Me engañas?	–Cariño, ¿me engañas con otra?#–Vale.
3	Quitar el autocorrector	–¿¿¿Qué queremos???#–¡¡¡Quitar el autocorrector al movil!!!#–¿¿¿Cuándo lo queremos???#–¡Ahorca!#–¡Ahorro!#–¡Aborda!#–¡Albora!
4	Sigues siendo	– Cariño, está lloviendo y sigues siendo una tonta.#– ¿Qué?#– Me dijiste que con el tiempo cambiarias...
5	Como un moco	La vida es como un moco: intragable, dura y a veces no te deja respirar.
6	En el bote	-No te das cuenta pero ¡TE TENGO EN EL BOTE!#-¡Deja de decir gilipolleces y REMA! ¡Subnormal!
7	El trozo pequeño	-¿Me das un trozo pequeño de pastel? Pero así, súper chiquitito, que estoy a dieta.#-¿Así? ¿como los otros siete?.#-Sí, porfa
8	Me siento solo	-Me siento solo.#-Yo también, sentarse es fácil.
9	Llamadas del oftalmólogo	Tengo tres llamadas perdidas de mi oftalmólogo. El de ver me llama...
10	A la rumana	-Camarero, ponga una de calamares a la rumana.#-Perdón, señor, será a la romana.#-Irina, cariño, dile al gilipollas éste de dónde eres...
11	Natural de..	- ¿Me pone un zumo de piña?#- ¿Natural...?#- De Pontevedra, pero no creo que eso importe mucho...
12	Signos	- ¿De qué signo es tu mujer?#- Debe ser de exclamación, porque se pasa el día gritándome...
13	Capital de España	- Capital de España?#- La mayor parte en Suiza.
14	Día del abuelo	-Papi, ¡feliz día del abuelo!#-Ay hijita pero no tengo ningún nieto.#-¡SORPRESAAA!
15	Chino	-Como se dice en chino marinero pobre?#-Chin chu lancha.
16	Manzanas crueles	Hay un montón de manzanas en un árbol y de repente una se cae. Todas las de arriba empiezan a reírse y a burlarse de la que se ha caído y ésta responde:#- No os riáis, ¡Inmaduras!
17	Qué me das por mi marido	Dos amigas hablando:#- María, ¿qué me das por mi marido?#- Nada.#- ¡Trato hecho!
18	El peo viajante	Primer acto: Un peo volando por Londres.#Segundo acto: El mismo peo volando por Berlín.#Tercer acto: El mismo peo volando por París.#¿Cómo se llama la película?#El europeo.
19	¿Soy adoptado?	- Papá, ¿soy adoptado?#- ¿Tú crees que te habríamos elegido a ti?
20	Sujetador a la vista	- Cariño, se te ve el sujetador.#- Eso no es malo. Hay quien lo enseña adrede.#- Lo que tú digas, pero así no salgo contigo a la calle, Juanjo.

```
colnames(data_raw)
```

```
## [1] "id"     "titulo" "texto"
```

```
text=data_raw$texto
tabla=table(unlist(lapply(text,FUN=function(x) Encoding(x))))
head(text)
```

```
## [1] "- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es
## [2] "-Cariño, ¿me engañas con otra?#-Vale."
## [3] "-¿¿¿Qué queremos???#-¡¡¡Quitar el autocorrector al movil!!!#-¿¿¿Cuándo lo queremos???#-¡Ahorca!#
## [4] "- Cariño, está lloviendo y sigues siendo una tonta.#- ¿Qué?#- Me dijiste que con el tiempo cambi
```

```
## [5] "La vida es como un moco: intragable, dura y a veces no te deja respirar."
## [6] "-No te das cuenta pero ¡TE TENGO EN EL BOTE!#-¡Deja de decir gilipolleces y REMA! ¡Subnormal!"
```

```
head(tabla)
```

```
##
## unknown   UTF-8
##      62     778
```

```
library(dplyr)
text_df <- tibble(line = 1:length(text), text_raw =text)%>%
  mutate(Enconding=Encoding(text_raw),text_utf8=enc2utf8(text))
```

1.2 Extracción del diccionario raw empírico desde los chistes

Extraemos al dic_raw_1 todas las palabras que aparecen con separación espacio.

