Minería de texto y procesamiento de lenguaje natural

Universidad internacional de la rioja

Rogerio Orlando Beltran Castro

Procesadores de lenguajes

Oscar David Bocanegra Capera

10/junio/2024

Primer Libro - The Raven by Edgar Allan Poe

Sacamos el corpus del libro

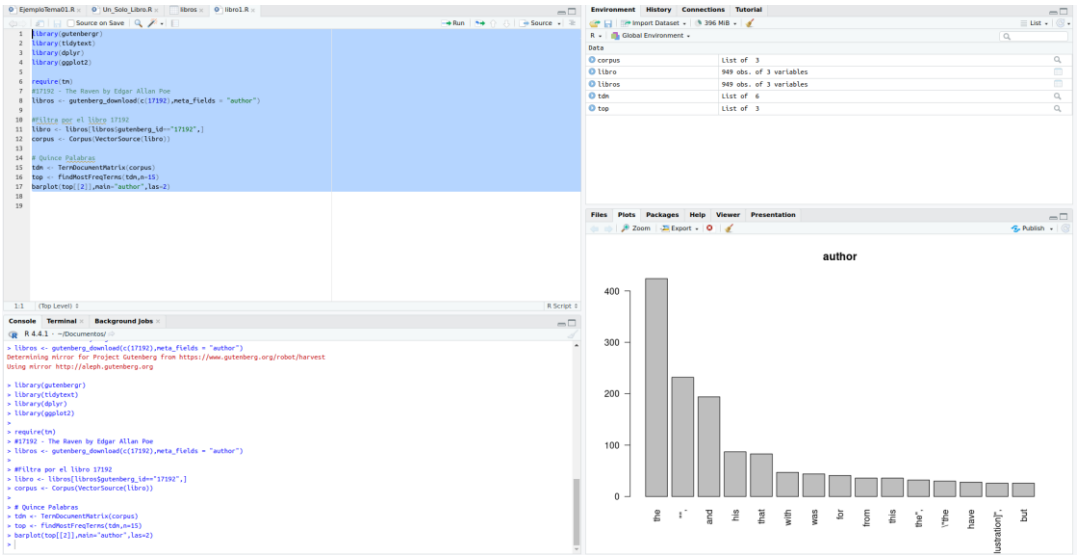
```
library(gutenbergr)
library(tidytext)
library(dplyr)
library(ggplot2)

require(tm)
#17192 - The Raven by Edgar Allan Poe
libros <- gutenberg_download(c(17192),meta_fields = "author")
```

Donde obtenemos el siguiente resultado

	gutenberg_id	text	author
1	17192	Transcriber's Notes:	Poe, Edgar Allan
2	17192		Poe, Edgar Allan
3	17192	In the List of Illustrations I restored a missing single ...	Poe, Edgar Allan
4	17192	"'Wretch,' I cried, 'thy God hath lent thee--by these...	Poe, Edgar Allan
5	17192	Respite--respite and nepenthe from thy memories ...	Poe, Edgar Allan
6	17192		Poe, Edgar Allan
7	17192	The List of Illustrations uses 'visitor' where the poe...	Poe, Edgar Allan
8	17192	illustration use 'visiter'.	Poe, Edgar Allan
9	17192		Poe, Edgar Allan
10	17192		Poe, Edgar Allan
11	17192	* * * * *	Poe, Edgar Allan
12	17192		Poe, Edgar Allan
13	17192		Poe, Edgar Allan
14	17192		Poe, Edgar Allan
15	17192		Poe, Edgar Allan
16	17192	THE RAVEN	Poe, Edgar Allan
17	17192		Poe, Edgar Allan
18	17192	BY	Poe, Edgar Allan
19	17192	EDGAR ALLAN POE	Poe, Edgar Allan
20	17192		Poe, Edgar Allan
21	17192	ILLUSTRATED	Poe, Edgar Allan
22	17192	BY GUSTAVE DORE	Poe, Edgar Allan
23	17192		Poe, Edgar Allan

