

Frecuencias

Alberto Benavides
14 de septiembre de 2020

1. INTRODUCCIÓN

Un análisis estadístico básico de un libro consiste en analizar la frecuencia de las palabras con que está escrito. Una guía bastante completa para lograrlo en R [5] puede consultarse en <https://tinyurl.com/y64jld27>.

2. RECURSOS

El libro de Dracula [4] se encuentra disponible de manera gratuita en el sitio de Project Gutenberg que alberga libros cuya propiedad intelectual ha caducado y, por lo tanto, se trata de obras que forman parte del dominio público. Se puede acceder a los textos planos de estas obras mediante la librería `gutenbergr` [2] de R y la descomposición en palabras de la obra se facilita con el uso de `tidytext` [3] que, a su vez, utiliza para su funcionamiento la librería `dplyr` [6]. Existe un cuaderno de Jupyter [1] donde se trabajaron los datos disponible en <https://tinyurl.com/y3dp3ngl>.

3. PREPROCESAMIENTO

Después de obtenido el libro mediante la función `gutenberg_download`, se analiza su contenido inicial y final con las funciones `head` y `tail`. Esto muestra la existencia de apartados introductorios y anexos que se escapan del contenido del libro en sí, por lo que se procede a remover del contenido esos elementos. Destaca aquí el uso de la función `grep` que permite identificar expresiones regulares y, en este caso, la ubicación del final del libro que coincide con las palabras “THE END”.

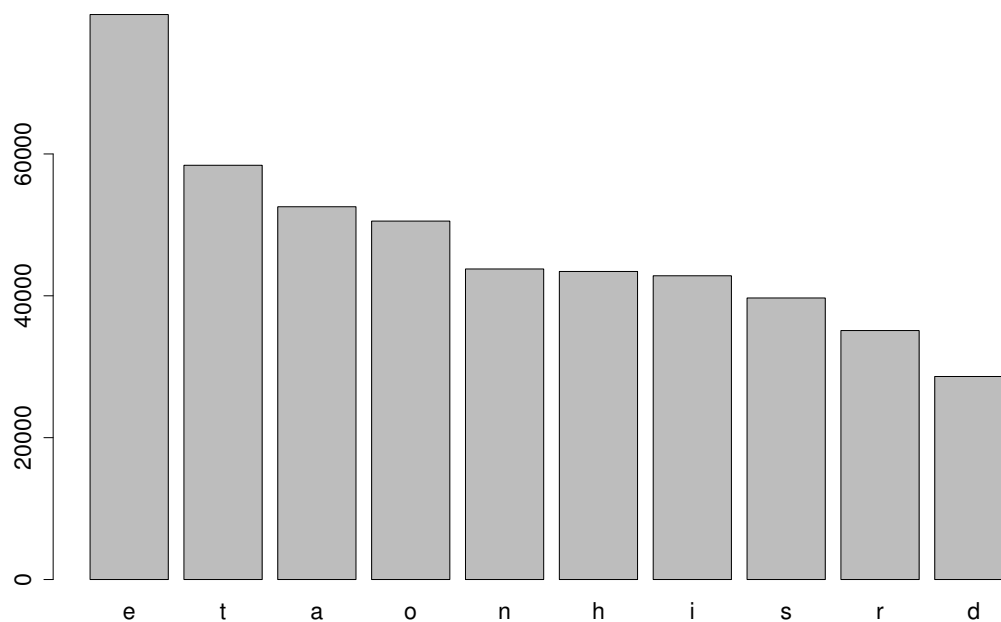


Figura 4.1: Diagrama de barras de las frecuencias de aparición de los caracteres más usados en el libro de *Dracula*.

4. RESULTADOS Y ANÁLISIS

Una vez identificado y aislado el contenido del libro, se extraen las letras y palabras que contiene con la función `unnest_tokens` de la librería `tidytext`. Luego se genera una tabla de frecuencias en orden descendiente por conteo de aparición de las diez letras y palabras más usadas mediante la función `order`. Dichas tablas se grafican en los diagramas de barras de las figuras 4.1 (p. 2) y 4.2 (p. 3).