Criterios iniciales:

- Decidimos enconding a UTF-8 columna text_utf8 si hay que depurar por enconding habrá que ver cómo.
- Hay que decidir qué se hace con los CARACTERES SPECIALES:{,:; () ¿?!!}. De momento los voy a eliminar
- Todas las MAYÚSCULAS a MINÚSCULAS
- De momento NO SE ELIMINAN DIGITOS: se quedan tal cual, hay que distinguir los de los dígitos de años.
- No catalogamos idiomas.... se supone que todo está en castellano o términos técnicos que añadiremos
- Castellano es toda palabra o derivado de palabra que se encuentre en un spelling estándar de castellano que podemos ir adaptando.

```
library(tidytext)
glimpse(text_df)
```

```
## Rows: 840
## Columns: 4
## $ line      <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1~
## $ text_raw  <chr> "- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fuera~
## $ Enconding <chr> "UTF-8", "UTF-8", "UTF-8", "UTF-8", "unknown", "UTF-8", "UTF~
## $ text_utf8 <chr> "- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fuera~
```

```
text_raw=text_df %>% unnest_tokens(word, text_utf8)
glimpse(text_raw)
```

```
## Rows: 20,254
## Columns: 4
## $ line      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ text_raw  <chr> "- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fuera~
## $ Enconding <chr> "UTF-8", "UTF-8", "UTF-8", "UTF-8", "UTF-8", "UTF-8", "UTF-8~
## $ word      <chr> "que", "niño", "tan", "feo", "es", "mi", "hija", "ah", "no",~
```

```
knitr::kable(head(text_raw,20))
```

line	text_raw	Enconding	word
1	- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es verdad, si te vi embarazada.#- Es adoptada#- Mejor me voy	UTF-8	que
1	- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es verdad, si te vi embarazada.#- Es adoptada#- Mejor me voy	UTF-8	niño

line	text_raw	Encoding	word
1	- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es verdad, si te vi embarazada.#- Es adoptada#- Mejor me voy	UTF-8	tan
1	- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es verdad, si te vi embarazada.#- Es adoptada#- Mejor me voy	UTF-8	feo
1	- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es verdad, si te vi embarazada.#- Es adoptada#- Mejor me voy	UTF-8	es
1	- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es verdad, si te vi embarazada.#- Es adoptada#- Mejor me voy	UTF-8	mi
1	- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es verdad, si te vi embarazada.#- Es adoptada#- Mejor me voy	UTF-8	hija
1	- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es verdad, si te vi embarazada.#- Es adoptada#- Mejor me voy	UTF-8	ah
1	- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es verdad, si te vi embarazada.#- Es adoptada#- Mejor me voy	UTF-8	no
1	- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es verdad, si te vi embarazada.#- Es adoptada#- Mejor me voy	UTF-8	sabía
1	- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es verdad, si te vi embarazada.#- Es adoptada#- Mejor me voy	UTF-8	que
1	- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es verdad, si te vi embarazada.#- Es adoptada#- Mejor me voy	UTF-8	fueras
1	- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es verdad, si te vi embarazada.#- Es adoptada#- Mejor me voy	UTF-8	padre
1	- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es verdad, si te vi embarazada.#- Es adoptada#- Mejor me voy	UTF-8	soy
1	- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es verdad, si te vi embarazada.#- Es adoptada#- Mejor me voy	UTF-8	madre
1	- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es verdad, si te vi embarazada.#- Es adoptada#- Mejor me voy	UTF-8	ah
1	- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es verdad, si te vi embarazada.#- Es adoptada#- Mejor me voy	UTF-8	si
1	- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es verdad, si te vi embarazada.#- Es adoptada#- Mejor me voy	UTF-8	es
1	- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es verdad, si te vi embarazada.#- Es adoptada#- Mejor me voy	UTF-8	verdad
1	- Que niño tan feo#- Es mi hija...#- Ah! no sabía que fueras padre.#- Soy madre...#- Ah! si! es verdad, si te vi embarazada.#- Es adoptada#- Mejor me voy	UTF-8	si

```
dic_raw_1=sort(unique(text_raw$word))
nw=length(dic_raw_1)# Hay 1627 palabras
nw
```

```
## [1] 3948
```

1.3 Construcción del modelo de diccionario

Construiremos una tabla de modelado del corpus de palabras de los chistes:

- Como primary key la word (las `nw` words) (desde el text_raw en utf8)
- Su frecuencia: número de veces que aparece en los diagnósticos
- Si es correcta según un spelling de español de España (hay que buscar... qué hay mejor)

```
count_freq=text_raw %>% group_by(word) %>% summarise(N=n())
```

```
dic_raw_1 = tibble(word=dic_raw_1) %>% left_join(count_freq,by="word")
```

Ahora vemos claramente cómo podemos mejorar las words para UNIFICARLAS en un único "léxico" que nos permita un tratamiento unificado, auqnue las variantes escritas podrían tener significado humorístico.