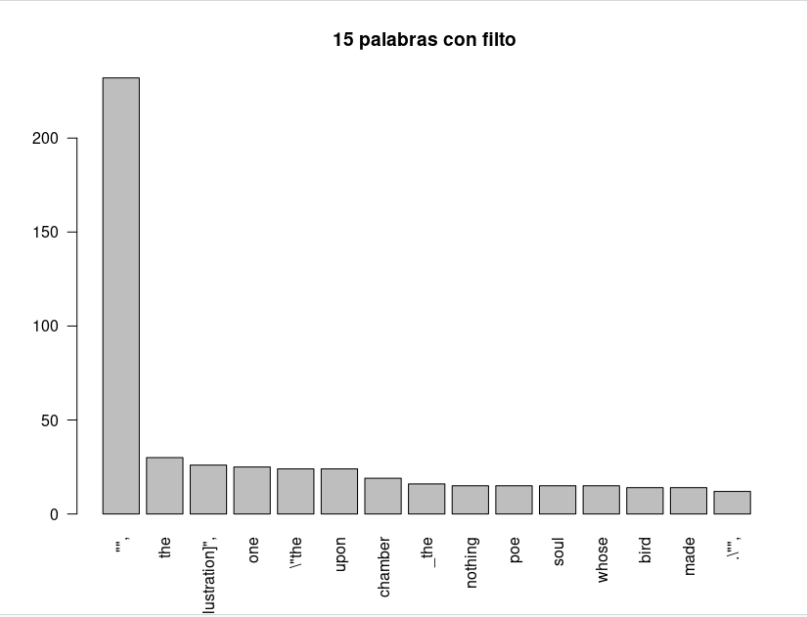
Showing 1 to 23 of 949 entries, 3 total columns

Luego de esto pasamos a presentar los quince términos más frecuentes y su nube de palabras



En donde podemos evidenciar una gráfica con las 15 palabras más repetidas, pero ahora lo que vamos a hacer es eliminar las palabras que no generan gran relevancia como lo pueden llegar a ser los conectores, para así de esta manera centrarnos un poco más en las palabras que están enfocadas en el libro

Y de esa manera logramos obtener la siguiente grafica



Ahora para hacer la nube de palabras ejecutamos estas líneas teniendo en cuanta la info anterior y obtenemos la siguiente imagen o nube

```
# generar esta nube de palabras.
require(wordcloud)
wordcloud(corpus_filtered)
```

Ahora realizamos el mismo procedimiento con el segundo libro el cual es Colossus of Chaos by Nelson S. Bond y obtenemos los siguientes resultados

