Tarea 2 de Modelos Probabilistas Aplicados

Frecuencias y histogramas

5271

10 de enero de 2021

1. Origen de los datos

En este trabajo se utilizo como fuente de los datos el libro "The Adventures of Sherlock Holmes", del escritor y médico británico Arthur Conan Doyle. Este libro se encuentra disponible en la biblioteca virtual gratuita Project Gutenberg, al que se puede acceder desde el siguiente enlace: https://www.gutenberg.org.

2. Sobre el libro

Este libro es una colección de doce cuentos de Arthur Conan Doyle , los cuales fueron publicados por primera vez el 14 de octubre de 1892. El mismo agrupa los primeros cuentos con el detective consultor Sherlock Holmes , que se habían publicado en doce números mensuales de The Strand Magazine de Julio de 1891 a junio de 1892.

3. Análisis y tratamiento de los datos

Al libro seleccionado se le realiza un estudio de frecuencia de ocurrencia tanto de las letras como de las palabras que compone el texto. El análisis será realizado en el programa R versión 4.0.2 [1] en el entorno de desarrollo Rstudio [2].

3.1. Letras

De le libro en cuestión se extrajeron todas las letras que componen el texto, teniendo el mismo 432064 caracteres alfanuméricos que fueron almacenados en un *Data frame*, como se muestra en el cuadro 1 de la página 2. De los datos obtenidos solo necesitamos las letras por lo cual procedemos a un filtrado del *Data frame*. A partir de los datos filtrados se realiza un análisis de frecuencia de ocurrencia de las letras en el texto, como se muestra en el cuadro 2 de la página 2. Para una mejor comprensión de los datos obtenidos se realiza una representación gráfica de los mismos mediante un histograma, como se puede observar en la figura 1 de la página 3.

Cuadro 1: fragmento del *Data frame* que contiene todos los caracteres que aparecen en el texto.

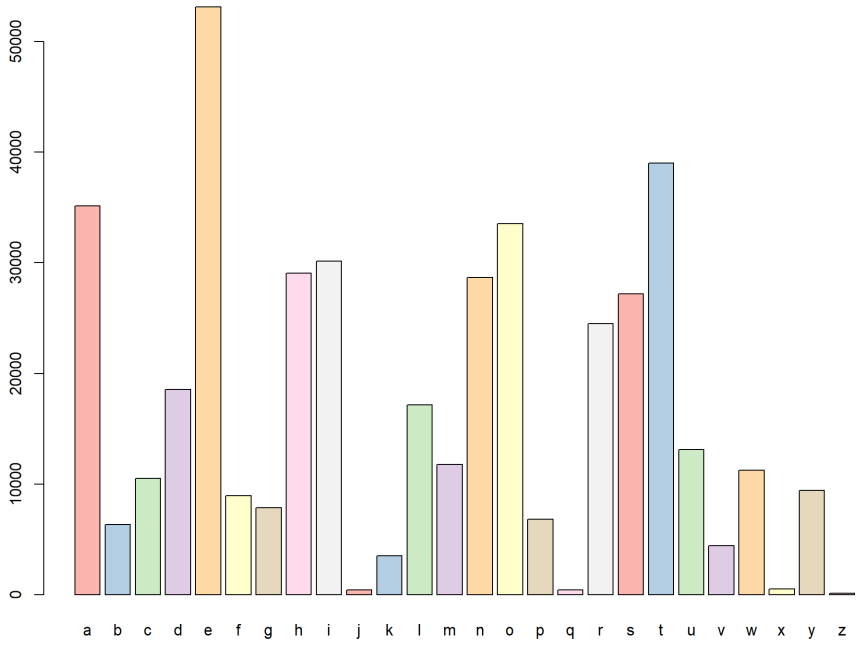
	gutenberg_id	letra
1	1661	t
2	1661	h
3	1661	e
4	1661	a
5	1661	d
6	1661	v
7	1661	e
8	1661	n
9	1661	t
10	1661	u

Cuadro 2: fragmento de *Data frame* que contiene la frecuencia de ocurrencia de cada letra en el texto

