Tarea 2
Análisis de texto

Erick Cervantes Mendieta
Matrícula: 2032430

Modelos Probabilistas Aplicados

Septiembre 2020

1. Presentación de los datos

Para esta tarea se analizó el libro titulado "The Singing Mouse Stories", el cual se encuentra disponible de manera gratuita en el sitio Web: *Project Gutenberg* [1]. Se utilizaron algunas técnicas para análisis de texto, como lo es las frecuencias de las letras y de las palabras, ideal para identificar rápidamente los temas comunes, y por otra parte, el agrupamiento de palabras, ya que a veces un grupo de palabras puede proporcionar una mejor perspectiva que una sola palabra.

En el cuadro 1 se pueden observar las cinco letras más utilizadas en el libro y en el cuadro 2, se pueden observar las palabras más frecuentes en el texto.

2. Análisis de los datos

Tanto las palabras como las letras utilizadas en la redacción del libro fueron analizadas, en el sentido de que se pudieron ordenaran según la frecuencia con la que aparecieron, por otra parte, los dígitos como las preposiciones, los artículos y las conjunciones fueron excluidos(as) del análisis, ya que estos no aportaban gran información.

Cuadro 1: Letras más utilizadas en el libro

Letra	Frecuencia
e	10395
t	7894
a	6741
o	6306
n	5857

Cuadro 2: Palabras más utilizadas en el libro

Letra	Frecuencia
the	1775
and	895
of	731
in	367
it	350

La figura 1 muestra el número de veces que fueron utilizadas las letras del abecedario en el texto en orden descendente, en donde se puede observar que la letra más utilizada fue la **e** y la menos utilizada fue la **x**.

En la figura 2 se pueden visualizar las palabras cuya frecuencia es mayor a cien, es decir, estas palabras aparecieron más de cien veces en el texto del libro, el listado es pequeño debido al tamaño del texto. Se observa que la palabra más utilizada fue el artículo **the**, seguida de la conjunción **and**, y en tercer lugar la preposición **of**, estas palabras son conocidas como *palabras huecas* [2], ya que aportan poca información semántica. Afortunadamente el programa R (Versión 4.0.2) [3] nos ofrece una herramienta para eliminar este tipo de palabras, lo que me permitió poder obtener bigramas para poder extraer información relevante del texto.

La figura 3 expone una red semántica con los bigrama obtenidos, los cuales no contienen palabras huecas. Esta red muestra la intensidad con la que se relacionan las palabras, es decir, la frecuencia con la que parejas de palabras aparecen en el texto. De dicha figura, podemos observar que aparece la palabra *singing mouse*, lo que no da la referencia al personaje principal (debido al título también), otro conjunto de palabras que me llamó la atención fue *lake belle marie*, por lo que supongo la historia se desarrolla cerca de un lago. Así mismo, *little river* y *delectable mountains* nos hacen sospechar, que la historia se desarrolla cerca de un paisaje lleno de naturaleza.

Referencias

[1] Emerson Hough. The Singing Mouse Stories. https://www.gutenberg.org/ebooks/21004, 2007.

[2] Juan Mendoza. Redes Semánticas con R, 2018. [Web; accedido el 08-09-2020].

[3] R Core Team. R: A Language and Environment for Statistical Computing. https://www.R-project.org/, 2020.