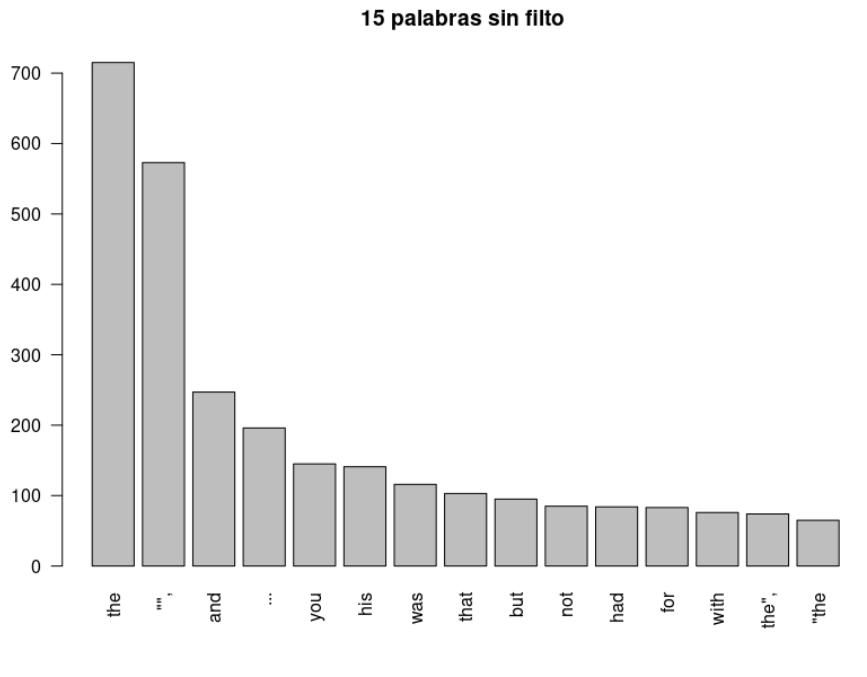
El corpus

Filter

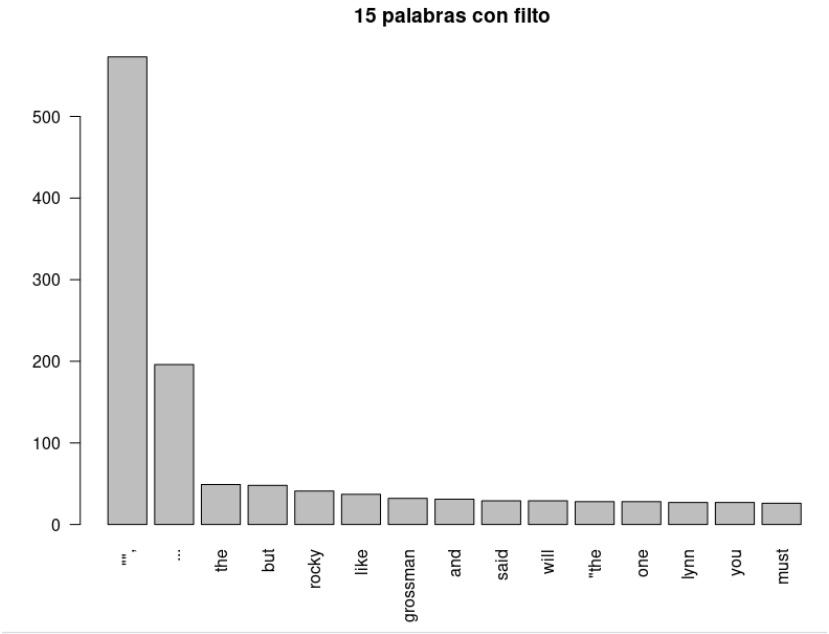
	gutenberg_id	text	author
1	62246	Colossus of Chaos	Bond, Nelson S.
2	62246		Bond, Nelson S.
3	62246	By NELSON S. BOND	Bond, Nelson S.
4	62246		Bond, Nelson S.
5	62246	IT was the evil spawn of lifeless space,	Bond, Nelson S.
6	62246	drifting aimlessly until ITs sinister	Bond, Nelson S.
7	62246	birthing place should come. And finding	Bond, Nelson S.
8	62246	that abode for life, IT grew, sucking	Bond, Nelson S.
9	62246	energy from Terra itself--gathering	Bond, Nelson S.
10	62246	strength for that time when all should	Bond, Nelson S.
11	62246	flee before ITs malign wrath.	Bond, Nelson S.
12	62246		Bond, Nelson S.
13	62246	[Transcriber's Note: This etext was produced ...	Bond, Nelson S.
14	62246	Planet Stories Winter 1942.	Bond, Nelson S.
15	62246	Extensive research did not uncover any evide...	Bond, Nelson S.
16	62246	the U.S. copyright on this publication was ren...	Bond, Nelson S.
17	62246		Bond, Nelson S.
18	62246		Bond, Nelson S.
19	62246	_Out of the darkness It came. Out of the grim, bleak...	Bond, Nelson S.
20	62246	incalculable depths of outer space, into the empire ...	Bond, Nelson S.
21	62246	warmth ... and life._	Bond, Nelson S.
22	62246		Bond, Nelson S.

