

La frecuencia de uso de palabras en el libro: Precepts in Practice; or, Stories Illustrating the Proverbs by A. L. O. E.

TAREA 2

Alumno:
Joaquín Arturo Velarde Moreno

Cambios

1. Summary con textit,
2. Se agrega referencia de la figura 2.3,
3. se añadio referencia al repositorio,
4. se corrige ortografía.

1. Introducción:

En el presente trabajo se hizo el análisis de la frecuencia en el uso de una selección de palabras y letras contenidas en el libro de Precepts in Practice; or, Stories Illustrating the Proverbs del autor A. L. O. [1] Este libro se encuentra en la biblioteca digital de Gutenberg [2], el cual es el repositorio de más de 63,127 títulos en formato ebook.

Esta obra que revisamos está escrita en inglés y se conforma por un conjunto de relatos cuyas ilustraciones tienen el objetivo de complementar visualmente cada proverbio y, de acuerdo con este género literario, cada historia conlleva una enseñanza de la que podemos aprender en nuestra vida cotidiana, según las diferentes situaciones y asuntos de que se trate, este documento se encuentra alojado en el repositorio [4] como recurso libre.

2. Metodología:

Para cumplir con nuestro objetivo, primeramente, procedimos a descargar el texto con el programa R- 4.0.2 [5], dado que la biblioteca Gutenberg [2] permite tal descarga libremente. Como siguiente paso, se removió todo carácter especial no alfa numérico, es decir, los símbolos tipográficos (guiones, cursivas, negritas, etc.) para limpiar el texto. Debido a que la frecuencia de datos era dominada principalmente por artículos, pronombres personales, conjunciones y números, se hizo una hoja de datos en Microsoft Excel [3] con este tipo de palabras. Posteriormente, se le ordenó al programa limpiar el texto de dichos caracteres. Con el resto del léxico, se pudo calcular la frecuencia de dichas palabras en dos áreas:

1. Análisis de letras ,
2. Análisis de palabras .

2.1. Análisis de grafías:

Para efectuar el análisis de grafías, se creó un diagrama de cajas y bigotes que representan la frecuencia de aparición de las letras **en el texto** (Figura 2.1), obteniendo los siguientes datos: mínimo, máximo, media, primer cuartil y tercer cuartil utilizando, como se señaló antes, el programa R 4.0.2 [5].

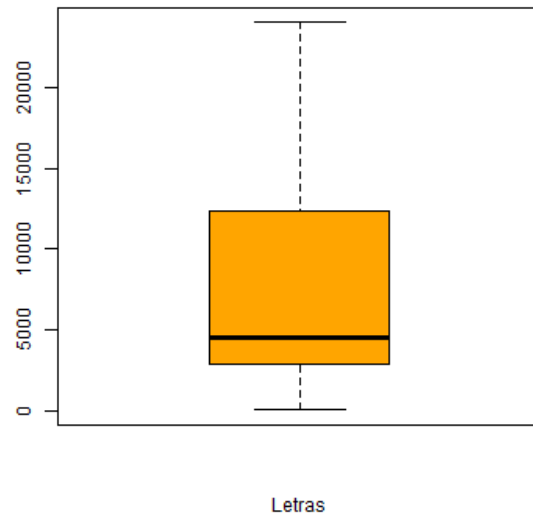


Figura 2.1: Frecuencia de letras representado en diagrama de cajas

Se decidió, para mayor claridad, dar un enfoque al conjunto que inicia desde el segundo cuartil y cuarto cuartil obtenidas con la función *summary()* para, de este modo, concentrar las frecuencias en un rango más cercano. Con esta información se elaboró una figura de barras en la que se muestra la frecuencia de las letras, descubriendo un predominio de una serie de consonantes como son: s, r, d, l, u, w, m, f, c, g, y, p ordenadas de mayor a menor frecuencia **(Figura 2.2)**.

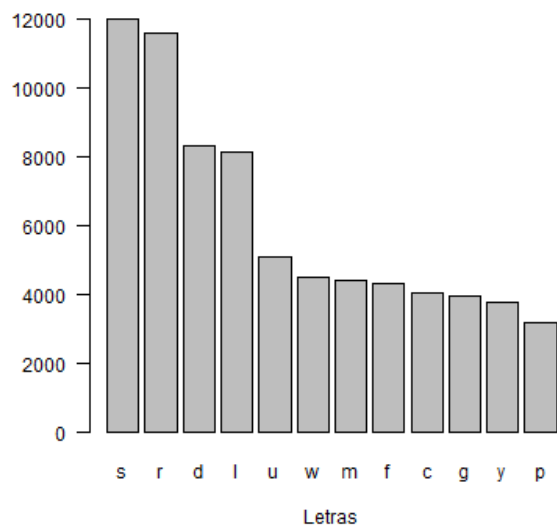


Figura 2.2: Frecuencia de letras representado en diagrama

2.2. Análisis de palabras:

Con la utilización del programa mencionado, también pudimos obtener información acerca de la frecuencia de las palabras y su cantidad contenidas del segundo quantil hasta el cuarto quantil del resumen dado por la function `summary()` para así concentrar el léxico más frecuente (Figura 2.3).

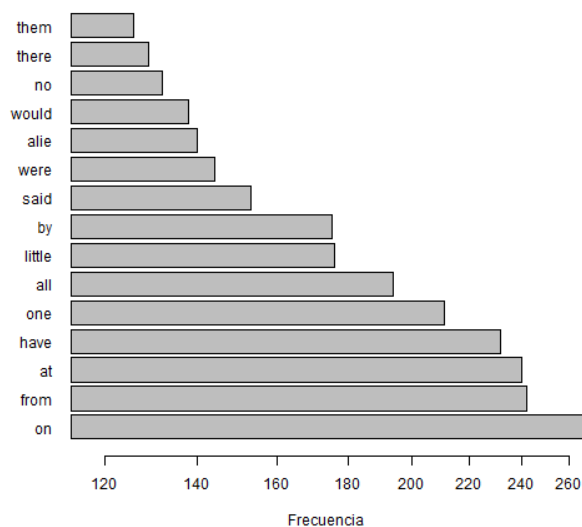


Figura 2.3: Frecuencia de Palabras representado en diagrama

Referencias

- [1] A. L. O. E. *Precepts in Practice; or, Stories Illustrating the Proverbs*. <https://www.gutenberg.org/files/58791/58791-0.txt>. 2019.
- [2] Michael Hart. *Project Gutenberg*. <https://www.gutenberg.org/>. 1971.
- [3] Microsoft Office. *Microsoft Excel*. <https://www.microsoft.com/es-mx/microsoft-365/excel>. 2011.
- [4] Joaquin Arturo Velarde Moreno. *Repositorio con material de la clase de probabilidad*. Recursos libre, disponible en https://github.com/joaquin3600/Modelos_Probabilistas_Aplicados. 2020.
- [5] The R Foundation. *The R Project for Statistical Computing*. <https://www.r-project.org/>. 2019.