Análisis de Frecuencias de Letras en Libros para poder realizar el diagrama de Huffman

Dylan Canning

Contents

Introducción

En este documento explico como he extraido y analizado las frecuencias de letras de un conjunto de libros en español. Los libros los he descargado de Project Gutenberg utilizando el gutenberg-bulk-downloader.

Al haberlo descargado del Project Gutenberg no tienen Copy-Right y se pueden utilizar de manera libre.

Función para leer archivos de forma segura

La siguiente función `read_file` se utiliza para leer el contenido de un archivo de texto. Esta función lee el archivo en codificación "iso-8859-1" (el default de la biblioteca gutenberg.

```
read_file <- function(file_path) {
  result <- tryCatch({
    data <- readr::read_file(file_path, locale = readr::locale(encoding = "iso-8859-1"))
    return(tibble(content = data, file = basename(file_path)))
  }, warning = function(warning_msg) {
    return(NULL)
  }, error = function(error_msg) {
    return(NULL)
  })
  return(result)
}
```

Leyendo todos los archivos de texto del directorio

A continuación, defino el directorio donde se encuentran almacenados los libros y leeo cada uno de ellos con la función de antes.

```r
directory <- "/Users/dylan/Desktop/ProyectoRedes/gutenberg-bulk-downloader/_storage/"
file_paths <- list.files(directory, pattern = "*.txt", full.names = TRUE)
all_books <- map_dfr(file_paths, read_file)
```

Preparación y limpieza de los datos de texto

Antes de calcular las frecuencias de las letras, hay que limpiar los datos. Elimino todos los caracteres que no son letras y convierto todo el texto a mayúsculas. (Me da error con las minisculas).

```r
cleaned_books <- all_books %>%
  mutate(cleaned_content = str_remove_all(content, "[^A-Za-z]")) %>%  # Remove everything except letter
  mutate(cleaned_content = str_to_upper(cleaned_content))  # Convert to uppercase
```

Cálculo de frecuencias de letras

Una vez el texto esta limpio, procedo a calcular la frecuencia de cada letra. Posteriormente saco el porcentaje en proporcion al total de letras de todos los libros leidos (100).

```r
letter_counts <- cleaned_books %>%
  unnest(Letra = tokenizers::tokenize_characters(cleaned_content, strip_non_alphanum = TRUE, simplify =
  count(Letra) %>%
  filter(str_length(Letra) == 1)
```

```
## Warning: `unnest()` has a new interface. See `?unnest` for details.
## i Try `df %>% unnest(c(Letra))`, with `mutate()` if needed.
```

```r
total_letters <- sum(letter_counts$n)
letter_counts <- letter_counts %>%
  mutate(Porcentaje = n / total_letters * 100)
kable(letter_counts, caption = "Frecuencias de Letras en Libros", format = "latex")
```

Table 1: Frecuencias de Letras en Libros

Letra	n	Porcentaje
a	3226766	11.7868890
b	435672	1.5914440
c	1098834	4.0138747
d	1335833	4.8795963
e	3573655	13.0540221
f	259643	0.9484367
g	358332	1.3089327
h	372637	1.3611867
i	1682594	6.1462618
j	128901	0.4708559
k	26475	0.0967092
l	1516538	5.5396843
m	779942	2.8490104
n	1914540	6.9935255
o	2487623	9.0869112
p	693079	2.5317129
q	280329	1.0239995
r	1817473	6.6389544
s	2104553	7.6876143
t	1336437	4.8818026
u	1131922	4.1347402
v	297812	1.0878623
w	57884	0.2114415
x	44138	0.1612294
y	315622	1.1529195
z	98658	0.3603828