Showing 1 to 22 of 2,263 entries, 3 total columns

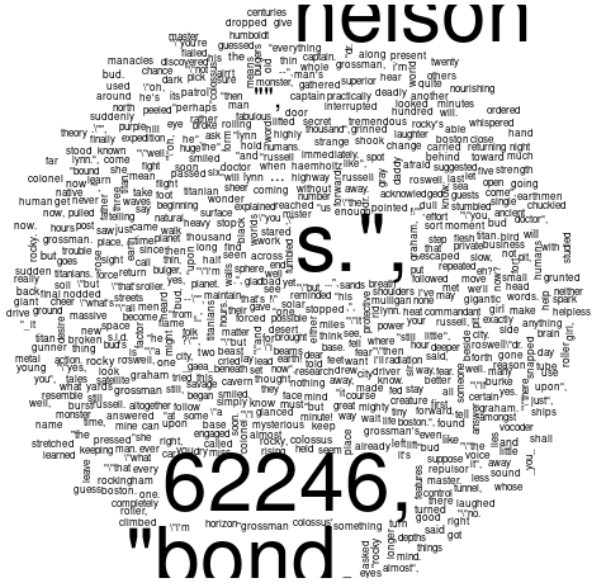
Las 15 palabras sin filtro



Las 15 palabras con filtro



Y la respectiva nube de palabras



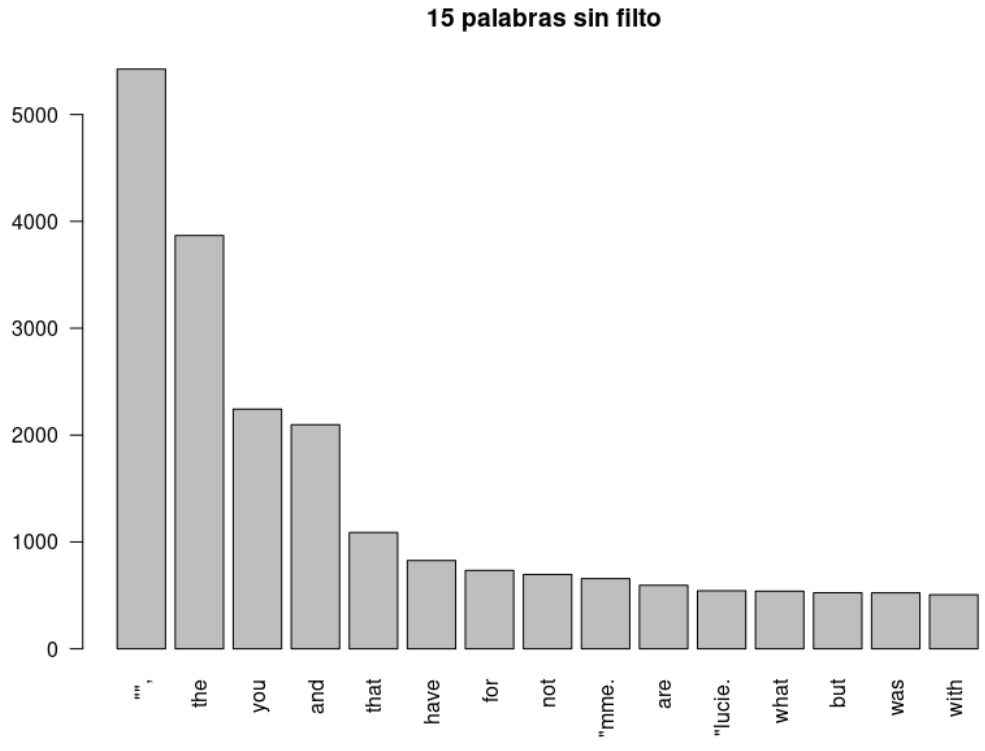
Ahora realizamos el mismo procedimiento con el 3er libro el cual va a ser Three Plays by Brieux by Eugène Brieux