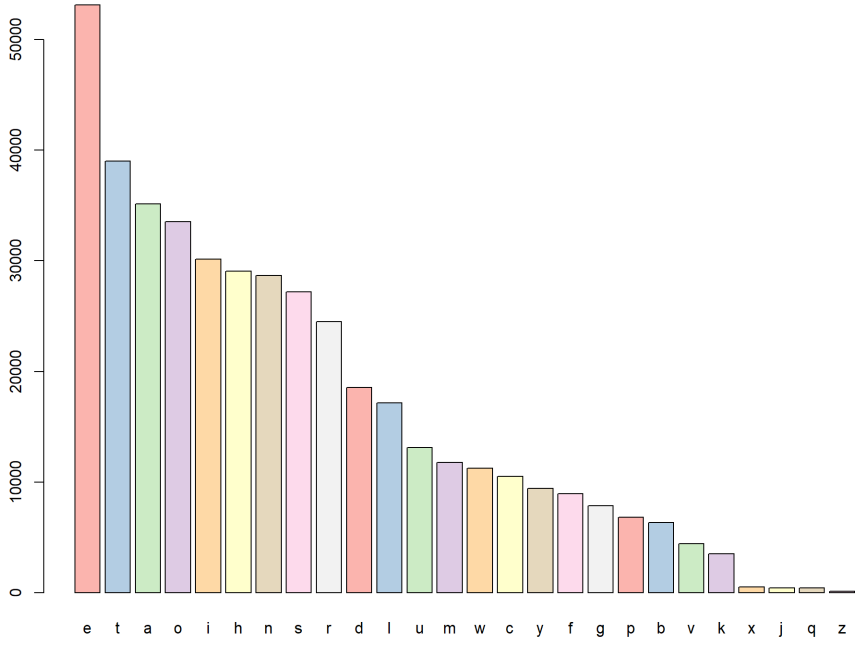
	Var1	Freq
11	a	35137
12	b	6362
13	c	10499
14	d	18563
15	e	53111
16	f	8975
17	g	7887
18	h	29047
19	i	30140
20	j	452
21	k	3543
22	l	17145
23	m	11787

```r
Conan = gutenberg_download(c(1661))

letras = Conan %>% unnest_tokens(letra, text, "characters")
write.csv(letras, "letras.csv")

palabras = Conan %>% unnest_tokens(palabra, text, "words")
write.csv(palabras,"palabras.csv")

oraciones = Conan %>% unnest_tokens(oraciones, text, "sentences")
write.csv(oraciones,"oraciones.csv")

ngrams= Conan %>% unnest_tokens(ngram, text,"ngrams", n = 2)



barplot(sort(table(palabras$palabra), decreasing=TRUE),log="y")

frcl = as.data.frame(table(letras$letra))
```

Tarea2.R



(a) Frecuencia de ocurrencia de las letras en el texto



(b) Frecuencia de ocurrencia de las letras en el texto ordenada de manera decreciente

Figura 1: Histogramas de frecuencia de ocurrencia de las letras en el texto

Cuadro 3: fragmento del *Data frame* que contiene todas las palabras que aparecen en el texto.

	Gutenberg_id	Palabra
1	1661	the
2	1661	adventures
3	1661	of
4	1661	sherlock
5	1661	holmes
6	1661	by
7	1661	sir
8	1661	arthur
9	1661	conan
10	1661	doyle

Cuadro 4: fragmento de *Data frame* que contiene la frecuencia de ocurrencia de cada palabra en el texto

	Var1	Freq
79	8s	1
80	9	1
81	90	1
82	9th	2
83	a	2641
84	abandoned	3
85	abandons	1
86	abbots	1
87	aberdeen	2

3.2. Palabras

Para el trabajo con las palabras, se extraen todas las presentes en el texto, teniendo el mismo 432064 palabras que fueron almacenadas en un *Data frame*, como se muestra en el cuadro 3 de la página 4. A partir de los datos obtenidos se realiza un análisis de frecuencia de ocurrencia de las palabras en el texto, como se muestra en el cuadro 4 de la página 4. Para una mejor comprensión de los datos obtenidos se realiza una representación gráfica de los mismos mediante un histograma, como se puede observar en la figura 2 de la página 5.

Como se puede observar en la figura 2, hay palabras que aparecen en pocas ocasiones en el texto y otras que se repiten gran cantidad de veces, estos extremos nos impiden hacer un análisis mas a fondo del contenido del texto, por lo que se realiza un filtrado de las palabras por frecuencia de ocurrencia tomando solamente aquellas que aparecen más de 200 y menos de 450 veces, el resultado de este filtro se muestra en el cuadro 5 de la página 5 y gráficamente en la figura 3 de la página 6 y en la nube de palabras que se muestra en la figura 4 de la página 7.