Ejemplos

Palabras que contienen "zq"

```
dic_raw_1[grep("zq",dic_raw_1$word),]
```

```
## # A tibble: 2 x 2
##   word          N
##   <chr>     <int>
## 1 izquierda     2
## 2 izquierdo     1
```

Palabras que contienen "ch"

```
dic_raw_1[grep("(ch)",dic_raw_1$word),]
```

```
## # A tibble: 109 x 2
##    word           N
##    <chr>      <int>
##  1 agaché         1
##  2 ancho          1
##  3 anchoas        1
##  4 anoche         4
##  5 aprovecha      1
##  6 baches         1
##  7 bachiller      1
##  8 borracho      13
##  9 borrachos      4
## 10 cachichien     1
## # ... with 99 more rows
```

Palabras (dos palabras) con :

```
dic_raw_1[grep(":",dic_raw_1$word),]
```

```
## # A tibble: 3 x 2
##   word          N
##   <chr>     <int>
## 1 1acto:un      1
## 2 2acto:una     1
## 3 3acto:el      1
```

1.3.1 Añadimos columna de spelling al diccionario

Primero veamos algunos ejemplos de las sugerencias: ver manual en de hunspell. Github diccionarios open office

```
library("spelling")
library("hunspell")
#https://github.com/titoBouzout/Dictionaries # do
#es=dictionary(lang = "diccionarios/es_ES.dic", affix = "diccionarios/es_ES.dic", add_words = NULL,  ca
es_ES<- dictionary("diccionarios/es_ES.dic")
#print(es_ES)
list_dictionaries()# estos son los que  vienen por defecto
```

```
## [1] "en_AU" "en_CA" "en_GB" "en_US"
```

```
hunspell_check(c("bieja","colon","colón"),dic= es_ES)
```

```
## [1] FALSE  TRUE FALSE
```

```
hunspell_suggest(c("bieja","colon","colón"),dic=es_ES)
```

```
## [[1]]
## [1] "vieja" "biela"
##
## [[2]]
## [1] "colon"  "clono"  "colo"   "colona" "colono" "colan"  "colen"  "color"
##
## [[3]]
## [1] "colon" "clonó" "coló"  "colan" "colen"
```

```
palabras=c("amor", "amoroso", "amorosamente", "amado", "amante", "amador")
hunspell_analyze(palabras,dic=es_ES)
```

```
## [[1]]
## [1] " st:amor"      "a st:mor fl:a"
##
## [[2]]
## [1] "a st:moroso fl:a"
##
## [[3]]
## [1] "a st:morosamente fl:a"
##
## [[4]]
## [1] " st:amar fl:D"
##
## [[5]]
## [1] " st:amante"       " st:amantar fl:E"
##
## [[6]]
## [1] " st:amador"      "a st:mador fl:a"
```

de momento tomaremos sólo la primera sugerencia, aunque guardaremos todas.

```
list_sugerences= sapply(dic_raw_1$word, FUN=function(x) hunspell_suggest(x,dic=es_ES))



dic_raw_1$list_sugerence_first=sapply(list_sugerences, FUN=function(x) x[1])
dic_raw_1$list_sugerence_all=sapply(list_sugerences,
                                    FUN=function(x){
                                      if(length(x)>=1) {return(paste(x,collapse=","))}
                                      if(length(x)==0){return(NA)}
                                      })
glimpse(dic_raw_1)
```

```
## Rows: 3,948
## Columns: 4
## $ word                 <chr> "0", "1", "10", "100", "1000", "12", "120", "14",~
## $ N                    <int> 2, 9, 5, 4, 2, 2, 1, 1, 3, 1, 4, 3, 1, 2, 17, 1, ~
```

```
## $ list_sugerence_first <chr> "a", "a", NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ list_sugerence_all   <chr> "a,e,o,d,u,y", "a,e,o,d,u,y", NA, NA, NA, NA, NA,~
```

2 Primer modelo de curado de los chistes

```
knitr::kable(head(dic_raw_1,20))
```

word	N	list_sugerence_first	list_sugerence_all
0	2	a	a,e,o,d,u,y
1	9	a	a,e,o,d,u,y
10	5	NA	NA
100	4	NA	NA
1000	2	NA	NA
12	2	NA	NA
120	1	NA	NA
14	1	NA	NA
15	3	NA	NA
16	1	NA	NA
17	4	NA	NA
18	3	NA	NA
1acto:un	1	tractoran	tractoran
1er	2	ser	ser,ter,fer,ver,her,ere
2	17	a	a,e,o,d,u,y
2,10	1	NA	NA
20	5	NA	NA
2012	1	NA	NA
2013	1	NA	NA
21	1	NA	NA

2.0.1 Salvar en excel

```
write_excel_csv2(x=dic_raw_1,file="data/dic_raw_1_chistes.csv")
```

```
dic_raw_1_long_chistes= dic_raw_1 %>% right_join(text_raw,by="word")
write_excel_csv2(x=dic_raw_1_long_chistes,file="data/dic_raw_1_2_long_chistes.csv")
```

2.1 Siguiente paso tratamineto de los datos curados y generación de las Document Term Matrix

Primera aproximación generación dela DTM del corpus de peticiones curadas. Cruzar estos datos con los tópicos/key words de losa chistes. Podéis hacerlo con tidytext o con tm (o con quanteda).

2.2 Más chistes con metadatos

En el fichero de este git “chistes_con_metadatos.csv” hay más chistes con dos columnas de metadatos para practicar.