En donde obtenemos el corpus

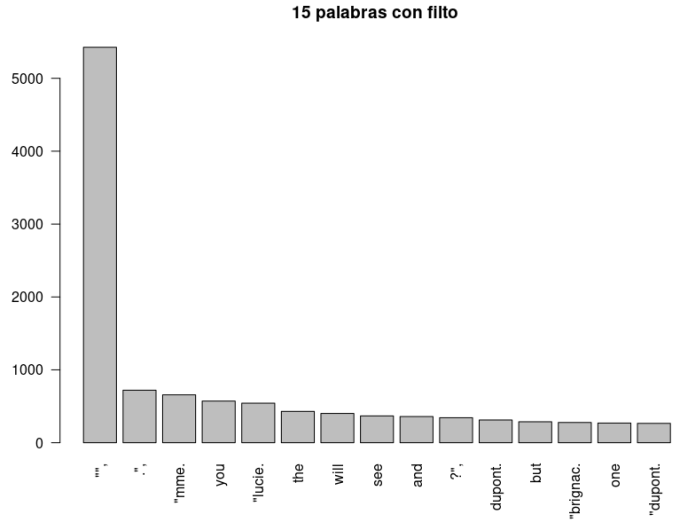
Filter

	gutenberg_id	text	author
1	46992	----------	Brieux, Eugène
2	46992	Transcriber's note:	Brieux, Eugène
3	46992		Brieux, Eugène
4	46992	Words in bold characters are enclosed within Plus (...	Brieux, Eugène
5	46992	----------	Brieux, Eugène
6	46992		Brieux, Eugène
7	46992		Brieux, Eugène
8	46992	Three Plays by Brieux	Brieux, Eugène
9	46992	Member of the French Academy	Brieux, Eugène
10	46992		Brieux, Eugène
11	46992	[Illustration: Brieux.]	Brieux, Eugène
12	46992		Brieux, Eugène
13	46992		Brieux, Eugène
14	46992		Brieux, Eugène
15	46992		Brieux, Eugène
16	46992	Three Plays by Brieux.	Brieux, Eugène
17	46992	With a Preface by Bernard	Brieux, Eugène
18	46992	Shaw. The English	Brieux, Eugène
19	46992	Versions by Mrs.	Brieux, Eugène
20	46992	Bernard Shaw, St. John	Brieux, Eugène
21	46992	Hankin and John Pollock.	Brieux, Eugène
22	46992		Brieux, Eugène

las 15 palabras sin filtro



15 palabras con filtro



Y la nube de palabras

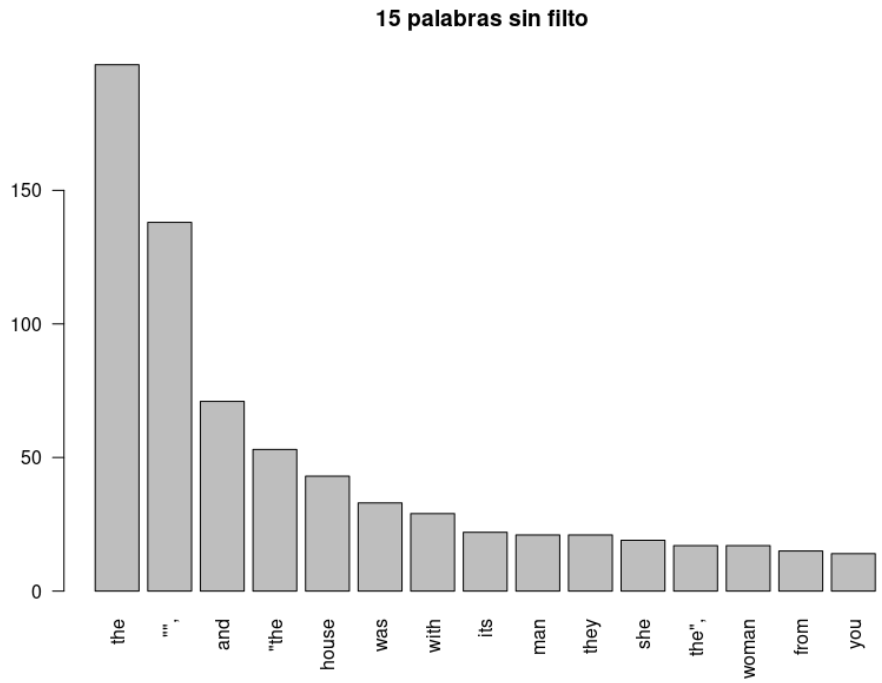
Y ahora repetimos el mismo proceso con el último libro el cual va a ser The angry house by Richard Rein Smith