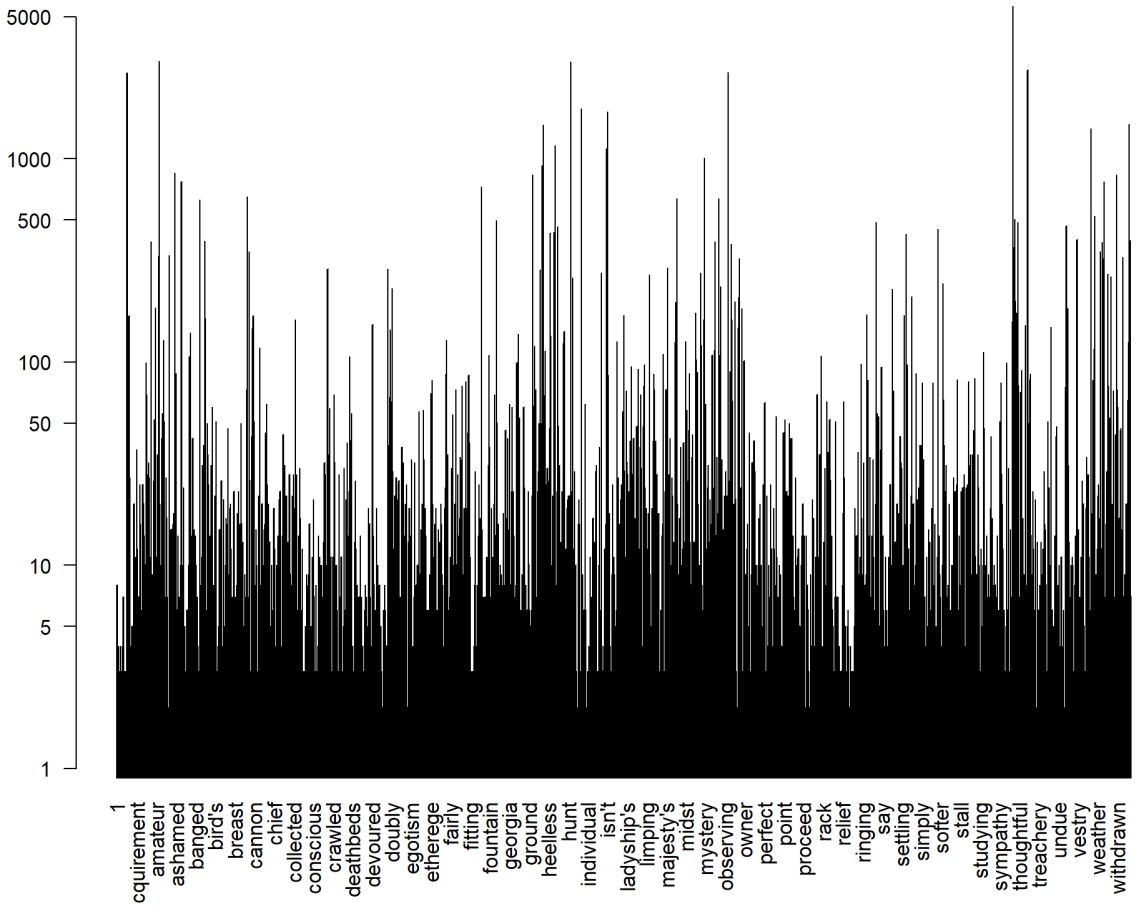
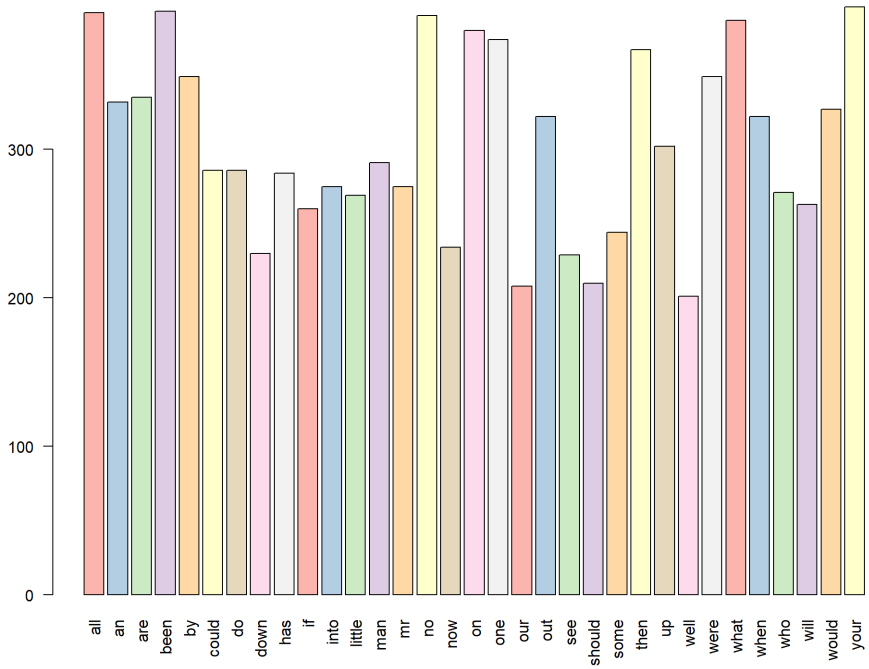