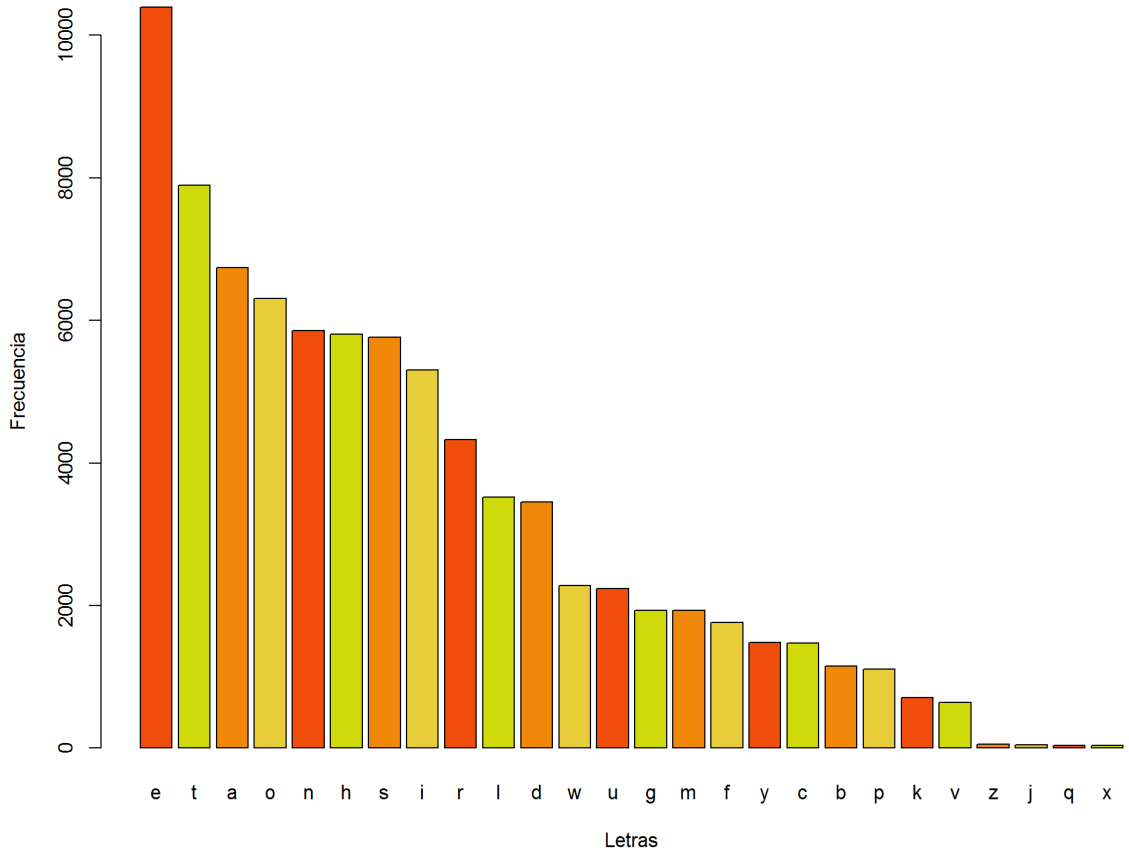

Figura 1: Frecuencia de las letras utilizadas en el texto

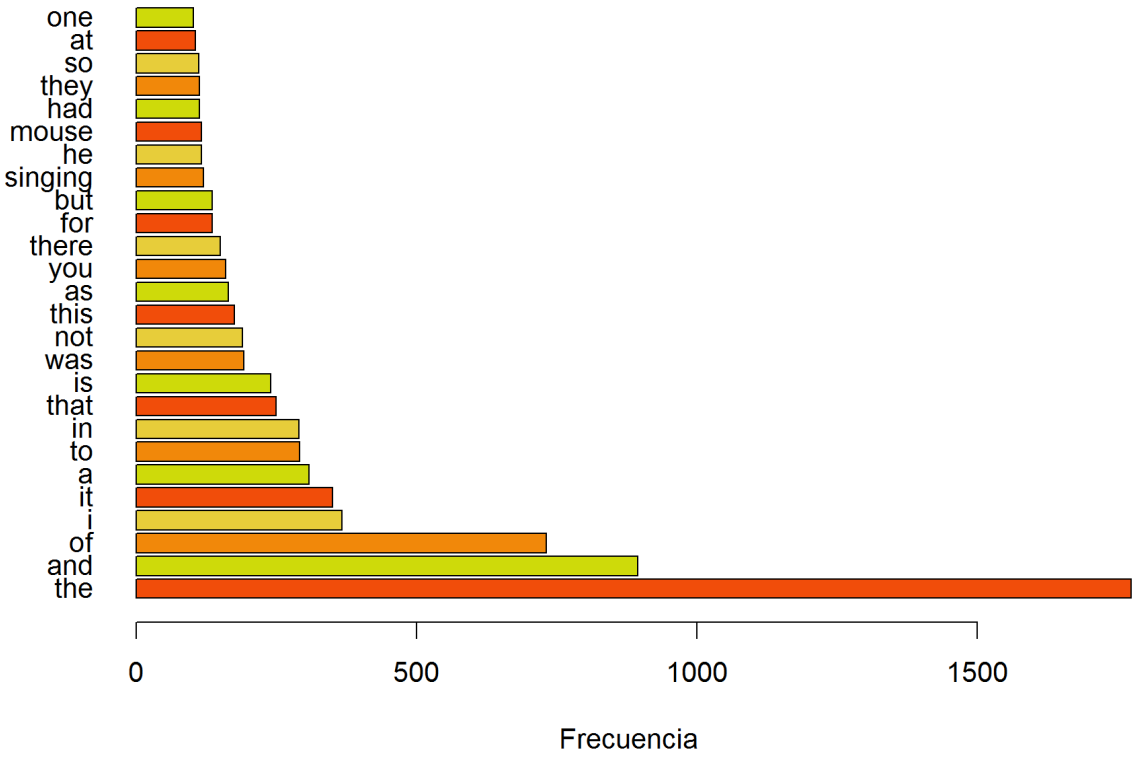