En donde obtenemos el corpus

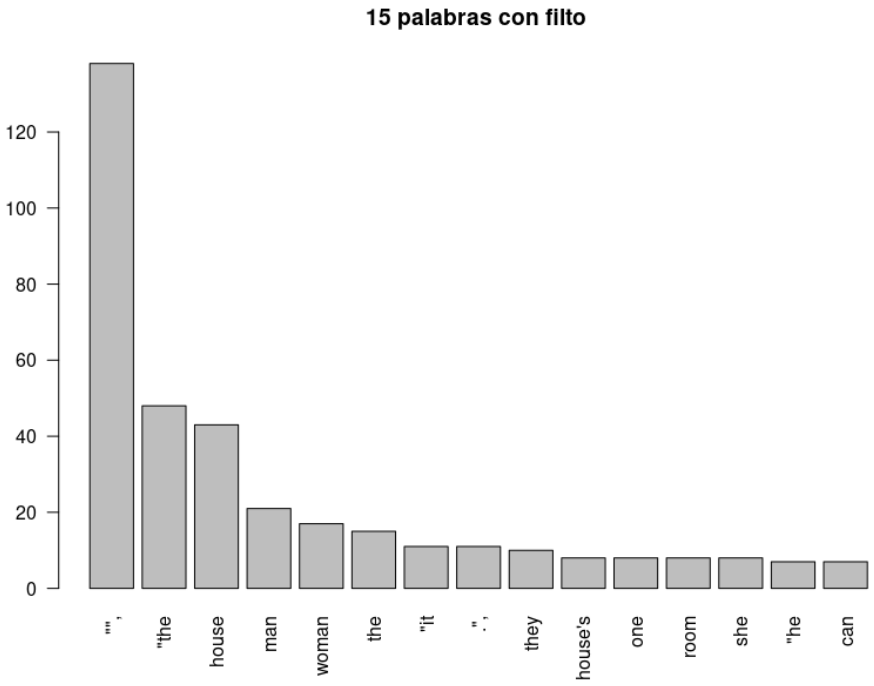
Filter

	gutenberg_id	text	author
1	69211	The ANGRY HOUSE	Smith, Richard Rein
2	69211		Smith, Richard Rein
3	69211	By RICHARD R. SMITH	Smith, Richard Rein
4	69211		Smith, Richard Rein
5	69211	[Transcriber's Note: This etext was produced ...	Smith, Richard Rein
6	69211	Startling Stories Summer 1955.	Smith, Richard Rein
7	69211	Extensive research did not uncover any evide...	Smith, Richard Rein
8	69211	the U.S. copyright on this publication was ren...	Smith, Richard Rein
9	69211		Smith, Richard Rein
10	69211		Smith, Richard Rein
11	69211	The house's electronic brain glowed with an intangi...	Smith, Richard Rein
12	69211	have been pride.	Smith, Richard Rein
13	69211		Smith, Richard Rein
14	69211	It thought, I am content. I am content because ther...	Smith, Richard Rein
15	69211	things I can do to make them happy. I can cook thei...	Smith, Richard Rein
16	69211	beds, scrub my floors, wash my windows. I can bath...	Smith, Richard Rein
17	69211	warm, give them a gentle, cool breeze. If they want ...	Smith, Richard Rein
18	69211	can rise hundreds of feet on my antigravity rays an...	Smith, Richard Rein
19	69211	view. I can give them soft music, entertaining TV pr...	Smith, Richard Rein
20	69211	surprises.	Smith, Richard Rein
21	69211		Smith, Richard Rein
22	69211	The house activated one of the many telescopic sca...	Smith, Richard Rein