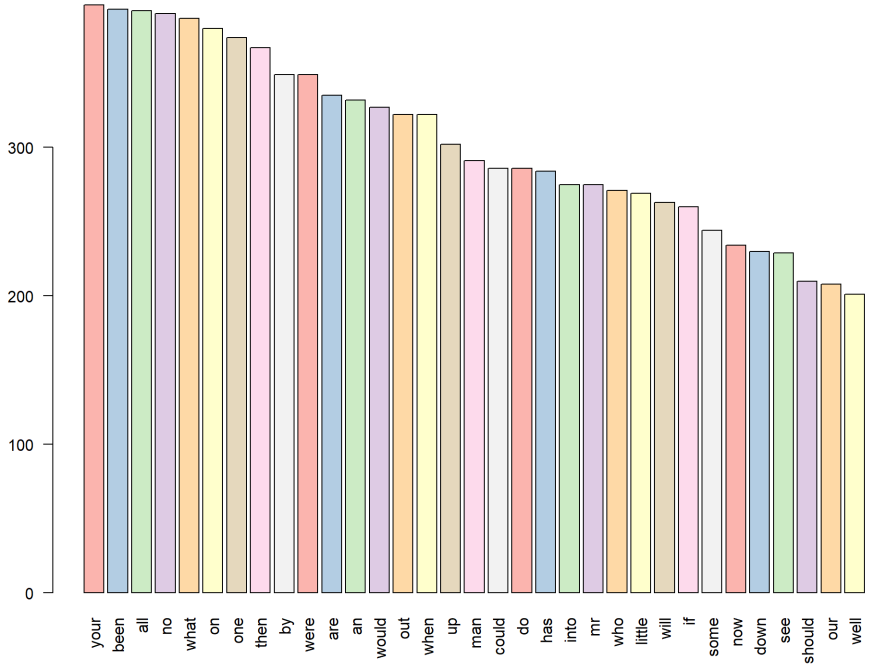
Figura 2: Histogramas de frecuencia de ocurrencia de las palabras en el texto en el texto, usando la escala logarítmica para mejor comprensión del gráfico

Cuadro 5: fragmento de *Data frame* frecuencia de ocurrencia de las palabras en el texto en un rango de 200 – 450 veces

	Var1	**Freq**
274	all	392
333	an	332
417	are	335
697	been	393
1046	by	349
1667	could	286
2143	do	286
2177	down	230
3350	has	284



(a) Frecuencia de ocurrencia de las palabras en el texto en un rango de 200–450 veces



(b) Frecuencia de ocurrencia de las palabras en el texto en un rango 200 – 450 veces, ordenada de manera decreciente

Figura 3: Frecuencia de ocurrencia de las palabras en el texto en un rango 200 – 450 veces