En cuanto al contexto del libro, esta información no es muy relevante puesto que las palabras que aparecen se denominan palabras vacías (o stop words en inglés) que se refiere a palabras que en un idioma carecen de significado fuera de contexto (más información en <https://tinyurl.com/y2pr7kcb>). En español se incluyen todas las preposiciones, artículos y adverbios. Al eliminar estas palabras, se obtiene una lista de palabras más significativas cuya representación gráfica en una nube de palabras se puede consultar en la figura 4.3 (p. 4).

Por último, se hizo una animación de la frecuencia acumulada de las cincuenta palabras más significativas que aparecen en la obra a partir del porcentaje de avance de la misma, de uno en uno por ciento. Esta animación puede verse en <https://tinyurl.com/y2dnxrv8>.

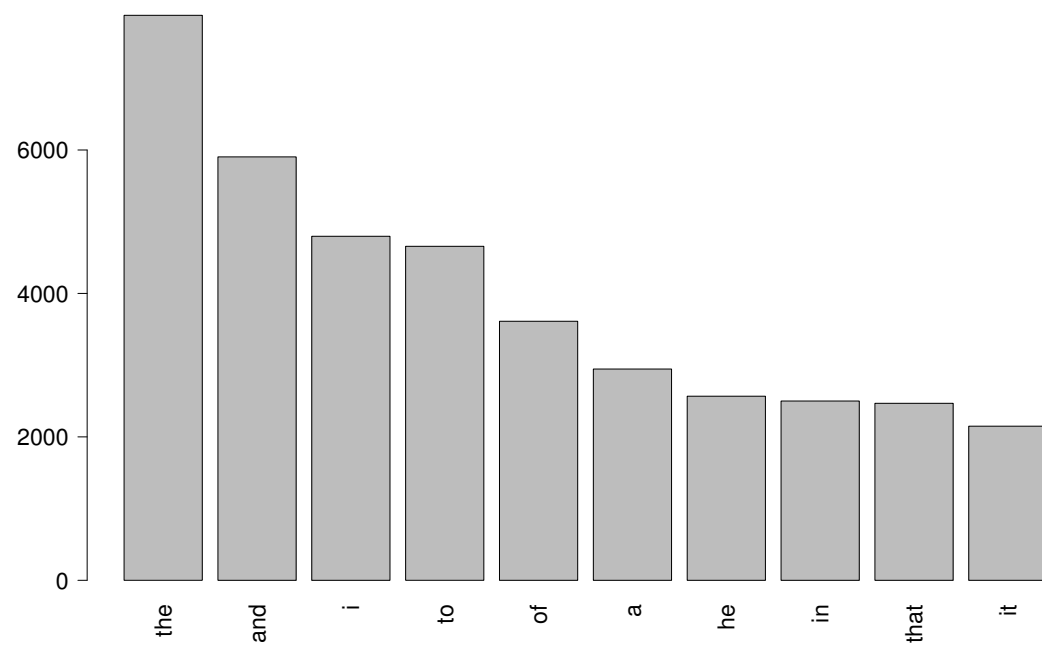


Figura 4.2: Diagrama de barras de las frecuencias de aparición de las palabras más usadas en el libro de *Dracula*.



REFERENCIAS

- [1] Jupyter. Project Jupyter. <https://jupyter.org/>, 2020.
- [2] David Robinson. gutenbergr. <https://www.rdocumentation.org/packages/gutenbergr/versions/0.1.5>, 2019.
- [3] Julia Silge. tidytext. <https://www.rdocumentation.org/packages/tidytext/versions/0.2.5>, 2020.
- [4] Bram Stoker. Dracula by Bram Stoker. <https://www.gutenberg.org/ebooks/345>, 2020.
- [5] The R Foundation. The R Project for Statistical Computing. <https://www.r-project.org/>, 2020.
- [6] Hadley Wickham. dplyr. <https://www.rdocumentation.org/packages/dplyr/versions/0.7.8>, 2018.