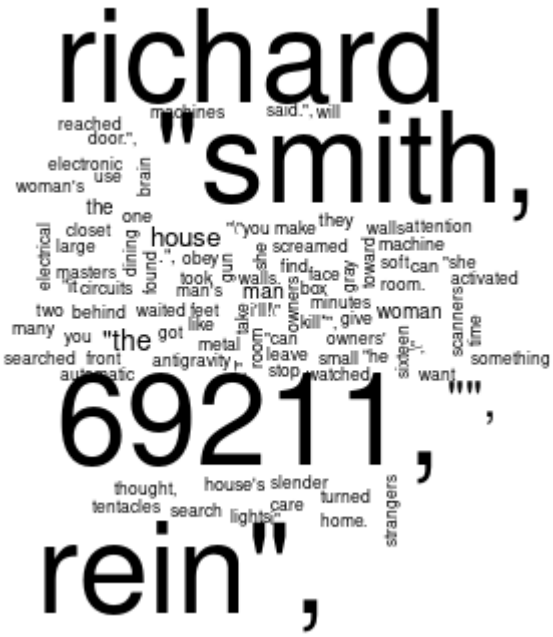
Las 15 palabras sin filtro



Las 15 palabras con filtro



Y la nube de palabras



Punto 3

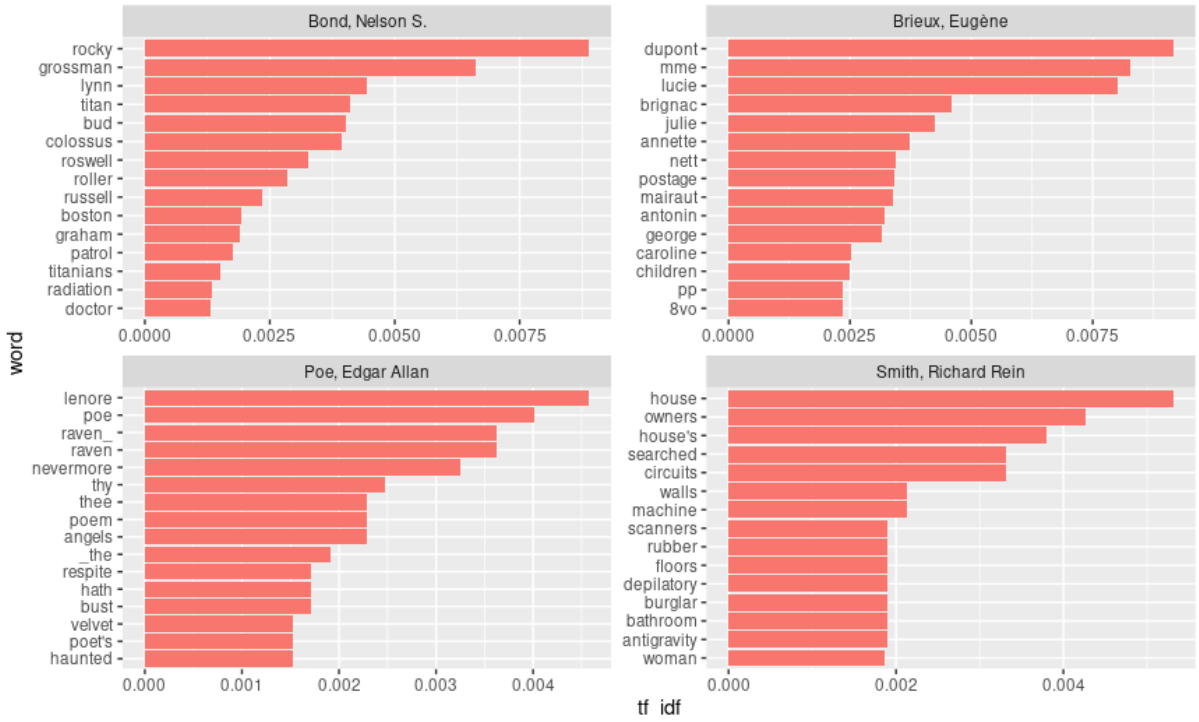
En este punto calculamos el índice TF-IDF de los cuatro libros en el cual obtenemos el siguiente resultado el cual hace referencia al cálculo para identificar términos que son

característicos de un documento en particular, lo que ayuda a resaltar palabras que son más importantes o representativas dentro del contexto del texto

	author	word	n	tf	idf	tf_idf
1	Brieux, Eugène	dupont	744	0.0066044100	1.3862944	0.0091556564
2	Bond, Nelson S.	rocky	106	0.0064168533	1.3862944	0.0088956476
3	Brieux, Eugène	mme	671	0.0059563967	1.3862944	0.0082573191
4	Brieux, Eugène	lucie	651	0.0057788588	1.3862944	0.0080111993
5	Bond, Nelson S.	grossman	79	0.0047823718	1.3862944	0.0066297751
6	Smith, Richard Rein	house	54	0.0184615385	0.2876821	0.0053110536
7	Brieux, Eugène	brignac	373	0.0033110819	1.3862944	0.0045901342
8	Poe, Edgar Allan	lenore	24	0.0033039648	1.3862944	0.0045802677
9	Bond, Nelson S.	lynn	53	0.0032084267	1.3862944	0.0044478238
10	Smith, Richard Rein	owners	9	0.0030769231	1.3862944	0.0042655211
11	Brieux, Eugène	julie	345	0.0030625288	1.3862944	0.0042455665
12	Bond, Nelson S.	titan	49	0.0029662813	1.3862944	0.0041121390
13	Bond, Nelson S.	bud	48	0.0029057449	1.3862944	0.0040282178
14	Poe, Edgar Allan	poe	21	0.0028909692	1.3862944	0.0040077342
15	Bond, Nelson S.	colossus	47	0.0028452085	1.3862944	0.0039442966
16	Smith, Richard Rein	house's	8	0.0027350427	1.3862944	0.0037915743
17	Brieux, Eugène	annette	304	0.0026985761	1.3862944	0.0037410209
18	Poe, Edgar Allan	raven	19	0.0026156388	1.3862944	0.0036260453
19	Poe, Edgar Allan	raven	19	0.0026156388	1.3862944	0.0036260453

Punto 4

Representar gráficamente los quince términos más característicos de cada libro en donde en la imagen podemos ver cuáles fueron los 15 temimos que destacaron en cada libro y cada uno tiene una gráfica la cual se puede interpretar como la intensidad de cada termino en dicho libro.