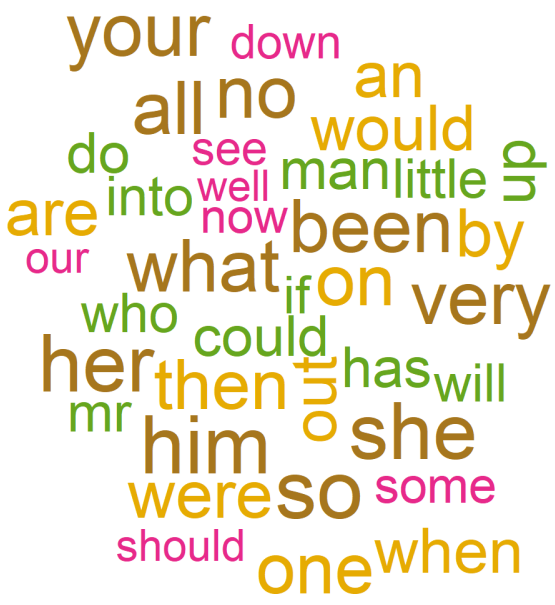


Figura 4: Nube de palabras utilizadas en el texto en un rango 200 – 450

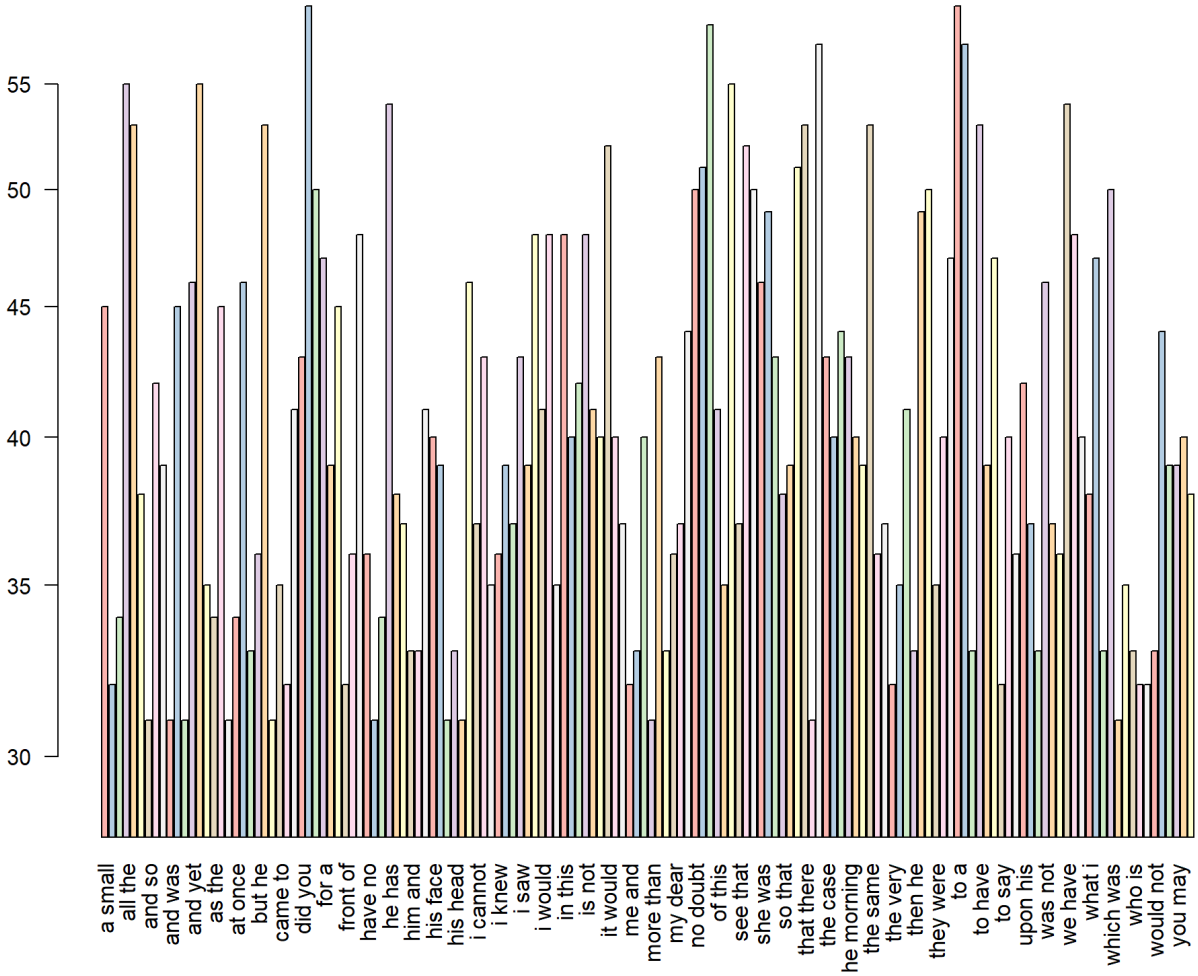


Figura 5: histograma de pares de palabras usadas en el texto en un rango de 30 – 60 veces

Además se analizan los pares de palabras que aparecen juntas en el texto, como se muestra en la figura 3 de la página 6

El código general se encuentra disponible en el repositorio https://github.com/Albertomnoa/Tareas_MPA/Tarea2

Referencias

[1] R Core Team. R: R: Un lenguaje y un entorno para la informática estadística, 2020.

[2] RStudio Team. Rstudio: Entorno de desarrollo integrado para r, 2020.