Figura 2: Frecuencia de las palabras más utilizadas en el texto

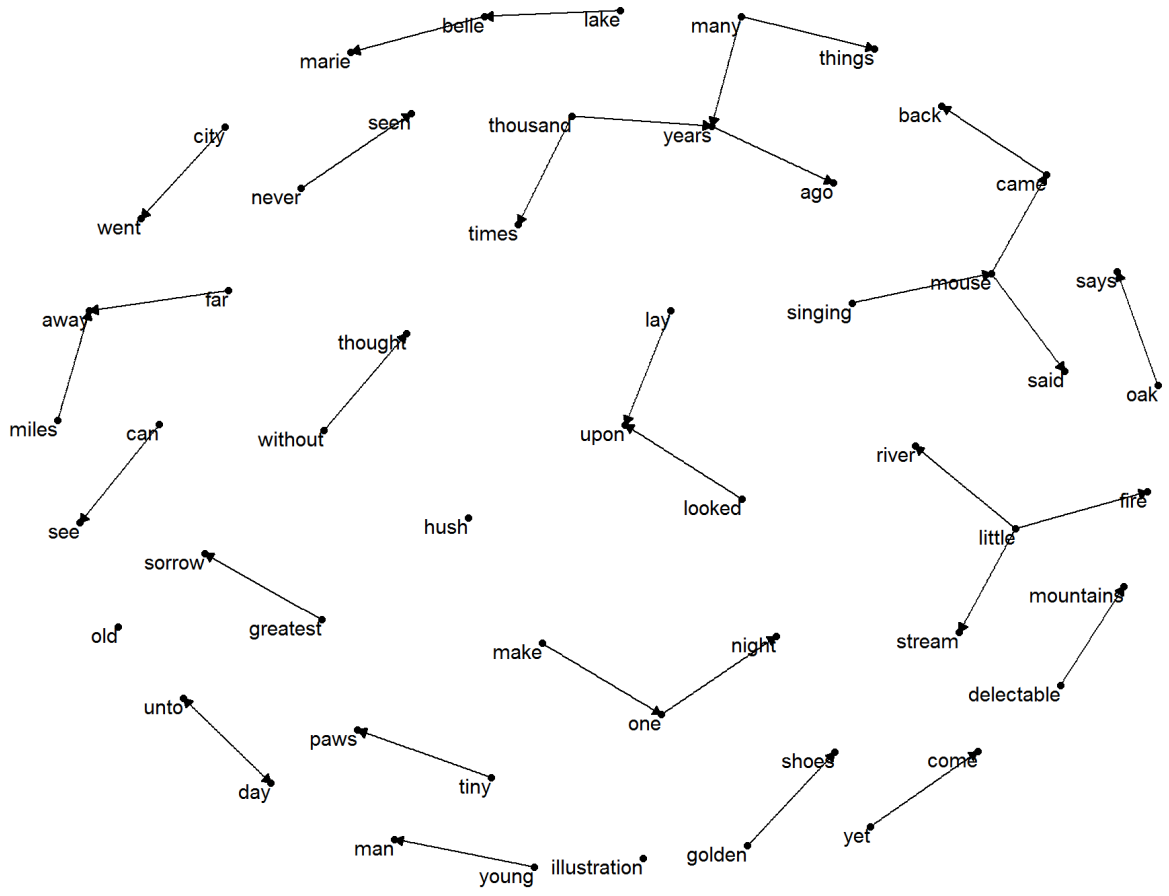

Figura 3: Red semántica de parejas de palabras utilizadas en el texto