Punto 5

En este punto realizamos una comparación entre los términos más frecuentes y términos más característicos por cada uno de los libros en donde podemos evidenciar que la tabla “top_words” se encarga de contener las 15 palabras más características por autor, basadas en su TF-IDF calculado anteriormente

	author	word	n	tf	idf	tf_idf
1	Bond, Nelson S.	rocky	106	0.0064168533	1.3862944	0.008895648
2	Bond, Nelson S.	grossman	79	0.0047823718	1.3862944	0.006629775
3	Bond, Nelson S.	lynn	53	0.0032084267	1.3862944	0.004447824
4	Bond, Nelson S.	titan	49	0.0029662813	1.3862944	0.004112139
5	Bond, Nelson S.	bud	48	0.0029057449	1.3862944	0.004028218
6	Bond, Nelson S.	colossus	47	0.0028452085	1.3862944	0.003944297
7	Bond, Nelson S.	roswell	39	0.0023609177	1.3862944	0.003272927
8	Bond, Nelson S.	roller	34	0.0020582360	1.3862944	0.002853321
9	Bond, Nelson S.	russell	28	0.0016950179	1.3862944	0.002349794
10	Bond, Nelson S.	boston	23	0.0013923361	1.3862944	0.001930188
11	Bond, Nelson S.	graham	45	0.0027241358	0.6931472	0.001888227
12	Bond, Nelson S.	patrol	21	0.0012712634	1.3862944	0.001762345
13	Bond, Nelson S.	titanians	18	0.0010896543	1.3862944	0.001510582
14	Bond, Nelson S.	radiation	16	0.0009685816	1.3862944	0.001342739
15	Bond, Nelson S.	doctor	31	0.0018766269	0.6931472	0.001300779
16	Brieux, Eugène	dupont	744	0.0066044100	1.3862944	0.009155656
17	Brieux, Eugène	mme	671	0.0059563967	1.3862944	0.008257319
18	Brieux, Eugène	lucie	651	0.0057788588	1.3862944	0.008011199
19	Brieux, Eugène	brignac	373	0.0033110819	1.3862944	0.004590134

Punto 6

Y ahora miramos nuestra tabla top_words como lo definimos en el codigo la cual se encarga de obtener las asociaciones entre palabras la cual se encarga de encontrar qué palabras suelen aparecer juntas y con qué frecuencia

libro1.R × rated_words × top_words ×

Filter

	author	word	n	tf	idf	tf_idf
1	Bond, Nelson S.	rocky	106	0.0064168533	1.3862944	0.008895648
2	Bond, Nelson S.	grossman	79	0.0047823718	1.3862944	0.006629775
3	Bond, Nelson S.	lynn	53	0.0032084267	1.3862944	0.004447824
4	Bond, Nelson S.	titan	49	0.0029662813	1.3862944	0.004112139
5	Bond, Nelson S.	bud	48	0.0029057449	1.3862944	0.004028218
6	Bond, Nelson S.	colossus	47	0.0028452085	1.3862944	0.003944297
7	Bond, Nelson S.	roswell	39	0.0023609177	1.3862944	0.003272927
8	Bond, Nelson S.	roller	34	0.0020582360	1.3862944	0.002853321
9	Bond, Nelson S.	russell	28	0.0016950179	1.3862944	0.002349794
10	Bond, Nelson S.	boston	23	0.0013923361	1.3862944	0.001930188
11	Bond, Nelson S.	graham	45	0.0027241358	0.6931472	0.001888227
12	Bond, Nelson S.	patrol	21	0.0012712634	1.3862944	0.001762345
13	Bond, Nelson S.	titanians	18	0.0010896543	1.3862944	0.001510582
14	Bond, Nelson S.	radiation	16	0.0009685816	1.3862944	0.001342739
15	Bond, Nelson S.	doctor	31	0.0018766269	0.6931472	0.001300779
16	Brieux, Eugène	dupont	744	0.0066044100	1.3862944	0.009155656
17	Brieux, Eugène	mme	671	0.0059563967	1.3862944	0.008257319
18	Brieux, Eugène	lucie	651	0.0057788588	1.3862944	0.008011199
19	Brieux, Eugène	brignac	373	0.0033110819	1.3862944	0.004590134
20	Brieux, Eugène	julie	345	0.0030625288	1.3862944	0.004245566
21	Brieux, Eugène	annette	304	0.0026985761	1.3862944	0.003741021